Eötvös Loránd Tudományegyetem
Társadalomtudományi Kar
MESTERKÉPZÉS

A látens Dirichlet allokáció társadalomtudományi alkalmazása

A kuruc.info romaellenes megnyilvánulásainak tematikus elemzése

Konzulens:
Dr. Simon Dávid

Készítette:
Balogh Kitti
HQ9LC7
Survey statisztika szak

2015. április

TARTALOMJEGYZÉK

KÖSZÖNETNYILVÁNÍTÁS

Ezúton szeretném megköszönni kollégáimnak, Szabó Martina Katalinnak, Régeni Annának, Varjú Zoltánnak és Morvay Gergelynek az annotálásban nyújtott segítséget, amellyel hozzájárultak a szakdolgozatom színvonalasabbá tételéhez. Külön szeretnék köszönetet mondani Varjú Zoltánnak, akinek szakértelme, hasznos tanácsai és az adatgyűjtési, adatfeldolgozási és adattisztítási folyamatokban nyújtott nagyfokú segítsége nélkülözhetetlenek voltak a szakdolgozatom megszületéséhez. Emellett általánosságban szeretném megköszönni a Precognoxnak az inspiráló légkört, amely alkalmas terepet nyújtott ahhoz, hogy a szakdolgozatom egy céges projektből fejlődhessen ki.

1. BEVEZETÉS

A 2000-es évek közepétől kezdve egyre inkább elterjedtek és elfogadottá váltak a cigányellenes megnyilvánulások a hazai társadalmi és médiabeli diskurzusokban. A médiában megjelenő cigányellenes, rasszista retorika amellett, hogy megteremtette a nemzeti radikális és a szélsőjobboldali politikai szerveződések diszkurzív alapját, a közbeszédben is legitimizálta a rasszista kijelentéseket és felszabadította a rasszizmus rejtett formáit [Feischmidt, 2013]. A szakdolgozatom célkitűzése ennek a romaellenes diskurzusnak a vizsgálata, amelyet a nemzeti radikális, szélsőjobboldali média zászlóshajójának számító kuruc.info online hírportál cikkein vizsgálok.

Az új információ- és kommunikációtechnológiai eszközök hatására egyre nagyobb mennyiségben termelődő adatot a hagyományos adatfeldolgozási módszerekkel képtelenek vagyunk feldolgozni, ezért a társadalomtudósoknak új módszerekre van szükségük az adatok elemzése és értelmezése céljából. A hagyományos kvalitatív diskurzuselemzéshez ezért egy a szövegbányászat, a természetes nyelvi feldolgozás és az információkinyerés területén az utóbbi évtizedben nagy népszerűségnek örvendő valószínűségi nyelvi modellt, a látens Dirichlet allokációt fogom igénybe venni. A látens Dirichlet allokáció az ún. topik modellek (topic models) legegyszerűbb tagja. Segítségével nagy mennyiségű szöveges adat látens szemantikai, tematikus struktúráját tárhatjuk fel hierarchikus bayesi elemzéssel.

A dolgozatom célja tehát kettős. Egyrészt szeretnék a magyar média romareprezentációit vizsgáló diskurzuselemzésekhez kapcsolódni és a szélsőjobboldali online média elmúlt nagyjából kilenc évében zajló diskurzus tematikus struktúráját elemezni. Másrészt egy alternatív, a kvalitatív diskurzuselemzést kvantitatív elemekkel ötvöző módszertant kívánok megvalósítani, amely során egy a romareprezentációs vizsgálatok hazai szakirodalmában eddig nem használt módszer, a látens Dirichlet allokáció alkalmazhatóságát próbálom ki. A kiválasztott módszer alkalmazásának előnye, hogy a kutató munkájának egyes lépéseit automatikussá teszi, ezzel gyorsabb, kevesebb emberi erőforrást igénylő munkamenethez juthatunk. Emellett nagy mennyiségű szöveges adat vizsgálatához ad lehetőséget, amelyet tisztán kvalitatív diskurzuselemzés során nehéz lenne kivitelezni. A dolgozat második felében a módszer alkalmazásának társadalomtudományi relevanciáját szeretném megmutatni a

kuruc.info romaellenes cikkeinek tematikus elemzésén keresztül.

1.1. Romareprezentációs kutatások a magyar írott médiában

A magyar média romareprezentációs kutatások történetéről Kriza és Vidra 2010-es, valamint Munk 2013-as tanulmánya nyújt átfogó betekintést, a 2010-es években készült vizsgálatokról pedig Bogdán, Feischmidt és Guld szerkesztésében kiadott *„Csak másban". Romareprezentáció a magyar médiában* című 2013-as tanulmánykötetből tájékozódhatunk. A felsorolt írások nagy segítséget nyújtottak a most következő irodalmi áttekintés megszerkesztésében.

A cigányokkal kapcsolatos szöveges médiatartalmakat az 1960-as években kezdték Magyarországon is elemezni, melyeket kezdetben kvantitatív tartalomelemzési, a későbbiekben tartalomelemzési és kvalitatív diskurzuselemzési módszereket vegyítve tártak fel. A rendszerváltás előtti időszakban azonban a romák nem igazán jelentek meg a médiában, vagy ha igen, akkor pozitív képet rajzoltak fel a cigányok helyzetéről, jellemzően a munkavállalással, a lakhatással és az iskoláztatással kapcsolatban. [Kriza and Vidra, 2010, Munk, 2013]

A médiabeli diskurzusban az 1990-es évek közepétől váltak láthatóvá a romákkal kapcsolatos tartalmak, és a '90-es évek végéig a téma sajtóbeli jelenléte egyre nőtt. Ezzel közel egy időben a társadalmi diskurzusba is begyűrűzött a cigánykérdés. Vicsek Lilla 1996-ban végzett vizsgálata szerint 1995-ben az országos napilapokban három-négynaponta, 1998-ban már kétnaponta jelentettek meg cigányokkal kapcsolatos híreket. A cikkek feltárását kvantitatív tartalomelemzéssel végezte, amelyet a szövegkörnyezet kvalitatív elemzésével egészített ki. Az általa vizsgált cikkek a romákról egyöntetűen negatív képet mutattak, a cikkek 35%-ában jelent meg az előítélet, a diszkrimináció és az etnikai konfliktus témaköre, és egynegyedükben számoltak be bűnözésről és egyéb deviáns viselkedésről. Vicsek emellett azt állapította meg, hogy a cigánysággal kapcsolatos témák a cikkek több mint kétharmadában jelennek meg hírként, amely az összefüggéseket feltáratlanul hagyja és „elbagatellizálja" a témát. A problémaorientált reprezentáció Bernáth és Messing 1998-as felmérése alapján sem változott. Ezt a vizsgálatot hat országos napilapon végezték, amelyet a módszertan viszonylagos állandósága mellett 2003-ban és 2012-ben is elvégeztek, és éppen jelenleg folytatnak egy átfogó kutatást, melyben a magyar média ropmareprezentációját vizsgálják 1988-tól 2015-ig[1]. Az 1998-ban vizsgált cikkek egynegyede kötötte össze a romákat a bűnözéssel és további egynegyede számolt be helyi inter-

[1] http://szociologia.tk.mta.hu/a-roma-mediakep-1988-2015

etnikus konfliktusokról. Kutatásuk szerint az 1990-es évek végére megszaporodtak a cigányság témájával foglalkozó hírek és a roma tematika begyűrűzött a társadalmi diskurzusba. Bernáth és Messing a kutatás módszeréül a tartalomelemzést választották, amit interjúkkal egészítettek ki, a kutatás zárásaként pedig javaslatokat fogalmaztak meg egy elfogadóbb média érdekében. Vicsek, valamint Bernáth és Messing is megállapították, hogy az általuk elemzett médiumok romamegjelenítéseikben a romákat nem individualizáltan, hanem általánosságban, egybefüggő „masszaszerű" csoportként jelenítik meg. [Bernáth and Messing, 1998, Kriza and Vidra, 2010, Munk, 2013, Pócsik, 2013]

Bernáth és Messing a következő években több szűkebb témával kapcsolatos kutatást közöltek, pl. 2000-ben a székesfehárvári, zámolyi gettóügyről vagy 2004-ben Messing a jászladányi iskolaügyről. Az 1998-as kutatásukhoz hasonló tartalomelemzést Bernáth végzett 2003-ban, szintén hat országos napilap cikkein. Az 1998-as és a 2003-as vizsgálat eredményeinek összevetése alapján az derült ki, hogy a 2000-es évek elejére valamivel visszaszorult az arctalan romabemutatások aránya. Azonban a romákat a bűnözéssel és a konfliktusokkal összekapcsoló cikkek aránya továbbra is magas maradt, a hírek 37%-a foglalkozott konfliktusokkal és diszkriminációval, 21%-a pedig bűnözéssel. Emellett a szociális kérdések, szegénység témakör aránya is megnőtt (20%-ról 32%-ra), és a kormányzati politika, támogatások témája is népszerűbb lett. Közel egy időben, 2004-ben Terestyéni is folytatott tartalomelemzést az összes magyar heti- és napilap négy hónapot felölelő mintáján. A kedvezőtlen jegyet említő hírek mindössze hét százalékában vélte felfedezni a bűnözés témáját, ami az összes hír 2%-át jelentette. Mindezt úgy mérte, hogy a mintavétel idején zajlott a zámolyi gyilkosság nyomozása és bírósági eljárása. A két kutatás eltérése a korpuszok különbözőségéből és a bűnözés témájának eltérő kódolásából adódhatott. A sikert említő cikkekben leggyakrabban roma sikerekről esett szó, legfőképpen kulturális teljesítmények, zenei kvalitások és tanulási eredmények kapcsán. Azonban neki is általános megállapítása, hogy a sajtó inkább negatív színezetű képet fest a romákkal kapcsolatban, a szegénységgel, az elesettséggel, a kirekesztettséggel és a deviáns életmóddal köti össze a cigányságot. A 2000-es évek első feléből még érdemes megemlíteni a 2005-ben minisztériumi megbízásból elkészült Zöld könyvet, amely a médiában egyre gyakrabban megjelenő cigányság témájára összpontosított. A Zöld könyv nem új kutatási eredményeket, hanem a könyvet készítő médiaszakemberek észrevételeit és ajánlásait tartalmazza, valamint az addigi tudományos eredmények áttekintését. [Bernáth, 2003, Terestyéni, 2004, Kriza and Vidra, 2010, Pócsik, 2013, Munk, 2013]

Az említett kutatások alapján azt a következtetést vonhatjuk le, hogy a magyar médiában az 1990-es évek közepétől 2006-ig tartó időszakban a nyílt cigányellenes, rasszista megnyilvánulások nem voltak kifejezetten jellemzőek. A változás 2006-ban jött az olaszliszkai lincselés hatására. Az esemény kiváltotta cigányellenes megnyilvánulások szinte berobbantak a médiába és a közbeszédbe, és a 2000-es évek második felére a nyíltan rasszista, cigányellenes beszéd széles körben elterjedt, és elfogadottá vált. Ennek a folyamatnak a talán legjellemzőbb példája, hogy az addig tabunak és politikailag inkorrektnek számító „cigánybűnözés" kifejezés rohamosan terjedni kezdett a médiabeli és a köznapi beszédben. Juhász szerint egy igen széles rétegnek a cigánybűnözés kifejezés az igazság kimondásának szimbólumává vált és az olaszliszkai események után visszatérő jelképe lett a cigány-nem cigány konfliktusoknak [Munk, 2013]. A cigányellenes hullámot pedig még jobban felverte a 2009. február 8-i Cozma-gyilkosság, ami szintén óriási gyűlöletet és felháborodást váltott ki a magyar médiában és a közbeszédben. Az esetet a médiában jellemzően úgy tálalták, hogy az elkövetőket hol egyértelműen, hol burkoltan összemosták a teljes cigánysággal. Pócsik Andrea 2007-es elemzésében arra mutatott rá, hogy az alapvetően cigányellenes attitűddel rendelkező magyar társadalomban a rasszista interpretációk egy romaellenes befogadóban nézeteinek helyeslését fejezhetik ki, ezzel megerősítve azokat. [Kriza and Vidra, 2010, Munk, 2013]

A tartalom- és diskurzuselemzésekkel párhuzamban érdemes kitekinteni a cigányellenes attitűdöket mérő társadalomtudományi survey-kre is, ugyanis ezek alapján a cigányellenes retorika nyílt elterjedése nem vonható párhuzamba a cigányellenes társadalmi attitűdökkel [Kriza and Vidra, 2010, Munk, 2013]. A 2013-ban megjelent Társadalmi Riport egyik tanulmánya ([Bernát et al., 2013]) egy 2011-es felmérés adatait közli. A felmérés szerint tízből egy magyar ért azzal egyet, hogy támogatást kellene adni a cigányoknak, tízből hatan értenek egyet azzal, hogy „a cigányok vérében van a bűnözés", tízből nyolcan vélik úgy, hogy „ha végre munkába állnának a romák, akkor a gondjaik megoldódnának" és tízből négyen helyeslik, hogy „léteznek még szórakozóhelyek, amelyek diszkriminálják a romákat". Ez az eredmény az 1994-es, az 1997-es, a 2000-es, a 2002-es és a 2008-as eredményektől nem igazán tér el, ami a magyar társadalom cigányellenes attitűdjének stabilitását és általános negativitását jelzi. Az Ipsos 2014. júliusi adatfelvétele alapján sem történt változás a magyar népesség cigányellenes attitűdjében[2]. A 2000-es évek elején tapasztalható enyhülés a sajtó diskurzusában tehát feltehetően csak annak volt köszönhető, hogy ebben az időszakban a politikai korrektség szellemisége általánosan elterjedtebb volt

[2] http://pcblog.atlatszo.hu/2014/07/17/a-jobbik-taboraban-merseklodott-a-ciganyellenesseg/

[Munk, 2013]. A cigányokkal kapcsolatos attitűdök, vélemények tehát nem változtak igazán, de a cigányokkal kapcsolatos rasszista beszéd megjelent és elfogadottá vált a közbeszédben és a politikai beszédben is. Bernáth és Messing 2011-es kutatása szerint mindennapossá váltak a romákkal kapcsolatos témák, és a híreket leginkább a bűnözés és a politika témákban írták. A szélsőjobboldali diskurzus szóhasználata és tematizációi bekerültek a nyilvánosságba, tehát az előítéletes kijelentések nem csak a szélsőjobboldali csoportokat jellemzik, hanem szélesebb körben megfigyelhetőek. 2011-ben mérésük alapján a hírek 31%-a kapcsolta össze a romákat a bűnözéssel, 11%-ban a szegénységről és a roma sztárokról szóltak. [Bernáth and Messing, 2012, Munk, 2013] Ez a változás egy összetettebb társadalmi és gazdasági helyzetbe ágyazódik bele, amelynek okait a szélsőjobboldali politikai erők (Jobbik, Magyar Gárda és egyéb nem hivatalos polgárőr szerveződések) előretörésében és hatékony kommunikációjában, a gazdasági válság okozta egzisztencális bizonytalanságban, a roma érdekvédelmi mozgalmak gyengülésében és a roma közösségek szegregációjának felerősödésében kereshetjük [Bernáth and Messing, 2012], de ezek elemzése túlmutat a szakdolgozat keretein.

1.2. A dolgozat felépítése

A fent tárgyalt hazai szakirodalomban a magyar írott média dikurzuselemzését általában kvalitatívan végezték, talán csak Bernáth és Messing 2012-es felmérése említhető meg, amely során a kvalitatív elemzés mellett három csoportot különítettek el a cikkek képzeteiről klaszterelemzéssel. A feltárásra váró adatok mennyisége azonban új elemzési módszerek bevonását igényelné, amelyek ugyan nem váltják fel a kvalitatív elemzést, de egy részét automatizáltabbá, gyorsabbá tennék és a kutatás kevesebb humán és pénzbeli erőforrásból válna megoldhatóvá. Emellett az alkalmazandó módszer bevonásával csökkenthető a szubjektivitás és javítható a kutatás reprodukálhatósága. A szakdolgozatban ezért a problémát vegyes kvantitatív-kvalitatív diskurzuselemzéssel szeretném feldolgozni, amelyet a lehetőségekhez mérten legnagyobb mértékben szeretnék kvantitatívan végezni. Hazai viszonylatban nem találtam erre példát szigorúan társadalomtudományi területről, azonban a külföldi szakirodalomban olvashatunk ilyen vállalkozásokról, például [Baker et al., 2013] könyve, amelyben a kritikai diskurzuselemzést házasították össze korpusznyelvészeti módszerekkel az iszlámról alkotott sajtókép elemzése céljából.

A dolgozat második fejezetében felvázolom a látens Dirichlet allokáció elméleti és történelmi keretét, és értelmezem a feldolgozás során használt fogalmakat. Röviden

bemutatom a bayesi következtetéselmélet legfontosabb elemeit és a vegyes tagságú modellek (mixed membership models) családját, amelyhez a topik modellek és a látens Dirichlet allokáció tartoznak. Ezután áttekintem a látens Dirichlet allokáció matematikai hátterét, a kutatás jelenlegi állását és a módszer társadalomtudományi alkalmazásait.

A dolgozat harmadik fejezetében ismertetem a módszer számítógépes kivitelezhetőségét és a kuruc.info Cigánybűnözés rovatának cikkein végzett elemzés menetét és eredményeit. Az elemzés során bemutatom a látens Dirichlet allokáció társadalomtudományos diskurzuselemzésének relevanciáját. Végül a cikkek egy kisebb mintáján összehasonlítom az általam alkalmazott módszer eredményét a Bernáth és Messing által kidolgozott témastruktúra [Bernáth and Messing, 2012] szerint készült kvalitatív elemzés eredményével.

2. MATEMATIKAI ÉS TÖRTÉNETI HÁTTÉR

A fejezet során a látens Dirichlet allokáció matematikai hátterét, a topik modellek kutatásának jelenlegi állását és gyakorlati jelentőségét szeretném ismertetni. Előtte azonban a látens Dirichlet allokáció tágabb statisztikai elméleti keretébe szeretnék betekintést nyújtani, a bayesi következtetéselméletbe, valamint a vegyes tagságú modellcsalád jellemzőibe, amelybe a topik modellek és a látens Dirichlet allokáció tartozik. Ezután a látens Dirichlet allokáció közvetlen előzményeit mutatom be, a látens szemantikus indexelést és a valószínűségi látens szemantikus indexelést. A szakdolgozat keretei csak futólagos betekintést engednek utóbbi témákba, azonban a logikus felépítés és a tiszta megértés érdekében szükségesnek tartom a használt fogalmak és gondolati sémák tisztázását.

A *látens Dirichlet allokáció* egy valószínűségi modell, amely hierarchikus (többszintű) bayesi adatelemzésen alapulva tárja fel egy szöveggyűjtemény, vagy más néven korpusz látens szemantikai struktúráját. A *bayesi adatelemzés* alatt olyan módszereket értünk, amelyekkel a megfigyelt és a minket érdeklő változók valószínűségi modelljét alapul véve vonunk le következtetéseket egy megfigyelt minta feltétele mellett [Gelman et al., 2014]. A látens Dirichlet allokáció esetében a minta a korpusz, amelyben a dokumentumok szavai jelentik a megfigyelt változókat. A vizsgálat céljául szolgáló látens változó pedig a szemantikai struktúra, azaz a topikok struktúrája [Blei and Lafferty, 2009].

Következő lépésként tekinstük át a statisztikai következtetések bayesi megközelítését.

2.1. Bayesi következtetéselmélet

A bayesi megközelítés központi eleme a Bayes-tétel, amelyet Thomas Bayes angol statisztikus és filozófus vetett papírra az XVIII. század közepén. A XX. és XXI. századra e gondolatból nőtte ki magát a bayesi statisztika, amely egy a mai napig uralkodó a frekventista statisztikát többé-kevésbé tagadó alternatív módszertan [Hu-

nyadi, 2011].[1] A következőkben a Bayes-tétel egyszerű formáját, valamint a bayesi következtetés során értelmezett alakját írom le [Hunyadi, 2011] és [Gelman et al., 2014] alapján, ezután a bayesi következtéselmélet fogalmait tárgyalom [Hunyadi, 2011], [Gelman et al., 2014] és [Kehl and Várpalotai, 2013] munkákat alapul véve.

Ha A és B két eseményt jelöl, a $P(B|A)$ feltételes valószínűség a *Bayes-tétel* szerint a következő formában írható fel:

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A)}.$$

Ha $B_1, B_2, \ldots, B_n$ teljes eseményrendszert alkotnak, akkor bármely pozitív valószínűségű A eseményre érvényes, hogy:

$$P(B_i|A) = \frac{P(A|B_i)\,P(B_i)}{\sum_{j=1}^{n} P(A|B_j)\,P(B_j)}.$$

A bayesi következtetés során a Bayes-tételt nem egyes események valószínűségére írjuk fel, hanem valószínűség-eloszlásokra, tehát diszkrét valószínűségi változók esetén a lehetséges értékekhez tartozó valószínűségek összegeire, folytonos valószínűségi változók esetén sűrűségfüggvényekre. Két mennyiségre vonatkoztatjuk a tételt, az y-nal jelölt megfigyelt mintára és a θ-val jelölt paraméterre:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \propto p(y|\theta)p(\theta). \tag{2.1}$$

Az y a mintát, a tapasztalati eredményeket jelenti, a θ paraméter pedig a vizsgálat tárgyát, amelyek általában többváltozós mennyiségeket reprezentálnak. A $p(\theta)$ a paraméterre vonatkozó a priori feltevéseinket jelöli, amelyet egyszerűbben *prior*nak nevezünk. A prior azokat az ismereteket, szubjektív vélekedéseket jelöli, amelyeket a minta megfigyelése előtt, azaz a mintától függetlenül feltételezünk. A bayesi elemzés során tehát intézményesen veszünk figyelembe mintától független információkat a frekventista megközelítéstől eltérően. A $p(y)$ az adott minta, tapasztalati eredmények előfordulását leíró valószínűségi eloszlás. Nagyobb hangsúlyt kapnak a bayesi következtetéselméletben a $p(y|\theta)$ és az $p(\theta|y)$ feltételes valószínűségek. A $p(y|\theta)$ a frekventista statisztikában is használatos likelihood függvény, a $p(\theta|y)$ pedig a posteriori eloszlás, egyszerűbben *posterior*, amely az az eloszlás, amelyet a megfigyelt adatok feltétele mellett kapunk. A posterior tehát azt mondja meg, hogy a megfigyelt minta ismeretében milyen a vizsgálat tárgyát képező paraméter eloszlása. Ez

[1] Manapság inkább mindkét módszertan ismeretének fontosságára és a két irány közötti szintézisteremtésre hívják fel a statisztikusok figyelmét.

alapján kitűnik, hogy a statisztikai becslés tárgya a frekventista statisztikától eltérően nem egy ismeretlen, de rögzített érték, hanem egy ismeretlen valószínűségi változó.

A 2.1 egyenletben a $\propto$ jel az arányosságot jelöli, azaz a posterior arányos a likelihood és a prior szorzatával. Az arányosságot az teszi lehetővé, hogy a frekventista megközelítéstől eltérően csak egyetlen minta kiválasztásában gondolkodunk. Ehhez az egy mintához tartozó valószínűségeloszlás pedig nem függ θ-tól, azaz egy konstans értéket határoz meg. Ettől a tényezőtől emiatt eltekinthetünk. Így az y minta a posterior következtetést csak a $p(y|\theta)$ likelihood függvényen keresztül befolyásolja.

A prior eloszlás, illetve sűrűségfüggvény nem magától adódik, hanem előzetes ismereteinket és vélekedéseinket fejezzük ki vele. A prior előállításánál három dologra kell figyelnünk. Az első, hogy a prior független legyen a mintától, azaz olyan információkat hordozzon, amelyek a mintában nem találhatóak meg, ugyanis az információismétléssel romlik a az elemzés értéke. A másik, amire érdemes figyelnünk, hogy olyan függvényformát válasszunk, amelyik a paraméter(ek) mozgatásával különböző feltételezéseket tud leírni. A prior paraméterét/paramétereit a bayesi elemzés során gyakran *hiperparaméterek*nek hívjuk. Ezek testesítik meg az a priori információkat, ezért gyakran az eloszlás momentumai, mint a várható érték vagy a variancia azok a jellemzők, amelyek segítségével a szubjektív, előzetes vélekedéseinket be tudjuk vezetni a modellbe. Harmadikként pedig arra kell ügyelnünk, hogy a prior és az előálló likelihood szorzata lehetőséget adjon a további lépésekre. Azt az esetet, amikor a posterior eloszlás ugyanazt a parametrikus formát követi, mint a prior eloszlás, *konjugáció*nak hívjuk, a priort pedig *természetes konjugált*nak.

A bayesi következtetés számításigényes fázisa a posterior értékeléséhez és a posterior alapján történő elemzésekhez köthető, amilyen például a konstans kiintegrálása, a sokdimenziós eloszlások marginálisainak meghatározása vagy a posterior esélyhányadosok meghatározása. A számításigényes bayesi elemzések emiatt csak a számítógépek és az egyszerűbb programozási nyelvek elterjedésével kaptak igazán lendületre a '80-as évek második felétől. Ekkor fejlesztették ki azokat az algoritmusokat, amelyek különböző mintavételeken alapuló szimulációs technikákkal kezelni képesek ezeket a feladatokat. Az egyik népszerű és nagyon hatékony algoritmuscsalád, a Markov-láncokon alapuló Monte-Carlo algoritmusok (Markov Chain Monte Carlo, MCMC), amelynek gyakran használt tagja a Gibbs-mintavétel. Az MCMC algoritmusokra és a Gibbs-mintavételre a dolgozat későbbi részében bővebben kitérek.

A bayesi adatelemzés folyamatát összefoglalásképp három lépésben írhatjuk le

[Gelman et al., 2014] alapján:

1. Első lépésben létrehozunk vagy kiválasztunk egy az adott kutatási problémához leginkább illeszkedő valószínűségi modellt, amely a megfigyelt minta és a kutatás tárgyát képező változók együttes valószínűségeloszlását fejezi ki. Ebben a lépésben határozzuk meg a priort. A létrehozott modellnek illeszkednie kell a problémáról alkotott tudományos tudáshoz és az adatgyűjtési folyamathoz.

2. Második lépésben kiszámítjuk a posterior eloszlást.

3. Harmadik lépésben pedig kiértékeljük a modell illeszkedését és levonjuk a következtetéseket a megfelelő posterior eloszlás szerint. A modell ellenőrzésnél azt vizsgáljuk meg, hogy mennyire illeszkedik a modell az adatokhoz, a valós következtetések megokolhatók-e és hogy mennyire érzékenyek az eredmények az első lépésben megalkotott vagy kiválasztott modell feltételeire. Ezektől függően megváltoztathatjuk vagy bővíthetjük a modellt és megismételhetjük a lépéseket.

A későbbi alfejezetekben a látens Dirichlet allokációt ezen lépések logikája alapján tárgyalom.

A bayesi következtetéselmélet áttekintése után áttérek a vegyes tagságú modellek rövid történeti és általános matematikai áttekintésére, majd a látens Dirichlet allokáció közvetlen előzményeit, a látens szemantikus indexelés és a valószínűségi látens szemantikus indexelést mutatom be.

2.2. Vegyes tagságú modellek

A vegyes tagságú modelleket (mixed membership models) olyan esetekben érdemes használnunk, amikor egy sokdimenziós többváltozós adathalmazt szeretnénk klaszterezni, azonban azt feltételezzük, hogy azok nem csak egy, hanem a populáció több kategóriájába is beletartozhatnak. A dolgozat témájához csatlakozva az adathalmazunk a dokumentumokból álló korpusz, amelyet különböző kategóriákba, topikokba szeretnénk sorolni, azonban egy dokumentum több topikhoz is tartozhat. Ezt a helyzetet a vegyes tagságú modellek úgy kezelik, hogy minden egyes egyedről vagy megfigyelési egységről (dokumentumról) azzal a feltételezéssel élnek, hogy az összes klaszterbe beletartoznak, és a beletartozás mértékét egy a megfigyelésekhez tartozó tagsági vektorként reprezentálják, azaz úgynevezett fuzzy klaszterezést

(fuzzy clustering/soft clustering) hajtanak végre. A tagság mértéke a látens változók folytonos nemnegatív vektora, amelyek összege 1. Mivel a megfigyelésekhez tartozó tagsági vektorban lesz néhány olyan klaszter, amely jóval nagyobb súlyt/valószínűséget kap a többi klaszterhez képest, és sok olyan klaszter lesz, amely kicsi, nullához közeli súlyt/valószínűséget kap, a vegyes tagságú modellek egyben dimenziócsökkentő eljárásnak is tekinthetők. A vegyes tagságú modellek tehát egységes keretben kezelik a fuzzy klaszterezést és a dimenziócsökkentést, emellett bizonyos esetekben nagy előnyük lehet, hogy nem-felügyelt módszerek, azaz nincs szükségünk annotált adatokra, amikkel a modellt be kellene tanítanunk. [Airoldi et al., 2014] A vegyes (mixed) szó az alternatív látens osztályozó modellektől jön, ahol a vegyes azt jelenti, hogy minden attribútum a saját eloszlásának megfelelően generálódik [Crain et al., 2012].

Az első vegyes tagságú modelleket különböző tudományterületeken kezdték el használni egymástól függetlenül, tipikusan kategoriális adatok elemzése céljából. A korai vegyes tagságú modellek közé tartozik a gyógyászati klasszifikációt szolgáló Grade of Membership modell (Woodbury et al., 1978), az admixture modell (Pritchard et al., 2000), amelyet genetikai vizsgálatokhoz alkottak meg, és a gépi tanulás területén létrejött látens Dirichlet allokáció modell (Bleit et al., 2003), amely modellt a dolgozatban magyar nyelvű romaellenes cikkek tematikus struktúrájának vizsgálatára fogok felhasználni. [Airoldi et al., 2014]

A vegyes tagságú modellek eredeti ötlete Max Woodbury matematikustól származik, aki fuzzy osztályozással szerette volna megoldani az orvosi diagnózisok bizonyos problémáit. Az ötlet azonban nem kapott sok figyelmet a statisztikusoktól egészen a 2000-es évek elejéig, amikor a bayesi módszertan szélesebb körben elterjedt és Erosheva egy új bayesi megközelítést fejlesztett ki a modellhez. Az ezzel majdnem egyidőben és függetlenül létrehozott admixture modell és látens Dirichlet allokáció szintén a bayesi következtetésre és közelítő eljárásokra támaszkodnak. [Airoldi et al., 2014]

A vegyes tagságú modellek egy közös keretben egyesítik ezeket a modelleket, lehetőséget adva további modellek létrehozására a populációra vonatkozó feltételek, a mintavételi egységek és a látens változók szintjeinek, valamint a mintavételi sémák variálásával. A vegyes tagságú modellek eltérő gyökerei miatt a modelleknek két általános interpretációja alakult ki, a vegyes tagságú modellek általános alakja és a generatív folyamat. A dolgozat szempontjából a generatív folyamat a jelentősebb, ugyanis a látens Dirichlet allokáció irodalmában ezt az interpretációt szokás használni. Emellett a generatív folyamat a vegyes tagságú modellek definiálásnak egy

intuitívabb módját prezentálja. [Airoldi et al., 2014, Galyardt, 2014]

2.2.1. A generatív folyamat

A gépi tanulás szakirodalomban a vegyes tagságot leggyakrabban generatív folyamatként reprezentálják. Ennek a leírásnak az elterjedtsége a látens Dirichlet allokáció népszerűségének köszönhető, amelynek szakirodalmában konzisztensen a generatív folyamatként való reprezentációt használják [Galyardt, 2014]. A generatív folyamat leírásához [Galyardt, 2014] tanulmánya szolgált alapként.

A generatív leírás szerint a populáció K profilt tartalmaz, ahol $k = 1, \ldots, K$, és minden $i = 1, \ldots, N$ egyén különböző mértékben tartozik a profilokhoz. Ha például a populáció egy dokumentumokból álló korpusz, a profilok a dokumentumokban rejlő topikokat/témákat reprezentálják.

Minden egyénnek van egy *tagsági vektor*a, $\theta_i = (\theta_{i1}, \ldots, \theta_{iK})$, amely azt határozza meg, hogy egy egyén milyen mértékben tartozik bele a profilokba. Az *egyén* egyszerűen a populáció egy tagjára utal, például egy dokumentumra, egy képre, egy génre vagy egy személyre. A θ komponensei nemnegatívak és összegük 1, tehát θ egy valószínűségi vektornak tekinthető. Például ha egy dokumentum 40%-ban romák és nem romák közötti konfliktusokról szól és 60%-ban a kisebbségi oktatásról, akkor a dokumentum tagsági vektora $\theta_i = (0.4, 0.6, 0, \ldots, 0)$ lehet, attól függően, hogy a témák k indexe milyen értéket vesz fel.

Minden megfigyelt x_j változónak, ahol $j = 1, \ldots, J$, különbözik a valószínűségi eloszlása a profilokon belül. Például a kisebbségi oktatás témán belül a szavak eloszlása más lesz, mint a romák és nem romák közötti konfliktus topikon belül. Egy x_j változó maga is lehet többdimenziós, és ez esetben megfigyelhetjük az x_j változó $r = 1, \ldots, r_{ij}$ megvalósulásait minden i egyén esetében, amiket x_{ijr} jelöl. Az X_j eloszlását k profilon belül az F_{kj} valószínűségeloszlás-függvény adja meg.

A z_{ijr} indikátorvektor azt jelöli, hogy melyik profilhoz tartozik a j megfigyelt változó i egyénhez tartozó r megvalósulása. Például, a z_{ijr} szövegelemzés esetén azt mutatja meg, hogy az i-dik dokumentumban lévő j szó r megvalósulása milyen topikhoz tartozik.[2]

A θ_i tagsági vektor azt jelzi, hogy az egyének milyen mértékben tartoznak az egyes profilokhoz, ha $z_{ijr} \sim \text{Mult}(\theta_i)$. Ha a j megfigyelt változó r megvalósulása az i egyénen belül a k profilhoz tartozik, akkor $z_{ijr} = k$. Az x_{ijr} eloszlása z_{ijr}

[2] A látens Dirichlet allokáció esetén a szavaknak nincsenek különböző megvalósulásai, de el lehet képzelni például egy tájnyelvi kutatást, ahol a „megy” és a „megyen” a „megy” szó megvalósulásai.

indikátorvektorra mint feltételre nézve

$$x_{ijr}|z_{ijr} = k \sim F_{kj}.$$

A teljes generatív eljárás i egyén esetében a következő:

1. Válasszunk véletlenszerűen egy θ_i profileloszlást, ahol $\theta_i \sim D(\theta)$.
2. Minden $j = 1, \ldots, J$ változóra:
   (a) Minden $r = 1, \ldots, r_{ij}$ megvalósulásra:
      i. Válasszunk egy z_{ijr} profilt, ahol $z_{ijr} \sim \text{Mult}(\theta_i)$.
      ii. Válasszunk egy x_{ijr} megfigyelést, ahol $x_{ijr} \sim F_{z_{ijr},j}(x_j)$ és z_{ijr} profilhoz rendelt x_j eloszlásból származik.

A generatív folyamat tehát azt a folyamatot írja le, ahogyan minden i egyén esetében kiválasztódik az egyén k profilhoz tartozásának mértéke és az egyén j attribútumai az i k profilhoz tartozásának mértéke alapján generálódnak. Például ha a populáció dokumentumok gyűjteménye, akkor a generatív folyamat azt írja le, ahogyan egy dokumentum megírásánál kiválasztjuk a dokumentum témájául szolgáló topikokat, és a dokumentum szavait a topikok függvényében határozzuk meg. Ha ezt minden egyes dokumentum minden egyes szavának esetében elvégezzük, előáll a populáció egésze, azaz a korpusz. A generatív folyamat tehát egy nagyon intuitív módját írja le egy korpusz keletkezésének.

A téma szempontjából elégségesnek vélem csupán a generatív folyamat reprezentáció bemutatását, azonban a Függelék A részében a vegyes tagságú modellek általános alakja szerinti reprezentáció is olvasható.

A következő alfejezetben a látens Dirichlet allokáció közvetlen előzményeit, a látens szemantikus indexelést és a valószínűségi látens szemantikus indexelést mutatom be. Utóbbi a látens Dirichlet allokációval együtt a vegyes tagságú modellek szöveges adatokra alkalmazott változataihoz tartoznak, melyeket leggyakrabban *topik modelleknek* (topic models) hívnak a szakirodalomban.

2.3. A látens szemantikus indexeléstől a topik modellekig

Ahogy az előző fejezetben a vegyes tagságú modellek történeti fejlődésénél láttuk, a látens Dirichlet allokáció alkotja a vegyes tagságú modellek egyik alappillérét. Azonban a látens Dirichlet allokáció nem előzmény nélküli a szöveges adatok látens

szemantikai struktúrájának kinyerése terén, két fontos módszer, a látens szemantikus indexelés (Latent Semanic Indexing, LSI) és a valószínűségi látens szemantikus indexelés (probabilistic Latent Semanic Indexing, pLSI) előzi meg történetileg. A fejezet során először az LSI, a pLSI és az LDA használatához szükséges dokumentumreprezentációs modellt, a vektortérmodellt mutatom be, ez ugyanis szükséges a módszerek matematikai leírásának megértéséhez. Ezután az LSI és a pLSI matematikai működését ismertetem. A fejezetet [Crain et al., 2012], [Tikk et al., 2007], [Bolla and Krámli, 2012] és [Hofmann, 1999] alapján építem fel. A [Tikk et al., 2007] szövegbányászati tankönyvet a Vektortérmodell és a Látens szemantikus indexelés fejezeteknél vettem igénybe, a [Bolla and Krámli, 2012] statisztikai tankönyvet a Látens szemantikus indexelés fejezetnél, a [Crain et al., 2012] és a [Hofmann, 1999] tanulmányt pedig A valószínűségi látens szemantikus indexelés fejezetnél.

2.3.1. Vektortérmodell

A látens szemantikus indexelést, a valószínűségi látens szemantikus indexelést és a látens Dirichlet allokációt dokumentumok rendszerezéséhez használjuk, azaz segítségükkel dokumentumokat hasonlítunk össze és előre nem rögzített csoportokba soroljuk azokat. Ehhez a szövegbányászati feladathoz a szóban forgó három módszer a széles körben használt *vektortérmodell* (vector space model) nevű dokumentumreprezentációs modellt alkalmazza, amely a dokumentumokat egy sokdimenziós vektortérben ábrázolja. A vektortérmodell dimenzióit a korpuszban előforduló egyedi szavak[3] alkotják, ezek feszítik ki a vektorteret. Tehát a vektortérmodellben minden szót egy tengellyel reprezentálunk, az egyes dokumentumok pedig az őket alkotó szavakból álló vektorokként állnak elő.

A vektortérmodellben egy $D = \{d_1, \ldots, d_N\}$ dokumentumgyűjteményt az úgynevezett *szó-dokumentum mátrix* (term-document matrix, tdm) formában ábrázoljuk, ahol $\mathbf{D} \in \mathbb{R}^{M \times N}$. A mátrixban az oszlopok száma (N) a dokumentumok számával, a sorok száma (M) a korpuszban előforduló egyedi szavak számával egyezik meg. A szó-dokumentum mátrix d_{ki} eleme a k-adik szó (t_k) relevanciáját jelöli az i-edik dokumentumban. A d_i dokumentumot megtestesítő vektort $d_i = [d_{i1}, \ldots, d_{iM}]$-vel jelöljük.

[3] Egyedi szavaknak nevezzük a korpuszban előforduló egyéni szavakat, amelyek a szótár elemeit alkotják.

$$\mathbf{D} = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1N} \\ d_{21} & d_{22} & \cdots & d_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{M1} & d_{M2} & \cdots & d_{MN} \end{pmatrix}$$

A **D** szó-dokumentum mátrixban az oszlopokat tehát a dokumentumvektorok alkotják, a sorokat pedig a szóvektorok.[4] A mátrix azokban a cellákban tartalmaz nulla értéket, amelyek által meghatározott szó nem szerepel adott dokumentumban vagy nulla a relevanciája. Az egyedi szavak összességét hívjuk *szótár*nak vagy *lexikon*nak. A szótár hossza megegyezik az egyedi szavak számával, M-mel. A szó-dokumentum mátrixok a természetes nyelv jellegénél fogva nagyon ritkák, ugyanis kevés olyan szó van, amely sok dokumentumban fordul elő és sok olyan szó található a korpuszban, amely csak kevés dokumentumban jelenik meg.[5] A szótár mérete nagyon nagy lehet, akár milliós nagyságrendű is, ezért érdemes a mátrix méretét csökkenteni, amelyben nyelvtechnológiai és matematikai eszközök lehetnek segítségünkre. A látens szemantikus indexelés és a valószínűségi látens szemantikus indexelés, valamint a következő fejezetben soron következő látens Dirichlet allokáció mind matematikai megoldást jelentenek erre a problémára, ugyanis dimenziócsökkentő eljárásokként működnek. Ezek használata azonban nem váltja ki a nyelvtechnológiai eszközöket, együttes használatuk vezet a leginkább kívánatos eredményhez.[6]

Érdemes kiemelni, hogy a vektortérmodell szerinti dokumentumreprezentáció a szavak dokumentumon belüli helyét és sorrendjét figyelmen kívül hagyja. Ezt a szakirodalomban *szózsákmodell*nek (bag of words) szokás nevezni. A *zsák* egy általánosított halmazfogalom, amely egy olyan matematikai objektumot jelöl, amelyben egy elem akár többször is megjelenhet. A vektortérmodellt leggyakrabban pont amiatt a tulajdonsága miatt éri kritika, hogy a szavak egymás utáni sorrendje elvész. Azonban ez az egyszerűsítés általában nem okoz problémát, a gyakorlati feladatok nagy részében ugyanis nincsen szükség a szavak sorrendjére vonatkozó információra.[7] Ugyanígy van ezzel a látens szemantikus indexelés, a valószínűségi látens szemantikus indexelés és a látens Dirichlet allokáció is, amelyek nem a szavak szövegbeli helye, hanem

[4] Egyesek a szó-dokumentum mátrix transzponáltját, a dokumentum-szó mátrixot használják, amelynek soraiban találhatók a dokumentumvektorok és az oszlopaiban a szóvektorokat.

[5] A természetes nyelvekben a szavak eloszlása az úgynevezett Zipf-eloszlást követi, amely egy diszkrét hatványfüggvény-jellegű eloszlás. Az eloszlás elnevezése George Kingsley Zipf amerikai nyelvészhez kötődik.([Tikk et al., 2007])

[6] A dokumentumok előfeldolgozásának további lépéseit pl. tokenizálás, szótövezés, stopszószűrés a dolgozat 3. fejezetében, a kutatás kivitelezésénél, a gyakorlatban mutatom be.

[7] A szavak sorrendjére vonatkozó információra általában szintaktikai feladatoknál van szükség.

a szavak együttes előfordulása alapján reprezentálják a dokumentumok szemantikai struktúráját.

2.3.2. Látens szemantikus indexelés

A látens szemantikus indexelés (latent semanic indexing, LSI) a dokumentumvektorokat kisebb dimenziójú vektorokká alakítja át szinguláris értékfelbontás segítségével. Az új csökkentett dimenziók a dokumentumokból kinyerhető együttes szóelőfordulási mintázatok alapján az eredeti dimenziók kombinációjaként jönnek létre. Az LSI-vel kapott dimenziók közvetlenül nem értelmezhetők, azonban meglehetősen eredményesen adják vissza a dokumentumokban rejlő látens szemantikai struktúrát. Ennek köszönhetően olyan dokumentumok is egy kategóriába kerülhetnek, amelyek nem tartalmaznak egy szót sem a kategóriára jellemző szavak közül, azonban gyakran fordulnak elő együtt adott kategória szavaival. Emiatt a tulajdonsága miatt az LSI a szinonimák és a többértelmű szavak problémájára is megoldást jelent.[8]

A szinguláris értékfelbontás során a $\mathbf{D} \in \mathbb{R}^{M \times N}$ szó-dokumentum mátrix, ahol N a dokumentumok, M a szavak száma, felírható

$$\mathbf{D} = \mathbf{U}\mathbf{S}\mathbf{V}^T$$

alakban. Az $\mathbf{U}$ mátrix egy $M \times R$-es mátrix, és oszlopvektorai a $\mathbf{D}^T \cdot \mathbf{D}$ sajátvektorai, a $\mathbf{V}$ mátrix pedig egy $N \times R$-es mátrix, amely $\mathbf{D} \cdot \mathbf{D}^T$ sajátvektorainak felel meg. $\mathbf{U}$ és $\mathbf{V}$ mátrixok oszlopai ortonormáltak, azaz egymásra ortogonálisak. Az $\mathbf{S}$ $N \times R$-es mátrix diagonálisa tartalmazza a $\mathbf{D}^T \cdot \mathbf{D}$ sajátértékeinek négyzetgyökét. $R \leq min(M, N)$ a $\mathbf{D}$ mátrix rangja. Ezután átrendezzük az $\mathbf{S}$, valamint az $\mathbf{U}$ és a $\mathbf{V}$ mátrixokat úgy, hogy a szinguláris értékeket csökkenő sorrendben tartalmazzák. Az $\mathbf{S}$ nulla vagy nulla közeli értékei a mátrix jobb alsó részmátrixába kerülnek, amely értékekhez tartozó sorok és oszlopok elhagyhatók és a $\mathbf{D}$ mátrix dimenziója lecsökken. Ezzel párhuzamosan az $\mathbf{U}$ és a $\mathbf{V}$ mátrixból is töröljük a megfelelő oszlopokat, illetve sorokat. Ha nem csak zérus értékeket hagyunk el, akkor a $\mathbf{D}_{red}$ mátrix az eredeti $\mathbf{D}$ mátrix R_{red} rangú közelítése lesz, ahol R_{red} a megmaradt értékek számával egyenlő:

$$\mathbf{D} \approx \mathbf{D}_{red}, \qquad \mathbf{D}_{red} = \mathbf{U}_{red}\mathbf{S}_{red}\mathbf{V}_{red}^T.$$

A $\mathbf{D}_{red}$ mátrix megtartja a $\mathbf{D}$ mátrix legfontosabb strukturális elemeit, miközben a zajt és szóhasználatból származó problémákat kiiktatja. $\mathbf{D}_{red}$ mátrix dokumentumvektorai a $\mathbf{V}_{red}\mathbf{S}_{red}$ sorai lesznek, a mátrix új dimenziói pedig a $\mathbf{U}_{red}\mathbf{S}_{red}$ sorai.

[8] Ahogyan a pLSI és az LDA is.

2.3.3. A topik modellek és a valószínűségi látens szemantikai indexelés

A vegyes tagságú modellek szövegekre alkalmazott tagjait *topik modellek*nek hívjuk, amelyeket széles körben használnak dokumentumgyűjtemények szemantikai struktúrájának kinyeréséhez.[9] A topik modellek feltételei természetes módon ragadják meg a nyelv heterogenitását, azt a jelenséget, hogy a dokumentumok több szemantikailag koherens témát tartalmazhatnak, méghozzá különböző mértékben. A topik modellezés során a topikokat a szótár szavai feletti eloszlásként írjuk le, a dokumentumokat pedig a szótár szavainak egy halmazaként kezeljük, ahol a szavak az adott témához tartozó eloszlásból jönnek. A topik modellek tulajdonképpen azt a feltételes valószínűséget modellezik, ahogyan egy író a kiválasztott téma vagy témák függvényében a szavakat papírra veti. Ha megfigyelünk egy korpuszt, ez a folyamat a posterior eloszlás kikövetkeztetésével rekonstruálható, amely közben előtűnnek a dokumentumokban rejlő topikok, a szavak együttes előfordulásának jellemző mintázatai.

A *valószínűségi látens szemantikus indexelés* (probabilistic latent semantic indexing, pLSI) fontos lépésnek tekinthető a valószínűségi topik modellezés felé, tulajdonképpen a pLSI tekinthető az első topik modellnek és egyben a közvetlen előfutára a legszélesebb körben használt topik modellnek, a látens Dirichlet allokációnak. Az LSI-hez képest a pLSI és a topik modellek előrelépést jelentenek több szempontból is. Egyrészt az LSI alkalmazásakor a korpusz bővülése esetén a transzformált vektortérmodellt újra kell számolnunk, míg a pLSI és a topik modellek esetében az új dokumentumhoz is tudunk topikeloszlást illeszteni, bár a pLSI esetében kicsit komplikáltabb módon. Másrészt a pLSI és a topik modellek valószínűségi modellek, ezért az elemzésük a hagyományos statisztikai eszköztárral történik, sőt a bayesi statisztika módszereivel a modellek még robusztusabbá tehetők. Emellett a topik modellek rugalmasak, könnyen bővíthetők, ha a feladat jellege megkívánja.

Az LSI a korpuszt reprezentáló szó-dokumentum mátrixot szinguláris értékfelbontás segítségével vetíti egy kisebb dimenziójú szemantikus térbe, azonban a pLSI, valamint topik modellek valószínűségi megközelítése az LSI-nél egy még jobban értelmezhető látens szemantikus teret hoznak létre [Hofmann, 1999].

A pLSI-t Thomas Hofmann dolgozta ki 1999-ben. A módszert többek között szöveges adatok információkinyeréséhez és klaszterezéséhez alkalmazzák. A pLSI ugyanazokra a tartalmi előfeltételekre épül, mint az LSI, de egy egészen különböző valószínűségi generatív folyamatot használ egy korpuszban található dokumentumok

[9] Az angol topic model kifejezésnek nincsen meghonosodott magyar fordítása, ezért a továbbiakban a modellekre vonatkozóan a topik modell kifejezést használom, a szemantikus egységekre pedig felváltva hivatkozok topikként vagy témaként.

szavainak generálásához.

A pLSI magja az úgynevezett *aspektus modell*, a dokumentumgyűjtemények olyan látens változós modellje, amely minden megfigyeléshez egy $z \in Z = (z_1, \ldots, z_K)$ látens topik változót társít, azaz minden egyes $d \in D = (d_1, \ldots, d_N)$ dokumentum minden egyes $w \in W = (w_1, \ldots, w_m)$ szava esetén. A pLSI a következő generatív folyamatot írja le [Hofmann, 1999, Wei and Croft, 2006]:

1. Válasszunk ki egy $P(d) \sim \text{Mult}(d)$ vegyes topik eloszlást minden d dokumentum esetén.

2. Válasszunk ki egy z látens topikot $P(z|d)$ valószínűséggel minden w szó esetén.

3. Generáljunk egy w szót $P(w|z)$ valószínűséggel.

A generatív folyamat a következő együttes valószínűségeloszlásnak felel meg:

$$P(d, w) = P(d)P(w|d), ahol$$
$$P(w|d) = \sum_{j=1}^{K} P(w|z_j)P(z_j|d).$$

Tehát a pLSI egy dokumentum minden szavát egy keverék eloszlásból származó mintaként modellezi, ahol a keverék komponensek polinomiális eloszlású random változók. Ezek a polinomiális eloszlású random változó tekinthetők a topikok reprezentációinak. Minden szó egy adott topikból származik, és egy dokumentum különböző szavai különböző topikokból származhatnak. A pLSI minden dokumentumot úgy reprezentál, mint a topikok különböző arányú keveréke azáltal, hogy topikok fix halmazának valószínűségi eloszlására redukálja.

A pLSI modell megfelelő bázist nyújt szövegek elemzéséhez, de több probléma is adódik a modell alkalmazása közben. Az egyik gond, hogy a modell figyelembe veszi a dokumentumok sorrendjét, aminek az a következménye, hogy a paraméterek száma túl nagy lesz a modellben. Ez egyrészt azért gond, mert ahogy a dokumentumok száma növekszik, a paraméterek száma is lineárisan nő, másrészt a sok paraméter miatt a modell hajlamos az adatokat túlilleszteni. A másik probléma, hogy nincs természetes módja annak, ha egy új dokumentumhoz szeretnénk a topikeloszlást rendelni. A látens Dirichlet allokáció úgy oldja meg ezeket a problémákat, hogy nem csak a szavak sorrendjét nem veszi figyelembe, hanem a dokumentumok sorrendjét sem.

2.4. Látens Dirichlet allokáció

A látens Dirichlet allokációt (latent Dirichlet allocaton, LDA) David Blei, Andrew Y. Ng és Michael I. Jordan fejlesztették ki 2002-ben [Blei et al., 2003]. A látens Dirichlet allokáció a topik modellek egyik legegyszerűbb és egyben legnagyobb népszerűségnek örvendő modellje. A következő alfejezetekben bemutatásra kerül a módszer matematikai leírása, a topik modellek kutatásának jelenlegi állása és társadalomtudományi használatuk relevanciája. Ezek után a látens Dirichlet allokáció egy társadalomtudományi példán való gyakorlati alkalmazását próbálom ki diskurzuselemzés céljából, amelynek leírására a következő fejezetben kerül sor.

2.4.1. A generatív folyamat leírása

A generatív folyamat azt a képzeletbeli véletlen folyamatot írja le, ahogyan a megfigyelt adat, azaz egy adott korpusz keletkezik. A generatív folyamat leírásához [Blei et al., 2003], [Blei and Lafferty, 2009] és [Heinrich, 2005] irodalmakat használtam fel. Legyen D a dokumentumok száma, K a topikok száma, V pedig a szótár mérete, melyek mindegyike egy konstans skalár. Továbbá legyen α egy pozitív K-dimenziós vektor vagy egy skalár, és β pozitív V-dimenziós vektor vagy egy skalár. A leírás során a α-t vektorként, β-t skalárként kezelem.[10] $\mathrm{Dir}_K(\vec{\alpha})$ legyen egy K-dimenziós Dirichlet eloszlás az $\vec{\alpha}$ vektorparaméterrel és $\mathrm{Dir}_V(\beta)$ jelöljön egy V-dimenziós szimmetrikus Dirichlet eloszlást a β skalár paraméterrel. Legyen N_d a d dokumentum hossza, $\vec{\theta}_d$ a d dokumentumhoz tartozó topikeloszlás, ahol Θ egy $D \times K$ mátrix, $\vec{\phi}_k$ pedig a k topikhoz tartozó szóeloszlás, ahol Φ egy $K \times V$ mátrix. Valamint legyen $z_{d,n}$ a d dokumentum n-dik szavához tartozó topikhozzárendelés, mely egy indikátor, $w_{d,n}$ pedig a d dokumentum n-dik szavának szóindikátora.

A teljes generatív eljárás a következő:

1. Minden egyes k topik esetében
   (a) válasszunk egy $\vec{\phi}_k \sim \mathrm{Dir}(\beta)$ szóeloszlást.
2. Minden egyes d dokumentum esetében
   (a) válasszunk egy $\vec{\theta}_d \sim \mathrm{Dir}(\vec{\alpha})$ topikeloszlást.
   (b) Minden egyes w szó esetében

[10] Szimmetrikus Dirichlet-eloszlás esetén a paraméterek értéke egyenlő. Erről a későbbiekben lesz szó.

i. válasszunk egy $z_{d,n} \sim \text{Mult}(\vec{\theta}_d)$ topik hozzárendelést, ahol $z_{d,n} \in 1, \ldots, K$,

ii. válasszunk egy $w_{d,n} \sim \text{Mult}(\vec{\phi}_{z_{d,n}})$ szót, ahol $w_{d,n} \in 1, \ldots, V$

2.4.2. Dirichlet prior és posterior polinomiális eloszlás esetén

A látens Dirichlet allokáció a szavakat és a topikokat kategoriális változóként kezeli, ahol a kategóriák előbbi esetében a szótár szavai, utóbbi esetében a szemantikai jelentéssel bíró topikok. Az LDA azt feltételezi ezekről a változókról, hogy eloszlásuk polinomiális. A polinomiális eloszlás a binomiális eloszlás általánosítása, amikor a kísérletek során kettőnél több kimenetel lehetséges. Tegyük fel, hogy n független, azonos kísérletnek k lehetséges kimenetele lehet. Legyen $x_{i,j} = 1$, ha az i kísérlet eredménye a j kategóriába esik, és $x_{i,j} = 1$ egyébként. Ebben az esetben $x_i = (x_{i1}, \ldots, x_{ik})$ egy polinomiális kísérletet jelöl, ahol $\sum_{j=1}^{k} x_{ij} = 1$. Például a (0,0,1,0) a 3. kategória kimenetelét jelöli négy lehetséges kategória közül. Jelölje $n_j = \sum_{i=1}^{n} x_{ij}$ a j kategóriában kijött sikeres kísérletek számát. Az $(n_1, \ldots, n_k)$ összegek eloszlása polinomiális. [Agresti, 2002]

Jelölje $p_j = P(x_{ij} = 1)$ annak a valószínűségét, hogy egy kísérlet kimenetele a j kategóriába esik. A polinomiális eloszlás sűrűségfüggvénye

$$p(n_1, \ldots, n_k) = \frac{n!}{n_1! \ldots n_k!} p_1^{n_1} \ldots p_k^{n_k} = \frac{n!}{\prod_{j=1}^{k} n_j!} \prod_{j=1}^{k} p_j^{n_j}$$

A binomiális eloszlás a polinomiális eloszlás speciális esete $k = 2$ esetén. A polinomiális eloszlás tulajdonságai:

$$E(n_j) = np_j, \quad var(n_j) = np_j(1 - p_j), \quad cov(n_j, n_k) = -np_jp_k$$

A bayesi következtetés során egy paraméter eloszlása a megfigyelt adatra mint feltételre nézve, azaz a posterior eloszlás arányos a prior eloszlás és a likelihood függvény szorzatával. A polinomiális eloszlás természetes konjugáltja a Dirichlet eloszlás, amely a béta eloszlás többváltozós általánosítása. [11] A polinomiális eloszlás likelihoodja függvénye a

$$\prod_{i=1}^{k} p_i^{n_i}$$

szorzattal arányos. [12]

[11] A béta eloszlás a binomiális eloszlás természetes konjugáltja.

[12] Az előző jelöléshez képest a j futóindex helyett i futóindexet használok.

A Dirichlet eloszlás sűrűségfüggvénye a látens Dirichlet allokáció θ paraméterére felírva

$$p\left(\theta|\alpha_1,\cdots,\alpha_K\right) = \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_k^{\alpha_k-1},$$

ahol $0 \leq \theta_k \leq 1$ minden k esetén és $\sum_{k=1}^{K}\theta_k = 1$. A Γ a gamma függvényt jelöli, amely felfogható úgy, mint a faktoriális függvény valós számokra való kiterjesztése. Az eloszlás p paramétertől függő magja a $\prod_{k=1}^{K}p_k^{\alpha_k-1}$. A posterior eloszlás szintén Dirichlet a $\{n_k+\alpha_k\}$ paraméterekkel. A prior és a posterior eloszlás parametrikus formája megegyezik, ezért a Dirichlet eloszlás valóban természetes konjugáltja a polinomiális eloszlásnak.

A látens Dirichlet allokáció a Dirichlet eloszlás szimmetrikus formáját is használja a β paraméter esetében, az α paraméter esetében pedig opcionálisan. A szimmetrikus esetben minden komponens értéke egyenlő.

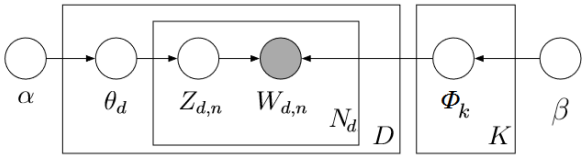
2.4.3. A látens Dirichlet allokáció együttes valószínűségeloszlása és grafikus modell reprezentációja

Az előző alfejezetben leírt generatív folyamat a megfigyelt és a látens valószínűségi változók együttes valószínűségeloszlását definiálja, amely átírható a hagyományos módon. A modellben a dokumentumok szavai a megfigyelt változók, a látens változókat pedig a topikstruktúra alkotja. A látens Dirichlet allokáció esetében a ϕ_k jelöli a topikokat, ahol minden egyes ϕ_k a szótár szavai feletti eloszlás. A d-edik dokumentum topik arányait a θ_d jelöli, ahol $\theta_{d,k}$ a d dokumentumban a k topikhoz tartozó topikarány. A d-edik dokumentumhoz rendelt topikot z_d jelöli, ahol $z_{d,n}$ a d dokumentum n-edik szavához rendelt topik. A d dokumentum megfigyelt szavait a w_d fejezi ki, ahol $w_{d,n}$ a d dokumentum n-edik szava, ami a szótár egy eleme.

Az LDA generatív folyamata a következő együttes valószínűségeloszlásnak felel meg:

$$p(\vec{\phi}_k,\vec{\theta}_d,\vec{z}_d,\vec{w}_d|\vec{\alpha},\beta) = \prod_{k=1}^{K}p(\vec{\phi}_k|\beta)\prod_{d=1}^{D}p(\vec{\theta}_d|\vec{\alpha})\left(\prod_{n=1}^{N_d}p(z_{d,n}|\vec{\theta}_d)p(w_{d,n}|\vec{\phi}_k,z_{d,n})\right)$$

Az együttes valószínűségeloszláson jobban leolvashatók a látens Dirichlet allokáció függőségi és feltételes függetlenségi feltételei. A $z_{d,n}$ topik hozzárendelés a dokumentumonkénti $\vec{\theta}_d$ topikaránytól függ, a megfigyelt $w_{d,n}$ szó a $z_{d,n}$ topikhozzárendeléstől és a $\vec{\phi}_k$ topikok összességétől. Műveletileg utóbbi úgy néz ki, hogy megnézzük, adott kifejezés melyik $z_{d,n}$ topikkal párosul és abban a topikban nézzük meg a $w_{d,n}$ szó valószínűségét.



2.1. ábra. LDA bayesi hálója, [Blei and Lafferty, 2009] alapján

A függőségi feltételek az LDA-hoz tartozó grafikus modellel is szemléletesen ábrázolhatók. (Ábra 2.1) A fenti grafikus modell egy *bayesi háló*, amely a grafikus modellnek egy olyan változata, amely egy irányított körmentes gráfként reprezentálja a valószínűségi változók együttes valószínűségeloszlását. A háló minden csúcsa egy valószínűségi változót jelöl. A látens változók üres körökkel, a megfigyelt változók szürke körrel vannak ábrázolva. A téglalapok a változók többszörös értékeit jelölik, azaz például az N_d-vel jelölt téglalap a szavak gyűjteményét jelöli a dokumentumokban, a D-vel jelölt a dokumentumokat a korpuszon belül. Bármely két csúcs feltételesen független egymástól a szüleikre mint feltételekre nézve.

A látens Dirichlet allokációnak léteznek egyéb interpretációi, mint a geometriai interpretáció és a mátrixfaktorizációs interpretáció. Előbbi megtalálható [Blei et al., 2003], [Steyvers and Griffiths, 2006] és [Ponweiser, 2012] munkákban, utóbbi [Steyvers and Griffiths, 2006] tanulmányban.

2.4.4. Posterior eloszlás közelítése Gibbs mintavétellel

Egy adott korpusz rejtett topik dekompozícióját úgy kaphatjuk meg, ha kiszámoljuk a posterior eloszlást, tehát a rejtett változók feltételes eloszlását a megfigyelt változókra, azaz a dokumentumok szavaira mint feltételre nézve [Blei and Lafferty, 2009, Blei, 2011]:

$$p(\vec{\phi}_k, \vec{\theta}_d, z_d | w_d) = \frac{p(\vec{\phi}_k, \vec{\theta}_d, z_d, w_d)}{p(w_d)}$$

A számlálóban a modellt alkotó valószínűségi változók együttes valószínűségeloszlása található, amely könnyedén kiszámolható a látens változók bármilyen értéke mellett. A nevező azonban a megfigyelések marginális valószínűségét tartalmazza, amelyet elméletileg úgy lehet kiszámolni, hogy összeadjuk az összes együttes valószínűségeloszlást az összes lehetséges látens topikstruktúra megtestesülése mellett. Ez azt jelenti, hogy az összes lehetséges módon hozzá kellene rendelnünk a szavakat a topikokhoz, ki kellene számolnunk ezek valószínűségét, amelyeket összeadunk és

integrálunk θ és ϕ tartományokon. Általában azonban a vizsgálni kívánt korpuszok szavainak száma legalább milliós nagyságrendű, emiatt az összes lehetséges topikstruktúra száma túl nagy, és gyakorlatilag kivitelezhetetlen a kiszámolása.

Ez a jelenség nem egyedüli, a bonyolultabb, sokváltozós valószínűségi modellek posterior eloszlását gyakran nem tudjuk kiszámolni az együttes eloszlások integrálása vagy a magas dimenziószám miatt. A modern valószínűségi modellezés egyik központi kutatási kérdése, hogy hatékony módszereket fejlesszenek ki a posterior eloszlás megfelelő közelítéséhez. A topik modellezéshez használt algoritmusok gyakran ilyen a posterior eloszlást közelítő általános célú módszerek adaptációi, amelyek például az előbbi egyenlet közelítését hajtják végre. A topik modellek posterior eloszlását közelítő algoritmusok általánosan két osztályba sorolhatók, a mintavételen alapuló algoritmusok és a variációs algoritmusok osztályába.

A dolgozatban a mintavételen alapuló eljárásokhoz tartozó Gibbs-mintavételt mutatom be, ugyanis a későbbiekben ez az algoritmus kerül alkalmazásra. A Gibbs-mintavétel a leggyakrabban alkalmazott Markov-lánc Monte Carlo (MCMC) módszer, amelyet [Kehl and Várpalotai, 2013] és [Steyvers and Griffiths, 2006] alapján dolgozok fel.

Az MCMC-technikák célkitűzése, hogy egy tipikusan összetett, sokdimenziós, akár ismeretlen sűrűségfüggvényből mintát tudjanak venni. Az MCMC-módszerek Markov-láncokat konstruálnak, amelyek időben függnek egymástól és egyensúlyi eloszlásuk a kívánt eloszlással egyezik meg. Minden iteráció utáni állapotot úgy tekint a kívánt eloszlásból származó mintaelemnek tekint. Ezek a mintaelemek azonban nem függetlenek egymástól a Markov-tulajdonság miatt. A lánc előállítása nem okoz különösebb problémát a gyakorlatban, inkább a konvergencia megállapítása okoz gondot.

A *Gibbs-mintavétel* a sokdimenziós problémákat kisebb dimenziós, egyszerűbb egységekre bontja le Markov-láncok felhasználásával. A megoldandó probléma a posterior marginális eloszlásainak, azaz az egyes paraméterek eloszlásainak, jellemzőinek megállapítása. A Gibbs-mintavétel a kívánt együttes eloszlásból indirekt módon, a feltételes eloszlásokból vesz mintát.

A látens Dirichlet allokáció modellben a minket érdeklő együttes eloszlások a $\vec{\phi}_k$ topik-szó és a $\vec{\theta}_d$ dokumentum-topik eloszlások. A Gibbs-mintavétel ahelyett, hogy közvetlenül a $\vec{\phi}_k$ ás a $\vec{\theta}_d$ eloszlásokból becsülne, kimarginalizálja a $\vec{\phi}_k$ és a $\vec{\theta}_d$ eloszlást és a $z_{d,n}$ topikhozzárendelések feltételes valószínűségéből becsülünk a $w_{d,n}$ megfigyelt szavakra mint feltételre nézve.

A Gibbs-algoritmus minden $w_{d,n}$ szón végigmegy a korpuszban, és megbecsüli az

adott szó k topikhoz való hozzárendelésének feltételes valószínűségét az összes többi szó topikhozzárendelésére mint feltételre nézve minden egyes k topik esetén. Ezekből a feltételes eloszlásokból választ ki az algoritmus véletlenszerűen egy topikot és azt rendeli hozzá adott $w_{d,n}$ szóhoz. Ezt a feltételes eloszlást a $p(z_{d,n} = k|z_{-d,n}, w_{d,n})$ alakban írjuk fel, ahol $z_{d,n} = k$ a $w_{d,n}$ szó k topikhoz való hozzárendelését, $z_{-d,n}$ az összes többi szó topikhozzárendelését jelenti. [Steyvers and Griffiths, 2006] a következőképp mutatta be a kiszámolást:

$$p(z_{d,n} = k|z_{-d,n}, w_{d,n}) \propto \frac{\mathbf{C}^{(j)}_{-(d,n),k} + \beta}{\sum_{j=1}^{V} \mathbf{C}^{(j)}_{-(d,n),k} + \beta} \frac{\mathbf{C}^{(d)}_{-(d,n),k} + \alpha}{\sum_{k=1}^{K} \mathbf{C}^{(d)}_{-(d,n),k} + \alpha}.$$

A j index azt jelöli, hogy a $w_{d,n}$ szó a szótár j elemével egyezik meg, ahol $j = 1, \ldots, K$. A $\mathbf{C}^{(j)}$ egy $V \times K$ dimenziós, a $\mathbf{C}^{(d)}$ pedig egy $D \times K$ dimenziós mátrix. $\mathbf{C}^{(j)}_{-(d,n),k}$ tartalmazza annak az összegét, ahányszor egy $w_{d,n}$ szó egy k topikhoz lett párosítva a jelenlegi d, n esetet nem tartalmazva. $\mathbf{C}^{(d)}_{-(d,n),k}$ tartalmazza annak az összegét, ahányszor egy d dokumentum valamelyik szavához egy k topik lett hozzárendelve a jelenlegi d, n esetet nem tartalmazva.

Egy szó topikhoz való hozzárendelését két faktor befolyásolja, amelyek az előző egyenlet két törtjét jelentik. A bal oldali tört egy $w_{d,n}$ szó k topikhoz való tartozásának valószínűségét jelöli, a jobb oldali egy k topik valószínűségét adja d dokumentumhoz tartozó aktuális topikeloszlás mellett. Ha sok szó rendelődik k topikhoz az összes dokumentum szintjén, nőni fog a valószínűsége, hogy további szavakat is ahhoz a topikhoz párosítson az algoritmus. Ezzel párhuzamosan ha egy dokumentumon belül több szó ugyanahhoz a topikhoz lett hozzárendelve, nőni fog a valószínűsége, hogy a dokumentumhoz tartozó többi szó is ahhoz a topikhoz legyen párosítva.

A Gibbs-algoritmus kezdő lépésként minden szót egy véletlenszerű k topikhoz rendel hozzá. Minden szó esetében a $\mathbf{C}^{(j)}$ és $\mathbf{C}^{(d)}$ gyakoriság mátrixok elemeit egyesével felváltja az aktuális topik hozzárendelésnek megfelelő elem. Ezután a fenti egyenletből egy új topikot választ véletlenszerűen az algoritmus és a gyakoriságmátrixokat feltölti az új topikhozzárendelésekkel. Minden Gibbs-minta minden szóhoz a korpuszban egy topikot párosít, amit azáltal ér el, hogy egyszer végigmegy az összes dokumentumon. A mintavétel kezdő lépéseinél, amit *beégési szakasz*nak (burnin period) hívnak, a Gibbs-mintáknak ki kell törlődniük, mivel azok gyenge közelítései a posterior eloszlásnak. A beégési szakasz után az egymást követő Gibbs-minták elkezdik közelíteni a céleloszlást, azaz a topik hozzárendelések posterior eloszlását. Bizonyos szabályos időközönként a Gibbs-mintákat eltárolja az algoritmus, hogy az eloszlásból reprezentatív mintákhoz jussunk kivédve a minták közötti korrelációt.

A Gibbs-algoritmus közvetlenül becsüli meg minden egyes szó z topikhozzárendelését. Ugyanakkor a $\vec{\phi}'$ topik-szó eloszlás és a $\vec{\theta}'$ dokumentum-topik eloszlás becsléseire is szükségünk van, amelyekhez a gyakoriságmátrixokból jutunk hozzá a

$$\hat{\phi}_k^{(j)} = \frac{\mathbf{C}_k^{(j)} + \beta}{\sum_{j=1}^{V} \mathbf{C}_k^{(j)} + \beta}, \qquad \hat{\theta}_k^{(d)} = \frac{\mathbf{C}_k^{(d)} + \alpha}{\sum_{k=1}^{K} \mathbf{C}_k^{(d)} + \alpha}$$

képletek alapján.

A konvergencia megállapítására több módszer létezik, amelyekbe [Cowles and Carlin, 2006] nyújt átfogó betekintést. A különböző szerzőknek különféle véleménye van arról, hogy hogyan kellene a konvergenciát beazonosítani, azonban minden eszköznek megvan a maga előnye és hátránya. Geweke egy hosszú Markov-lánc idősorelemzési eszközökkel történő diagnosztikáját javasolja. Gelman és Rubin több lánc használatát ajánlják, amelyeket különböző kezdőértékekről indítanak, és a láncokon belüli illetve közötti varianciákat hasonlítják össze. Gyakori diagnosztikai módszer, hogy a lánc egymás utáni lépéseiben generált paraméterértékeket ábrázolják, illetve ezek hisztogramját vizsgálják. Az elemzés során utóbbi módszereket fogom használni.

2.4.5. Látens Dirichlet allokáció modell illeszkedésének ellenőrzése

Ha sikerült a bayesi elemzés első két lépését elvégeznünk, azaz kiválasztottuk a feladathoz megfelelő valószínűségi modellt, valamint kiszámoltuk a paraméterek posterior eloszlását, harmadik lépésként ellenőriznünk kell a modell illeszkedésének jóságát (Goodness of Fit). Az elemzés során a probléma minden lényeges tulajdonságát egy valószínűségeloszlással ragadjuk meg, azonban, ennek a valószínűségeloszlásnak a helyes megválasztása nehéz feladat. Ezért érdemes megvizsgálni, hogy az illesztett valószínűségi modellnek sikerült-e a valóság minden lényeges aspektusát megragadnia. A modell ellenőrzése során a modell több részét is érdemes ellenőriznünk: a prior eloszlást, a mintavételi eloszlást, a hierarchikus struktúrát, és egyéb kérdéseket, például, hogy milyen magyarázó változókat foglaltunk bele a modellbe. [Gelman et al., 2014]

A látens Dirichlet eloszlás prior eloszlásának vizsgálata során arra jutottak, hogy aszimmetrikus α paraméter illesztésével robusztusabb modellhez jutunk, ugyanis a topikokat kevésbé fogják gyakori, szemantikailag kevésbé jelentős szavak uralni és a topikok stabilabbak maradnak a topikok számának növekedésével. Az aszimmetrikus Dirichlet eloszlású β paraméterrel illesztett modellek azonban nem mutattak

javulást a szimmetrikus eloszlású β paraméterrel illesztett modellekhez képest. A látens Dirichlet allokáció legjobb struktúráját tehát a dokumentum-topik eloszlások feletti aszimmetrikus prior és a topik-szó eloszlások feletti szimmetrikus prior adja. [Wallach et al., 2009a]

A látens Dirichlet allokáció modellfeltételeinek ellenőrzésében még nem történt sok előrelépés, noha David Blei, a modell egyik megalkotója több tanulmányában is a topik modellek egyik fő kutatási területeként szabta meg. [Blei, 2011] Egy próbálkozást lehet megemlíteni ebben az irányban, Mimno és Blei tanulmányát[Mimno and Blei, 2011], amelyben a szerzőpáros kidolgozott egy posterior prediktív ellenőrzést a látens Dirichlet allokáció egyik feltételes függetlenségi feltételének vizsgálatához. Azt a feltételt vizsgálták, miszerint egy $w_{d,n}$ szó független d dokumentum θ_d topik-eloszlásától a $z_{d,n}$ topik hozzárendelésre, mint feltételre nézve. Azaz, ha tudjuk egy w szó z topikhozzárendelését, az, hogy melyik dokumentumban van, nem ad plusz információt a w szóra nézve. Tehát ha egy szó csak a topikhozzárendeléstől függ, akkor az egy topikhoz tartozó szavak egymástól függetlenül származnak ugyanabból a polinomiális eloszlásból. A *posterior prediktív ellenőrzés* (posterior predicitive check, PPC) a modellfeltételek ellenőrzésének egy megbízható módját képviseli. Ha egy modell illeszkedik, akkor a prediktív eloszlás adataiból generált mintáknak és a modell által illesztett adatoknak "hasonlónak" kell lenniük. Ha valamilyen szisztematikus eltérést tapasztalunk a szimulációk és az eredeti adat között, az a modell elégtelenségét jelzi [Gelman et al., 2014]. Tehát a PPC-vel megállapíthatjuk, hogy a modellünk mikor illeszkedik a megfigyelésekhez, és beazonosíthatjuk a posterior azon részeit, amiket rosszul derít fel. Valós szöveges adatok elemzése során gyakran előfordul, hogy a korpusz nem illeszkedik azokhoz a függetlenségi feltételekhez, amiket a topik modellek megszabnak. A látens Dirichlet allokáció a topik modellek egyik legegyszerűbb változata, amelynek az utóbbi években sok a modellfeltételek szintjén kevésbé megalkuvó kiterjesztését, bővítését, módosítását fejlesztették ki, például a *Correlated Topic Modelt*, amely megengedi, hogy a topikok korreláljanak egymással. A modellfeltételek ellenőrzésével rájöhetünk, hogy mely feltételt nem elégítik ki az adataink, és ez alapján léphetünk tovább egy szűkebb vagy módosított modellre. Ezekről a módosított topik modellekről a következő alfejezetben lesz szó. Mimno és Blei kutatásának bővebb leírása az B fejezetben olvasható a függelék részeként. Mivel a módszer nincs implementálva a topik modellezéshez használt szoftverekben, a kutatásom során, melyet a következő fejezet tartalmaz, egy az R statisztikai programozási nyelvben írt kóddal kivitelezem a tesztet.

A valószínűségi nyelvi modellek teljesítményét, és így a látens Dirichlet alloká-

ció teljesítményét is loglikelihooddal vagy olyan loglikelihood alapú mérőszámokkal szokták mérni, mint a perplexitás. A teljesítmény kiszámolása történhet az egész korpuszon [Steyvers and Griffiths, 2006], esetleg egy másik (például dokumentumklasszifikációs vagy információkinyerési) feladaton [Wei and Croft, 2006], de a legáltalánosabb eljárás, hogy a korpusz tesztelés céljából megtartott (held-out) részén, a tesztkorpuszon számítják ki a modell teljesítményét. A topik modellezés során leggyakrabban a perplexitást (perplexity) [Blei et al., 2003], az empirikus likelihoodot (empirical likelihood) [Li and McCallum, 2008], a harmonikus átlag módszerét (harmonic mean method) [Griffiths and M. Steyvers, 2004], a Chib-stílusú becslést (Chib-style estimation) [Wallach et al., 2009b], a balról-jobbra mintavételt (left-to-right estimation) [Wallach et al., 2009b] és a fontossági mintavételt (importance sampling) [Li and McCallum, 2008] használják. A [Wallach et al., 2009b] tanulmány bővebb betekintést nyújt ezekről a teljesítményt mérő módszerekről, amelyek vizsgálata során arra a megállapításra jutottak, hogy ezek a módszerek egy része, mint a harmonikus átlag módszere vagy a fontossági mintavétel használhatónak bizonyulnak a modellek teljesítményen alapuló rangsorolására, azonban pontatlanok olyan szempontból, hogy a modellek választása közötti haszon mértékét tévesen reprezentálják. Ezek helyett a Chib-stílusú becslést és a balról-jobbra algoritmust ajánlják, amelyekkel a topik modellek közötti pontos értékelés is lehetséges.

A szakdolgozat kutatás részében ezek közül a teljesítményt mérő módszerek közül a harmonikus átlag módszerét fogom használni az optimális topikszám megtalálása érdekében. A kutatás során a topikok számáról vegyes módszerrel döntök, az ideális topikszámról kialakult a priori elképzelésem és a harmonikus átlag módszere adta eredmények alapján választom ki az optimális modellt.

A harmonikus átlag módszerét [Griffiths and M. Steyvers, 2004] használta először abból a célból, hogy az általuk elemzett PNAS természettudományos folyóirat 28,154 absztraktjának optimális topikszámát megtalálják. A módszer előnye, hogy a Gibbs-mintavétel állapotait elmentve könnyedén hozzájutunk és nem számításigényes. A módszer leírására [Griffiths and M. Steyvers, 2004] tanulmányát használom fel. A látens Dirichlet allokáció teljesítményének kiszámolása során a $p(w|K)$ likelihoodra vagyunk kíváncsiak, ahol w a korpusz szavai, K pedig a topikok száma. Ezt a likelihoodot nem tudjuk kiszámolni, mivel minden lehetséges szó-topik hozzárendelést összegeznünk kellene hozzá. Azonban meg tudjuk közelíteni azáltal, hogy amikor az algoritmus z-t véletlenszerűen kiválasztja a $p(z|w, K)$ posteriorból, elmentjük a $p(w|z, K)$ értékét. Ezt a műveletet bizonyos lépésenként megismételjük és a kapott

értékekből kiszámoljuk a $p(w|z, K)$ értékeket a következőképp:

$$P(w_{d,n}|z_{d,n}) = \left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V}\right)^K \prod_{k=1}^{K} \frac{\prod_V \Gamma\left(\mathbf{C}^{(j)}_{(d,n),k} + \beta\right)}{\Gamma\left(\sum_{j=1}^{V} \mathbf{C}^{(j)}_{(d,n),k} + \beta\right)},$$

ahol $\mathbf{C}^{(j)}_{(d,n),k}$ annak a gyakorisága, ahányszor a $w_{d,n}$ szó a k topikba került a $z_{d,n}$ hozzárendelések által és $\Gamma(\cdot)$ a standard gamma-függvény. Végül kiszámoljuk az így kapott értékek harmonikus átlagát.

A topik modellek illeszkedésének ellenőrzésére nincs kidolgozott, bevett módszertan, a meglévő módszerek között versengés folyik és a jelenlegi állás szerint a módszerek többsége nincsen a topik modellezéshez használt szoftverekben implementálva. Ezzel párhuzamosan arra sincsen meghatározott módszertan, hogy bizonyos modellfeltételek nem teljesülése esetén milyen más modellt érdemes választanunk. Ennek oka, hogy a topik modellezés egy friss és folyamatosan fejlődő területe a gépi tanulásnak. A következő fejezetben azokat a részterületeket szeretném áttekinteni, amelyek a látens Dirichlet allokáció és általában véve a topik modellek fontos kutatási területeinek számítanak.

2.4.6. A látens Dirichlet allokáción túl

A topik modellezés kutatása egy folyamatosan fejlődő terület, és egyes kutatási irányok még csak kezdetleges eredményeket tudnak felmutatni. Ezek közül a kutatási irányok közül csak az egyik az előző alfejezetben tárgyalt modelllilleszkedés és modellteljesítmény ellenőrzését szolgáló módszerek kifejlesztése. Az alfejezethez Blei egy 2011-es cikke [Blei, 2011] nyújt segítséget, melyben összefoglalja az aktuális és jövőbeli kutatási kérdéseket a topik modellezés területén. Az általa meghatározott kutatási irányok a mai napig meghatározónak számítanak.

A topik modellezés legaktívabb kutatási területe az olyan alternatív, szofisztikáltabb topik modellek kifejlesztése, amelyek a látens Dirichlet allokáció modellfeltételeit mérsékelik vagy éppen kiterjesztik. A következőkben ilyen modelleket veszünk sorra. Az LDA egyik feltétele, amely a szózsákmodell reprezentációból következik, hogy a szavak dokumentumon belüli sorrendje nem számít. Bár ez a feltételezés nyilvánvalóan nem fedi a valóságot, ha a célunk egy korpusz szemantikai struktúrájának kinyerése, akkor a szavak sorrendjét nem szükséges tudnunk. Azonban ha kifinomultabb céljaink vannak, például nyelvi generálás, akkor ez a feltétel nem tartható. A szavak nem-felcserélhetőségének kifejezésére több modellt is kidolgoztak az LDA kiterjesztéseként, például a *bigram topik modell* (Bigram Topic Model), amely

azt feltételezi, hogy a modell az előző szóra mint feltételre nézve generál egy új szót. [Wallach, 2006] Egy másik modell, a *kompozit modell* (Composite Model) két komponenst egyesít magában, egy rejtett Markov modellt (Hidden Markov Model, HMM) és egy topik modellt, amellyel egy szintaktikai és egy szemantikai generatív modellt párosítottak össze. [Griffiths et al., 2005] Egy másik LDA modellfeltétel szerint a dokumentumok egymás utáni sorrendje sem számít. Ez a feltétel akkor tűnhet a valóságnak nem megfelelőnek, amikor éveken, évtizedeken átnyúló korpuszokat elemzünk. Az ilyen korpuszok esetén joggal feltételezhetjük, hogy a topikok változnak az idő függvényében. Ennek a problémának az egyik megoldása a *dinamikus topik modell* (Dynamic Topic Model, DTM), amely figyelembe veszi a dokumentumok sorrendjét és gazdagabb topikstruktúrát nyújt az LDA-nál. [Blei and Lafferty, 2006] Egy harmadik feltétel, amellyel a látens Dirichlet allokáció él, hogy a topikok számát ismertnek és fixnek tekinti. Erre kínál megoldást a *bayesi nemparaméteres topik modell* (Bayesian Nonparametric Topic Model), amely a topikok számát a posterior következtetés során határozza meg. A bayesi nemparaméteres topik modelleket topikok hierarchiájára terjesztették ki, amelyek a topikoknak egy fáját alakítják ki, amely az általánosabb topikok felől halad a konkrétabb topikok felé. Egy következő feltétel, amelynek gyengítésére szintén több modellt kifejlesztették már, a topikok függetlensége. Ezek közé a modellek közé tartozik a *korrelált topik modell* (Correlated Topic Model, CTM) [Blei and Lafferty, 2006], és a *Pachinko allokáció* (Pachinko Allocation (Machine)) [Li and McCallum, 2008]. Létezik olyan topik modell is, amely megengedi, hogy a szavak kevésbé valószínűek legyenek egyes topikokban (Spherical topic model), továbbá olyan, amelyik a ritka topik modellekben további topik eloszlásokat juttat érvényre (Sparse Topic Model). Valamint olyan topik modellt is használnak, amely az ún. robbanásszerű (bursty) topik modellekben a szógyakoriságoknak egy a valóságnak jobban megfelelő modelljét nyújtja, ilyen a *Dirichlet compound Multinomial LDA*.

A kutatások egy másik iránya, amikor a topik modelleket olyan metaadatokkal egészítik ki a modellillesztés során, mint például a szerző neve, geográfiai helyek vagy linkek. Az egyik első ilyen metaadatot felhasználó modell a *szerző-topik modell* (Author-Topic Model), amely a szerzőkhöz különböző topikeloszlásokat illeszt. A szerző-topik eloszlások segítségével a szerzők közötti hasonlóságot is meg lehet állapítani. A másik sikeres metaadatot figyelembe vevő modell az *összefüggő topik modell* (Relational Topic Model), amely a dokumentumok közötti linkeket használja fel metaadatként. Az összefüggő topik modell a dokumentumokat az LDA feltételei szerint modellezi, és a linkeket úgy értelmezi, mint a dokumentumok topikeloszlása közötti

távolságokat. Utóbbi modell tehát egyszerre egy új topik- és hálózatmodell, amely a linkeket a csúcsok attribútumai (a dokumentumok szavai) szerint modellezi. A [Blei, 2011] cikkben további példákat hoz fel olyan metaadatok topik modellekbe foglalásáról, mint a nyelvi struktúra, a korpuszok közötti távolság vagy a névelemek. A [Mimno and Blei, 2011] tanulmányban három korpuszon próbálták ki a metaadatok hasznosságát, például hírek esetében a publikáció hónapját vagy a hírt tartalmazó rovatot, parlamenti felszólalások esetén az aktuális miniszterelnököt figyelembe véve. Roberts, Stewart és Tingley egy az egyes dokumentumokhoz tartozó metaadatokat egységes keretben kezelő modellt fejlesztettek ki, melynek a *strukturális topik model* (Structural Topic Model) nevet adták [Roberts et al., 2014a].

A [Blei, 2011] cikk a topik modellezés kutatásának három jövőbeli irányát emeli ki. Az egyik az előző alfejezetben említett probléma, amely a modellek kiértékelésére és a modell feltételek ellenőrzésére terjed ki. Ebben az alfejezetben érintettem, hogy a legáltalánosabb esetben a modelleket tesztkorpusz(ok)on értékelik ki, amely(ek)en kiszámolnak valamilyen kiértékelő mérőszámot. Azonban vannak olyan esetek, amikor erre nincsen különösebb technikai okunk. Például ha a topik modelleket arra szeretnénk használni, hogy segítségükkel egy nagyméretű korpuszt egységekbe tudjunk szervezni vagy összegezni tudjuk, egy az előbb felvázolt kiértékelés nem mondja meg, hogy egy modellnek milyen mértékben sikerült az adott korpuszt megszervezni vagy interpretálnia. Ezért olyan módszereket kell kidolgozni, amelyek segítségével meg tudjuk mondani, hogy mely modellfeltételek szükségesek, melyik modellt válasszuk és hogy hogyan döntsünk két ugyanarra a feladatra kifejlesztett topik modell között.

A második jövőbeli kutatási irányként Blei olyan új módszerek keresését határozta meg, amelyek alkalmasak a topikok és a korpuszok vizualizációjának minél explicitebb kivitelezésére. Ezzel összefüggésben szükség van olyan felhasználói felületek létrehozására, amelyeken a felhasználóknak lehetősége van a topik modellek interaktív kezelésére.

A harmadik még kevesek által gyakorolt kutatási irány a topik modellek olyan irányú felhasználása, amelyben a a topik modellek által feltárt topikokat, kulcsszavakat, dokumentum-topik eloszlásokat valamilyen történelmi, szociológiai, nyelvészeti, politikai, jogi vagy egyéb céllal szeretnénk felhasználni. Blei szerint ez egy új interdiszciplináris területet kínál a különböző területeken tevékenykedő kutatók számára.

A következő, a fejezetet záró alfejezetben ehhez az utóbbi kutatási irányhoz szeretnék csatlakozni, és a látens Dirichlet allokáció társadalomtudományi céllal felhasznált irodalmát szeretném áttekinteni. Az alfejezetet az indokolja, hogy a következő

nagy egységben egy rövid kutatásomat fogom bemutatni, amelyben a látens Dirichlet allokációt társadalomtudományi céllal alkalmazom a kuruc.info online hírportál Cigánybűnözés rovatán.

2.4.7. A látens Dirichlet allokáció és a topik modellek társadalomtudományi alkalmazása

Mintegy átkötésként a következő nagy fejezet és a mostani fejezet között, a látens Dirichlet allokáció és egyéb topik modellek társadalomtudományi alkalmazásának irodalmát szeretném áttekinteni. A társadalomtudományi alkalmazáson kívüli általános irodalomról a David Mimno által összegyűjtött irodalomjegyzék segíthet az eligazodásban, amely a http://mimno.infosci.cornell.edu/topics.html oldalon érhető el. Az alfejezetben olyan kutatásokról számolok be, amelyekben a topik modelleket arra használták, hogy egy társadalomtudományi, a modellből közvetlenül nem következő jelenséget magyarázzanak, derítsenek fel. Ezért az olyan vizsgálatok kikerültek a fókuszomból, amelyek ugyan egy társadalmi jelenséget vizsgáltak, de nem társadalomtudományi céllal, például [Yano et al., 2009] politikai blog posztokon végzett kutatása.

A topik modellek nyújtotta lehetőségeket elsősorban politikai és közpolitikai területeken fedezték fel maguknak a kutatók. A politikatudományi alkalmazások éllovasa az előző alfejezetben már említett Roberts, Stewart és Tingley szerzők, akik egy külön modellt fejlesztettek ki a dokumentumszintű metaadatok topik modellben való kezeléséhez. [Roberts et al., 2014a] A modellüket több társadalomtudományi példán kipróbálták különböző társszerzőkkel. A kínai események hírekben való prezentálását például két alkalommal is vizsgálták, egyszer öt hírügynökség (például az egyesült államokbeli Associated Press és a kínai Xihua) cikkein keresztül az 1997-2006-ig tartó időintervallumban [Roberts et al., 2014a], másodszor négy millió kínai cikken 2007-2014-ig [Roberts and Stewart, 2014]. Előzőben a hír forrását és az évet adták meg metaadatként és arra voltak kíváncsiak, hogy bizonyos események kapcsán mennyire tér el a hírforrások reprezentációja. Utóbbiban a tartományokat és a városokat adták meg metaadatként és azt elemezték, hogy az állami propaganda és cenzúra milyen stratégiái érvényesülnek tartományi és városi szinten.

Egy másik érdekes kutatásukban azt vizsgálták, hogy hogyan lehet kiváltani a survey-kben gyakori nyílt végű kérdések kvalitatív vizsgálatát topik modellezéssel. A kutatás tárgyául a kísérletben résztvevők politikai attitűdjei szolgáltak, amelyeket a kísérlet során egy beszélgetéssel próbáltak befolyásolni. A csoportokkal beszélgetést folytattak, amely során a kezelt csoportot igyekeztek érzelmileg bevonni a beván-

dorlók problémáival szemben, a kontroll csoporttal pedig egy objektív beszélgetést bonyolítottak le a bevándorlásról. Ezután egy kérdőívvel mérték meg a résztvevők politikai attitűdjeit. A strukturális topik modellnek metaadatként adták meg a szerzők, hogy a kezelt vagy a kontroll csoportba tartozott egy személy, a feltárt topikok pedig kimutatható eltérést mutattak a két csoport szóhasználata között. [Roberts et al., 2014b]

Egy szintén 2014-es tanulmányban Roberts, Stewart, Tingley és Lucas dzsihádista és nem-dzsihádista muszlim papok vallásos szövegeinek, főleg fatwáinak tematikus struktúráját tárták fel, amely során azt elemezték, hogy nézeteiktől függően milyen témákat fogalmaztak meg szövegeikben [Lucas et al., 2014].

Egy korábbi politikatudományi tanulmány [Grimmer, 2010] azt a kérdést járta körül, hogy az Amerikai Egyesült Államok szenátusának sajtónyilatkozataiban milyen témák jelennek meg, azaz a szenátus tagjai miről és hogyan kommunikálnak a választópolgárokkal. A vizsgálathoz Grimmer egy hierarchikus topik modellhez hasonló modellt használt, amelyben a szenátus tagjainak neve illetve az állam neve szerepelt metaadatként. Ezek alapján lehetősége nyílt a szenátus tagjainak tematizációit is összehasonlítani.

DiMaggio, Nag és Blei egy 2010-es vizsgálatuk során arra keresték a választ egy majdnem 8,000 cikkből álló korpuszon, hogy a cikkek hogyan keretezik nyelvileg a művészek illetve művészeti egyesületek Egyesült Államok kormányától kapott támogatásait, ösztöndíjait. A vizsgálat során látens Dirichlet allokáció segítségével határozták meg a művészeti támogatások leggyakrabban használt témáit. Ezenkívül különbségeket állapítottak meg a támogatások sajtóbeli reprezentációjáról a különböző kormányzatok alatt, illetve a különböző sajtótermékekben. [DiMaggio et al., 2013]

Az utolsó tanulmány, amit szeretnék a topik modellek társadalomtudományi alkalmazásai között megemlíteni, szintén egy az Egyesült Államokban folyt kutatást ír le. A vizsgálat során egy palesztin illetve izraeli szerzők szövegeiből, valamint egy demokrata és republikánus ideológiájú szövegekből álló korpusz elemeztek. A kutatáshoz az LDA egy kiterjesztését, az általuk joint topic and perspective modelként nevezett modellt használták. A topik modellel egyrészt az első korpuszban kirajzolódó izraeli-palesztin konfliktust, másrészt a demokrata és republikánus szerzők ideológiai ellentéteit elemezték. [Lin et al., 2008]

A topik modellek társadalomtudományi és egyéb humán tudományi alkalmazásai előtt tehát még nem nyílt meg teljesen az út és ma sem bővelkedünk a topik modelleket alkalmazó interdiszciplináris kutatásokban. Ramage, Rosen, Manning és Mc-

Farland két akadályt határoztak meg, amelyek a hasonló együttműködések útjában állnak. Az egyik akadályt az elérhetőségben látják, abban, hogy a topik modellezéshez használt technikai eszközöket nehéz különösebb hozzáértés nélkül használni. A másik akadályt pedig abban érzékelik, hogy a társadalomtudósok nehezen bíznak meg a topik modellek adta eredményekben anélkül, hogy végigolvasták volna a szövegeket. [Ramage et al., 2009] Azonban az alfejezetben vázolt alkalmazások szép példái annak, hogy sokrétű és izgalmas kutatási lehetőségek rejlenek a topik modellek társadalomtudományi célú felhasználásában, ezért mindenképpen érdemes az ez irányú kutatásokat folytatni a jövőben.

A következő fejezetben a látens Dirichlet allokáció egy társadalomtudományi célú felhasználását mutatom be. Az elemzés során a látens Dirichlet allokációt diskurzuselemzési céllal használom a kuruc.info online hírportál romaellenes cikkein. A célom a romaellenes cikkek tematikus struktúrájának automatikus feltárása, amely a társadalomtudós diskurzuselemzési munkáját segítheti.

3. A KURUC.INFO CIGÁNYBŰNÖZÉS ROVATÁNAK TEMATIKUS ELEMZÉSE

3.1. A korpusz bemutatása

A magyar nemzeti radikális, szélsőjobboldali média romaellenes megnyilvánulásainak vizsgálatához a kuruc.info online hírportál Cigánybűnözés rovatát választottam ki. A rovat cikkeiből látens Dirichlet allokáció segítségével azonosítom be a kuruc.info-n található legjellemzőbb romareprezentációs témákat.

A *kuruc.info* magát konzervatív, jobboldali, nemzeti, politikai párttól független tényfeltáró hírportálnak tartja[1], és a legolvasottabb jobboldali hírportálként hirdeti magát. A hírportálon szereplő cikkek politikai értékrendje azonban inkább a szélsőjobboldali, nemzeti radikális címkével illethetők. Választásom azért esett a kuruc.info-ra, ugyanis cikkeikben az etnikai kisebbségek, elsősorban a romák tematizálódnak legnagyobb gyakorisággal. [Glózer, 2013] Emellett a korpusz összeállítását az oldal szerzői maguk végezték, ugyanis külön rovatot vezetnek *Cigánybűnözés* névvel, amelyben a romákkal kapcsolatos híreket gyűjtik.[2] Ez megkönnyíti a korpuszgyűjtési munkát és elkerülhetők a kutatói félrekategorizálások.

A „cigánybűnözés" korpusz 2015. február 22-én lett letöltve. Ebben az időpontban a kuruc.info Cigánybűnözés rovata 11,805 darab cikket tartalmazott. Ebből a 11,805 cikkből a 10,453 darab cikk került begyűjtésre a Függelék D.1, D.2, D.3, D.4 és G.1 részében található kódok segítségével. A kódok bemutatására az adatgyűjtésről szóló alfejezetben kerül sor.

A szövegek alapvető statisztikáihoz tartozó kódokat a Függelék D.10 és E.1 alfejezetei tartalmazzák. A D.10 kód megnyitja a nyers és a normalizált, szótövezett, stopszavazott, gyakoriság szerint leszűrt szövegeket tartalmazó könyvtárakat és kiszűri azokat a dokumentumokat, amelyekhez nem tartozott dátum.[3] Így összesen 10,304 db cikk maradt. Ezután a kód dokumentumszinten normalizálja a nyers szö-

[1] https://kuruc.info/r/40/27066/

[2] A kuruc.info további rovatai az Antimagyarizmus, a Holokamu, Humor, Olvasói Levelek, Politikusbűnözés, Videók és Zsidóbűnözés nevet viselik.

[3] A későbbi elemzések miatt az időbélyeggel nem rendelkező cikkeket kizártam a vizsgálatból.

vegeket és megszámolja a karaktereket, a tokeneket[4], az egyedi szavakat és a mondatokat. Majd az elemzéshez használt normalizált, szótövezett, stopszavazott, gyakoriság szerint leszűrt dokumentumokon is végigmegy és megszámolja a karaktereket, a tokeneket és az egyedi szavakat. Végül kiírja az egyes dokumentumokhoz kapott értékeket a textstat.tsv-be. A E.1 behívja ezt a textstat.tsv fájlt és kiszámolja a leíró szövegstatisztikákat.

A nyers korpusz 36,753,335, a normalizált korpusz 34,853,393, a tisztított, elemzéshez használt korpusz pedig 13,673,147 karakterből áll. A normalizált korpusz 4,772,766 tokenből, 3,052,765 egyedi szóból és 313,560 mondatból tevődik össze, a tisztított, elemzéshez használt korpusz 1,594,606 tokenből és 1,142,896 egyedi szóból. A kuruc.info Cigánybűnözés rovatának átlagos cikke 3,384 karaktert, 463.5 tokent, 296.4 egyedi szót és 30.45 mondatot tartalmaz. Az elemzéshez használt korpusz dokumentumaiban átlagosan 1,328 karakter, 154.8 token és 111 egyedi szó található.

3.2. Az elemzés során használt eszközök bemutatása

Az adatgyűjtési és adatfeldolgozási fázishoz a Python, az R és a magyarlanc, a látens Dirichlet allokációval való elemzéshez a MALLET, illetve az R eszközöket, a további elemzésekhez az R programozási nyelvet használtam.

Az *R* [R Core Team, 2014] egy nyílt forráskódú funkcionális programozási nyelv és környezet, amely kiválóan alkalmas statisztikai elemzések és grafikai megjelenítések kivitelezésére. Az R egy GNU projekt (http://www.gnu.org/), amely nagyon hasonlít a John Chambers és kollégái által kifejlesztett S nyelvhez. A programozási nyelv elérhető Windowson, MacOS-on és az Unix-szerű operációs rendszereken és szabadon letölthető az http://www.r-project.org/ honlapról. Az eszköz szabadon használható tudományos és üzleti célból. Az R alapvető statisztikai opcióit különböző *csomag*ok (package) letöltésével lehet bővíteni. Jelenleg nagyjából 6,500 csomag érhető el a CRAN-en (Comprehensive R Archive Network) keresztül (http://cran.at.r-project.org/). Az R használatához igénybe vehetünk különböző grafikai felhasználói felületeket (Graphical User Interface, GUI), például az RStudio-t, különböző szövegszerkesztőket, például az Emacs-ot vagy irányíthatjuk parancssorból. A legutóbbi R verzió, az R 3.1.3 2015. március 9-én jelent meg. A dolgozathoz az R 3.1.3 verzióját használtam, az R-ben írt kódok a Függelék E részében találhatók.[5]

[4] Tokennek hívjuk a szavak konkrét korpuszbeli előfordulását.[Tikk et al., 2007]

[5] A leírás az R honlapján található információk (http://www.r-project.org/) alapján készült.

A *Python* egy nyílt forráskódú objektum-orientált programozási nyelv, amelyet eredetileg Guido van Rossum fejlesztett ki az 1990-es évek elején az ABC nyelv utódjaként. A programnyelvet azóta is folyamatos fejlesztik. A Python filozófiája, hogy tiszta, könnyen olvasható, egyszerű kódolást biztosítson, amely ideálissá teszi gyors prototípusok fejlesztésére és ad-hoc programozási feladatokhoz. A Python az R programozási nyelvnél általánosabb feladatokhoz nyújt lehetőséget, például web szerverekhez való kapcsolódáshoz vagy fájlok egyszerű módosításához. A Python elérhető Windowson, MacOS-on és az Unix-szerű operációs rendszereken és szabadon letölthető az https://www.python.org/ honlapról. A programnyelv alapvető lehetőségeit *modulok* letöltésével lehet kiegészíteni. Az eszköz szabadon használható tudományos és piaci célból egyaránt. A Python használatához igénybe vehetünk különböző grafikai felhasználói felületeket, például az Spyder-t, különböző szövegszerkesztőket, például az Emacs-ot vagy irányíthatjuk parancssorból. A legutóbbi verzió a Python 3.4.3 2015. február 25-én jelent meg, azonban sokan a Python 2.7 széria legutóbbi változatát, a 2014. december 10-i megjelenésű Python 2.7.9 verziót használják, ugyanis a modulok ezen a szérián érhetők el teljeskörűen. A szakdolgozathoz a Python 2.7.9 verzióját használtam, a Python-ban írt kódok a Függelék D részében találhatók. [6]

A *magyarlanc* [Zsibrita et al., 2013] egy a magyar nyelvre kifejlesztett Java-alapú morfológiai és szintaktikai elemző, amely a nyelvi elemzésen kívül lehetőséget nyújt a szövegek szegmentálására és a stopszószűrésére is. A magyarlancot a Szegedi Tudományegyetem Nyelvtechnológiai Csoportja fejlesztette, amelyhez a morphadoner, a Stanford POS-tagger és Bohner parser módosított változatait használták fel. A magyarlanc szabadon letölthető a Szegedi Tudományegyetem Nyelvtechnológiai Csoportjának honlapjáról és szabad felhasználást biztosítanak oktatási és kutatási célra. (http://www.inf.u-szeged.hu/rgai/nlp?lang=hu&page=magyarlanc) Az eszköz üzleti célú felhasználása azonban egyéb feltételekhez kötött. A magyarlanc könyvtárat parancssorból tudjuk használni, amelynek rövid leírása szintén a Szegedi Tudományegyetem Nyelvtechnológiai Csoportjának honlapján található. A magyarlanc szakdolgozatban való használatát a Függelék D.5 részében található Python kód tartalmazza. [7]

A *MALLET* [McCallum, 2002] egy nyílt forráskódú Java-alapú könyvtár, amely természetes nyelvi szövegek statisztikai feldolgozásához ad lehetőséget. A MALLET

[6] A leírás a https://www.python.org/, a https://wiki.python.org/moin/BeginnersGuide/Overview és [Rossum, 1995] források alapján készült.

[7] A leírás a http://www.inf.u-szeged.hu/rgai/nlp?lang=hu&page=magyarlanc forrás alapján készült.

könyvtárat Andrew McCallum írta diákjának és kollégájának, például David Mimno közreműködésével, a Massachusetti Amherst Egyetemről és a Pennsylvaniai Egyetemről. A MALLET megfelelő eszközt nyújt dokumentumok osztályozásához (naiv Bayes, maximum entrópia és döntési fa algoritmusok), szekvencia taggeléshez (rejtett Markov-modellek, maximum entrópia Markov-modellek, feltételes random mezők), topik modellekhez (látens Dirichlet allokáció, Pachinko allokáció és hierarchikus LDA) és grafikus modellekhez. A MALLET elérhető Windowson, MacOS-on és az Unix-szerű operációs rendszereken és szabadon letölthető az eszköz honlapjáról (http://mallet.cs.umass.edu/download.php). A könyvtár szabadon használható tudományos és üzleti célból egyaránt. A programot parancssorból lehet működtetni, amelyhez a szerzők részletes oktatói anyagot biztosítanak az eszköz honlapján. A kutatáshoz a MALLET könyvtár 2.0.7 verzióját használtam, a kódok a Függelék F részében találhatók. A MALLET-nek létezik egy R programnyelvre írt wrapperje, a *mallet* csomag (http://cran.r-project.org/web/packages/mallet/index.html). [8]

3.3. Adatgyűjtési és adatfeldolgozási folyamatok bemutatása

A jelenlegi alfejezetben a kutatás adatgyűjtési és adatfeldolgozási folyamatainak részleteit mutatom be, azaz a korpusz letöltésének és szervezésének, illetve a szöveg feldolgozásának és tisztításának lépéseit. Az adatgyűjtéshez tartozó kódokat a Függelék D.1-D.4 és a G.1 alfejezetek tartalmazzák, az adatfeldolgozáshoz pedig a D.5-D.9, illetve az E.2-E.3 kódok tartoznak. A következőkben ezek leírására kerül sor.

3.3.1. Adatgyűjtés

A Függelék D.1 részéhez tartozó kód feladata, hogy legenerálja a kuruc.info Cigánybűnözés rovatához tartozó összes linket. A cikkekhez tartozó linkeknek nincsen külön mintázata, ezért a Cigánybűnözés rovat cikkeit tizesével tartalmazó főoldalakról kellett a linkeket kinyerni, amelyek a „https.//kuruc.info/to/35/10” URL címmel kezdődnek és a „https.//kuruc.info/to/35/1810” URL címig tartanak. Az URL cím utolsó tagja 10-től 1810-ig tizesével lépeget, amiket kilistázunk az urls nevű listába. A kód a codecs modult, az urlparse modulból az urljoin-t, a boilerpipe.extract modulból az Extractor-t és a bs4 modulból a BeautifulSoup-ot hívja be. A Boilerpipe egy Java könyvtár, amely HTML oldalak főszövegét tisztítja meg és azok adatait

[8] A leírás a http://mallet.cs.umass.edu/index.php és a http://mallet.cs.umass.edu/about.php források alapján készült.

nyeri ki, az ehhez készült wrappert hívja be a kód. Ennek segítségével megtisztítjuk az urls listán lévő URL kódok HTML szövegét a felesleges elemektől és betöltjük mindet az extracted_html-be. A Beautiful Soup egy Python könyvtár, amely HTML és XML fájlokból nyeri ki az adatokat. A BeautifulSoup parancs használatával az extracted_html-ből kinyerjük a rovat főoldalainak HTML megtisztított szövegét, majd kiszedjük belőlük a linkeket, amik nagy valószínűséggel a Cigánybűnözés rovathoz tartozó cikkek linkjei. Ezeket a linkeket betöltjük az article_urls listába és a codecs modul használatával kiírjuk a cikkek URL címeit a rovat.tsv-be.

A Függelék D.2 részében található kód megnyitja a rovat.tsv-ben lévő URL címeket, beolvassa őket és lementi a HTML szövegeket. Sajnos a szövegek karakterkódolása nem volt egységes, sokszor egy szövegen belül is keveredett az UTF-8 és az ISO-8859-2 karakterkódolás. Ezért a G.1 függelékrészben található parancsok segítségével át kellett kódolni a fájlokat, gyakorlatilag el kellett rontani a fájlok karakterkódolását, hogy ki lehessen nyerni belőlük a megfelelő mintázatok alapján a Cigánybűnözés rovatba tartozó cikkeket. A mintázatok a G.1 alfejezetben találhatók a grep -i1R parancs után, amelyek azt jelezték egy HTML szövegben, hogy a cikk a Cigánybűnözés rovat része.

A Függelék D.3 alatt lévő kód beolvassa a cb mappából az átkonvertált, Cigánybűnözés rovatba tartozó cikkeket tartalmazó HTML szövegek fájlneveit és beteszi az articles.tsv-be. Ezután az articles.tsv elemeit betesszük egy listába, beolvassuk az eredeti kódolású, de nem csak Cigánybűnözés rovatba tartozó cikkeket tartalmazó HTML szövegek fájlneveit a html mappából. Ha egy a html mappában lévő fájl neve megegyezik a cb mappában lévő egy másik fájl nevével, tehát a Cigánybűnözés rovat egy cikke, akkor átmásoljuk a html mappa elemét a hate_html mappába. Így megkapjuk az eredeti kódolású, csak Cigánybűnözés rovatba tartozó cikkek HTML szövegeit, amik a hate_html mappába kerülnek.

A Függelék D.4 alfejezetében található kód beolvassa a hate_html mappából a HTML szövegeket tartalmazó fájlok neveit, megnyitja a fájlokat, beolvassa a tartalmukat a txt objektumba, majd az ISO-8859-2 kódolású részeket dekódolja és átkódolja UTF-8 karakterkódolásra. Ezután a txt objektumokból kinyeri az egyszerű szövegeket (plain text), és kiírja egy UTF-8 kódolású txt kiterjesztésű fájlként a hate_text mappába a hate_html mappában lévő fájlnévvel.

Ezzel az adatgyűjtés végére értünk, megkaptuk a 10,453 dokumentumból álló „cigánybűnözés” korpuszt. A korpusz tisztításához és az elemzéshez alkalmas formára hozásához azonban még szükség van néhány lépésre. Ezeket a lépéseket a Függelék D.5-D.9, illetve E.2-E.3 része tartalmazza.

3.3.2. Adatfeldolgozás és adattisztítás

A D.5 alfejezet a magyarlanc Python-ban való meghívását foglalja magában. A magyarlanc a hate_text mappában található szövegfájlokon megy végig. A fájlok neve 1-10477-ig tart, ha létezik ilyen nevű txt kiterjesztésű fájl a hate_text mappában, akkor a magyarlanc morfológiai elemzője elvégzi rajta a nyelvi elemzést és kiírja a processed nevű mappába az eredményt.

A Függelék D.6 alatt lévő kód megkeresi a hate_text mappában található szövegfájlokban a cikkekhez tartozó dátumokat, egységes formára hozza őket és kiírja a fájlok nevét és a hozzájuk tartozó dátumokat a code_date.tsv fájlba. Erre a lépésre azért van szükség, mert a későbbi elemzés során a dokumentumok témáinak időbeli változásait is elemzem, ezért azokat a dokumentumokat, amelyekben nem található időbélyeg, kizártam a további lépésekből és az elemzésből. Így 10,304 dokumentum maradt a 10,453-ból.

A Függelék D.8 részében található fájl megnyitja és beolvassa a magyarlanc által feldolgozott szövegeket, miközben a stoplist nevű listába kilistázza a stoplist.tsv-ből a stopszavakat[9] és a freq_delete.tsv-ből a korpuszban legfeljebb ötször előforduló szavakat. A stoplist.tsv tartalma a Függelék H alfejezete alatt található meg. Stoplistára főleg a magyar nyelvre jellemző tartalom nélküli funkciószavak kerültek, például az „az", az „ő", a „hogy", az „előtte", vagy a „van", valamint a cigánybűnözés korpuszra jellemző, az adott cikk szempontjából kevés szemantikai tartalommal bíró szavak. Utóbbiakra hozhatjuk fel példának a cikk műfajából fakadó szavakat (pl. „cikk", „január", „olvasó"), a kuruc.info egységeihez köthető szavakat (pl. „kuruc", „info", „külföld", „zsidóbűnözés", „friss"), személyek személynevei (pl. „péter", „ferenc", „marian"), a földrajzi neveket, amelyek ha megjelennek egy cikkben, akkor a melléknévi alakjaik is megjelennek (pl. „olaszország", „románia", „finnország") vagy a gyakori eseményleírásokból fakadó kevés jelentéssel bíró szavakat (pl. „férfi", „fiú", „éves"). A korpuszban legfeljebb ötször előforduló szavakat az E.2 kód segítségével nyertem ki, amely a már megtisztított korpusz szó-dokumentum mátrixából kiszedi a kívánt szavakat és kiírja a freq_delete.csv fájlba. Nagyjából 59,000 szót nyert ki a kód.[10] [11] Visszatérve a stem_filter.py fájlhoz, a kód a nyelvi elemzésen átesett szövegekből kinyeri a szótövezett alakokat, kisbetűsre alakítja őket, és a morfológiai

[9] Stopszavaknak hívjuk a tartalmi információt nem hordozó szavakat, amelyeket eltávolítunk a szövegekből [Tikk et al., 2007].

[10] Ezt a kódot a stem_filter.py első futtatása után futtattam le, hogy a már tisztított alakok alapján nyerje ki az alacsony gyakoriságú szavakat.

[11] A szavak listáját nem tudtam átmásolni a szövegszerkesztőbe a mérete miatt.

elemzés során a főnév, a melléknév vagy az ismeretlen szófaji kategóriába[12] kerülő elemeket megtartja és kiszűri belőle a stoplist listán lévő elemeket. A leírt folyamat alapján megmaradt elemeket dokumentumonként kiírja a stemmed_filtered nevű mappába. Az így kapott szótövezett, kisbetűsre alakított, valamint szófaj, jelentés és gyakoriság alapján szűrt elemeket tartalmazó dokumentumok kerülnek elemzésre.

A szövegek tisztításának azonban itt sem szakadt vége. Az első topik modellek eredményei ugyanis tartalmaztak néhány olyan topikot, amelyek különszedték a nem magyar nyelvű szavakat és azt egy vagy több különálló topikban helyezték el. Ez azért eshetett meg, mert a szófaji szűrésnél az ismeretlen szófajú elemeket is megtartottam és minden nem magyar nyelvű szó ebbe a kategóriába került a nyelvi elemzés során. Azonban semmiképp nem akartam elhagyni ezt a kategóriát, ugyanis a cigánybűnözés korpuszban használt sztenderd magyar nyelvtől eltérő kifejezések (pl. „cigányfajvédő", „ork", „cigánybűnöző") mind ebbe a kategóriába estek. Először felvettem stopszólistára a gyakori angol kifejezéseket, azonban ezzel a módszerrel nem sikerült eléggé megtisztítanom a szövegeket, és a modellillesztés során továbbra is kialakultak idegen nyelvű topikok. Ezért a Függelék E.3 részében található kód a Gibbs-mintavétel végső állapotát tartalmazó MALLET output alapján kiválogatta azokat a dokumentumokat, amelyekben több, mint húsz olyan szó fordul elő, amelyeket a nem magyar nyelvű topikokhoz rendelt az algoritmus. Az első látens Dirichlet allokáció modell szerint három olyan topik alakult ki, amely nem magyar nyelvű szavakat tartalmazott.[13] A beazonosított dokumentumokban a nem magyar nyelvű szövegek a magyar nyelvű szövegekkel keveredtek, ezért kézzel kellett kitörölnöm a nem magyar elemeket a dokumentumokból. Az első tisztítás után is létrejött egy nem magyar nyelvű topik, ezért az újabb tisztítás során azokat a dokumentumokat tisztítottam meg, amelyekben több mint tíz olyan szó fordult elő, amelyeket a nem magyar nyelvű topikhoz rendelt az algoritmus. Az eredeti, nem magyar nyelvű részeket tartalmazó dokumentumokat egy külön mappába helyeztem. Ezek alapján a D.7 kód kiszedte a fájlok azonosítóját. A kóddal kapott, nem magyar nyelvű szövegrészektől megtisztított szövegek listáját a Függelék I fejezete tartalmazza. A tisztítás során három fájlt töröltem teljesen, ugyanis ezek nem tartalmaztak magyar nyelvű részeket. Az így megtisztított dokumentumok topik modellje már nem tartalmazott

[12] A magyarlanc a feldolgozott dokumentumok 3. oszlopába írja ki a beazonosított szófajokat. Ezek közül csak a főnév (Noun), melléknév (Adjective) és egyéb (X) szófajú elemeket tartottam meg, tehát a kódban a többi szófaj kezdőbetűje szerepel.

[13] Az 1-es topik angol, a 10-es topik finn és német, a 20-as topik egyéb nyelvű szavakat tartalmazott. A számok természetesen változnak a Gibbs-mintavétel miatt, ezért a végső modellekhez pszeudórandom számgenerátort használtam.

nem magyar nyelvű szavakat, azonban előjött még egy probléma.

A nem magyar nyelvű szövegrészektől való tisztítás után újraillesztett látens Dirichlet allokációval kialakult egy olyan topik, amelyben a szavak az „ő" karakterek helyett „õ", az „ű" karakterek helyett „û" karaktereket tartalmaztak (pl. „cigánybûnözõ", „rendõr", „õk"). Ezt a karakterkódolási hibát a Függelékben található D.9 kóddal sikerült megoldani, amely kicseréli a rossz karaktereket a helyes karakterekre.[14] Ez után a tisztítási lépés után már készen állt a korpusz a látens Dirichlet allokáció illesztéséhez.

Az utolsó két lépés az adattisztítás során váratlanul ért, azonban két tanulságként is levonható. Az egyik tanulság, hogy a topik modellek alkalmasnak bizonyulnak eltérő nyelvű szövegek szétválasztására, illetve a másik tanulság, hogy olyan anomáliákat is detektálni tudnak, mint a karakterkódolási hiba.

3.4. A látens Dirichlet allokáció illesztése

A szövegek begyűjtése és feldolgozása után a korpusz tematikus elemzése következett. A tematikus struktúra kinyeréséhez a 2. fejezetben bemutatott látens Dirichlet allokáció (LDA) nevű topik modellt használtam. A látens Dirichlet allokáció illesztéséhez több eszköz is rendelkezésünkre áll, a legnépszerűbb programozási nyelvek és adatelemzéshez használt programozási nyelvek (pl. Java, C, C++, Python, Matlab, R) nagy része lehetőséget biztosít a látens Dirichlet allokáció illesztéséhez. Ezek közül az eszközök közül a MALLET Java könyvtárat, az R három csomagját, a topicmodels és az lda csomagokat, valamint a MALLET R wrapperjét, a mallet csomagot próbáltam ki.

A kipróbált eszközök közül a MALLET Java-alapú könyvtár bizonyult a legkönnyebben kezelhető és leggyorsabb futási idejű megoldásnak. A korpuszbeolvasás és a modellillesztés csupán néhány egyszerű parancs használatát igényli és a modellillesztés nagyjából hat percet vett igénybe, amely nyolcszor gyorsabb futást jelentett pl. a topicmodels R csomag modellillesztéséhez képest. A MALLET gyorsasága a fastLDA nevű algoritmusnak köszönhető, amely a Gibbs-algoritmus egy változata. A fastLDA algoritmusról a [Porteous et al., 2008] tanulmányban olvashatunk bővebben. Emellett a könyvtár lehetőséget biztosít aszimmetrikus α paraméter illesztéséhez és a különböző outputok kérése is egyszerűen kivitelezhető. A többi lehetőséggel kapcsolatos tapasztalataim a Függelék C részében olvashatók.

[14] Mivel ezt a hibát csak a topik modellezés során vettem észre, a szótövezett, szűrt korpuszon futtattam le a kódot, de célszerű lett volna egy előbbi fázisban végrehajtani.

A következő alfejezetekben a topikok számának kiválasztását, majd a MALLET használatát szeretném bemutatni a „cigánybűnözés" korpuszon.

3.4.1. Topikok számának kiválasztása

A topikok számának kiválasztásához több módszert együttesen vettem igénybe. Egyrészt a magyar romareprezentációs kutatások irodalma, valamint saját kutatói döntésem alapján a topikok számát minimum 15, maximum 30 topikban szabtam meg. A minimum érték megszabásában Bernáth és Messing 1998-as, 2002-as és 2011-es kutatására hagyatkoztam, akik 15 nagyobb tematikus kategóriát alakítottak ki. Hipotézisem szerint a „cigánybűnözés" korpuszban a bűnözés téma differenciáltabban jelenik meg, és az illesztett topikmodellek tapasztalatai alapján úgy találtam, hogy a 25-30 közötti topikszám lenne ideális a „cigánybűnözés" korpusz tematikus leírásához.[15] A tényleges topikszám kiválasztásánál egyrészt a 2.4.5 alfejezetben leírt harmonikus átlag módszert használtam, amely kiszámolásához a Függelék E.6 kódját alkalmaztam. Másrészt mivel ezzel a kóddal csak szimmetrikus α paraméter illesztéséhez tudtam harmonikus átlagot számolni, az aszimmetrikus α illesztéshez a MALLET kicsit eltérő, de hasonló mértékét vettem figyelembe. A MALLET a Gibbs-mintavételi állapotok loglikelihoodjainak mértani átlagát számolja ki, amit eloszt a tokenek számával.[16] A szimmetrikus α paraméterrel illesztett látens Dirichlet allokáció esetében 30, az aszimmetrikus α paraméterrel illesztett látens Dirichlet allokáció esetében 27 topik illesztését találtam megfelelőnek ezzel a vegyes megoldással.

3.4.2. A látens Dirichlet allokáció illesztése MALLET-tel

A MALLETT-tel összesen négy láncot illesztettem, egy LDA-t szimmetrikus α paraméterrel, és három LDA-t aszimmetrikus α paraméterrel. Az aszimmetrikus α paraméterrel illesztett modellhez a három láncot 5000 iterációval futtattam le eltérő kezdőparaméterekkel, a szimmetrikus α paraméterrel illesztettet szintén 5000 iterációval. Az elemzés során az aszimmetrikus α paraméterrel illesztett modelleket fogom figyelembe venni, mivel a szakirodalom szerint így robusztusabb modellhez jutunk [Wallach et al., 2009a], másrészt azért, mert feltételezem, hogy egy ilyen több évet felölelő korpusz olyan témákat is tartalmazhat, amik rövidebb időszakon

[15] Nem állítom, hogy kevesebb vagy akár jóval több topik nem lenne indokolt, azonban a jelenlegi elemzés kereteihez ezt a topikszámot találtam optimálisnak.

[16] A kipróbált példákon az szimmetrikus esettel összemérve ugyanazt a végeredményt adta a harmonikus és a mértani átlag. Ez persze nem jelenti, hogy minden esetben ugyanazt az eredményt adnák.

keresztül koncentráltabban jelentek meg a cikkekben. Az aszimmetrikus esetben az α paramétervektor alacsonyabb értékei az ilyen témákat jelenítik meg. Azonban kíváncsi voltam, hogy mennyire ad más eredményt ezen a feladaton, ha szimmetrikus α paraméterrel illesztem az LDA-t, ezért futtattam egy láncot szimmetrikus α paraméterrel is.

A MALLET-tel illesztett modellek kódja az F fejezetben található. Az F.1 kód behívja a szótövezett, megtisztított dokumentumokat és átalakítja a korpuszt olyan formára, amivel az LDA modellt illeszteni tudjuk. A folyamatnak ennél a lépésenél is kiszűrhetjük a stopszavakat a stoplist-file kapcsolóval.

Az F.2 és az F.4 alfejezet alatt olvasható kódok az F.1 kódban kapott mallet fájlt hívja be és a beállított paraméterekkel illeszt egy LDA modellt. A num-topics kapcsolóval a topikok számát, a num-iteration kapcsolóval az iterációk/lépések számát, az optimize-interval kapcsolóval azt szabjuk meg, hogy hány lépésenként optimalizálja a hiperparamétereket az algoritmus. A num-top-words kapcsolóval azt adjuk meg, hogy hány kulcsszót írjon ki topikonként a folyamat, a random-seed kapcsolóval a pszeudo-random számgenerátort állítjuk be egy bizonyos értékre, hogy a folyamat megismételhető legyen és a use-symmetric-alpha kapcsolóval azt adjuk meg a MALLET-nek, hogy szimmetrikus vagy aszimmetrikus α hiperparaméterrel illessze a modellt.

Az F.3 és az F.5 függelékrészben lévő kódokkal a konkrét modelleket illesztettem. Az előző kódokhoz képest új opciók a különböző output fájlokat kérő kapcsolók, mint a word-topic-counts-file, ami megszámolja, hogy a korpusz szavai melyik topikban hányszor szerepeltek, az output-state, ami a végső Gibbs-mintavétel állapotot összegzi, az output-state-interval, amely a Gibbs-mintavétel állapotait kéri le a megadott lépészámonként, az output-doc-topics, amely a dokumentumok topikeloszlását adja meg, az output-topic-keys, ami a topikok legnagyobb valószínűségű szavait összegzi[17], valamint a diagnostics-file, amely különböző modell ellenőrzésére szolgáló mértékeket számolja ki[18]. Az előbb felsorolt kapcsolók után a fájl nevét és kiterjesztését kell megadni, ahova az adott outputot írja ki a MALLET. A doc-topic-threshold kapcsoló azt határozza meg, hogy a doc-topics outputba kerülő topikoknak mi legyen a minimum valószínűsége, az alpha kapcsoló az α hiperparaméter kezdőértéket szabja meg, a beta kapcsoló a β hiperparaméter kezdő értéket. Előbbi esetében 5.0, utóbbi esetében 0.01 a default érték.

[17] Amennyi kulcsszót megadtunk a num-topics kapcsolónál.

[18] Ez a funkció még béta verzióban van, az output részeiről nem található leírás, ezért sajnos használni sem tudtam.

A következő fejezetben az illesztett modellek konvergenciáját ellenőrzöm, illetve a szavak dokumentumoktól való feltételes függetlenségét a topikhozzárendelésre mint feltételre nézve.

3.5. A konvergencia és az illeszkedés ellenőrzése

A fejezet során a második fejezetben leírtak szerint ellenőrzöm a Gibbs-mintavétel konvergenciáját, valamint a látens Dirichlet allokáció modell azon feltételét, miszerint a szavak feltételesen függetlenek a dokumentumoktól a topikhozzárendelésre mint feltételre nézve.

3.5.1. A Gibbs-mintavétel konvergenciájának ellenőrzése

Az aszimmetrikus α paraméterrel illesztett modellhez három láncot futtattam le 5000 iterációval, eltérő kezdőparaméterekkel, a szimmetrikus α paraméterrel illesztett modellhez egy láncot 5000 iterációval. A paramétereket mindkét modellnél először a default, ajánlott kezdőértékre állítottam be. Az aszimmetrikus modellnél a második lánc esetében ezeknél magasabb, a harmadik lánc esetében alacsonyabb értékeket választottam. A MALLET a lánc állapotait számomra nem megfelelő formátumban adja ki, ezért a fájlokat parancssorból a G.2 függelékrészben lévő paranccsal hoztam megfelelő formára. A paramétereket a D.11 és D.12 Függelékben lévő kódokkal nyertem ki utóbbi fájlokból, amely meghívja a fájlok első három sorát, kiszedi az α paraméter(eke)t a második sorból, a β paramétert a harmadik sorból, végül elmenti mindet egy csv fájlba. Az E.8 függelékrészben található kód behívja utóbbi fájlt és ábrázolja a paraméterek tíz lépésenkénti egymás utáni értékeit, valamint az értékek hisztogramját. Az ábrák a Függelék J fejezetében találhatók. Az eredmények alapján a láncok konvergenciája az esetek többségében megállapítható. Az aszimmetrikus α paraméterrel illesztett látens Dirichlet allokáció első láncában azonban találunk olyan topikokat (4., 7. 10., 19., 24.), amelyek szóeloszlása nem keveredik az elvártak alapján. A harmadik lánc esetében is vannak a hisztogram alapján kétes topikok (1., 8., 9., 21. topik), de 2000-3000 iteráció után a lánc egy helyre látszik konvergálni mind a négy topik esetén. A második lánc konvergenciája az alkalmazott elemzések tükrében megfelelőnek tűnik, valamint a szimmetrikus α paraméterrel illesztett LDA modellé is. Az elemzésnél a konvergenciát figyelembe véve a stabil topikokra fogok koncentrálni.

3.5.2. Posterior prediktív ellenőrzés

A 2.4.5 alfejezetben leírtak alapján a látens Dirichlet allokáció modellfeltételeinek ellenőrzésére még nem áll megfelelő eszköztár a rendelkezésünkre, ezért a [Mimno and Blei, 2011] által leírt posterior prediktív ellenőrzést próbáltam ki. Ez a szimulációs teszt nincsen az általam ismert eszközökben implementálva, ezért saját kódot írtam a módszerhez, amely az E.9 függelékrészben található. A kód a Függelék B részében ismertetett [Mimno and Blei, 2011] tanulmány alapján íródott, a szimulációk számában és az ábrázolás módjában a [Gelman and Meng, 2005] tanulmányhoz alkalmazkodtam. A posterior prediktív ellenőrzés során a látens Dirichlet allokáció azon feltételes függetlenségi feltételét ellenőriztem, miszerint egy $w_{d,n}$ szó független d dokumentum θ_d topikeloszlásától a $z_{d,n}$ topik hozzárendelésre, mint feltételre nézve. Mimno és Blei ehhez egy olyan diszkrepanciafüggvényt terveztek, amely a feltételes függetlenséget a W szavak és a D dokumentumindexek közötti feltételes kölcsönös információval méri k topikra, mint feltételre nézve. Az ellenőrzés során új szavakat sorsolunk véletlenszerűen minden egyes tokenre a topikok posterior eloszlásából, és újraszámoljuk a rangját és a kölcsönös információját minden egyes szónak. Az ezekből kirajzolódó intervallum egy referenciaeloszlást képez, ami a kölcsönös információ várható értékét adja minden rangra a polinomiális eloszlás feltétele mellett. A posterior prediktív tesztek eredményei a Függelék K fejezetében találhatók. Mimno és Blei azt tapasztalta, hogy a szemantikailag tartalmas szavakat tartalmazó topikok szavai és dokumentumindexei közötti kölcsönös információ valamivel magasabb, mint amit a polinomiális feltétel esetén várunk. Ugyanez figyelhető meg az aszimmetrikus α paraméterrel illesztett 1. lánchoz tartozó modell, valamint a szimmetrikus α paraméterrel illesztett modell esetében is. Az adatok tehát azt javasolják, hogy a feltétel nem tartható, azaz léteznek olyan dokumentumszintű információk, amelyek a dokumentumba kerülő szavakra hatással vannak. Ezen persze nem lepődünk meg, azonban továbbgondolkodásra késztet minket, hogy milyen dokumentumszintű információkkal bővíthetnénk a modellt, hogy az illeszkedésen javítsunk. Mimno és Blei is így tettek az említett tanulmányban és olyan dokumentumszintű metaadatokkal bővítették a modelljüket, mint a dokumentumokhoz tartozó dátum, a dokumentum rovata stb. Mivel a „cigánybűnözés” korpusz egy hosszabb időintervallumot átölelő korpusz, kézenfekvőnek tűnik a dokumentumokhoz tartozó dátumokkal bővíteni az illesztett látens Dirichlet allokáció modelleket. A következő alfejezetben erre kísérletet is teszek.

A következő alfejezetben a látens Dirichlet allokáció modellek topikjainak elemzem társadalomtudományi szempontból és a topikok időben való változásának vizs-

gálatára teszek próbát.

3.6. A kapott eredmények értelmezése

Az elemzés első lépésében meghatározom a topikokat a topikokhoz tartozó kulcsszavak és a dokumentumok topikeloszlásai alapján. A topikokat a kulcsszavak és a topikok legjellemzőbb dokumentumainak kvalitatív elemzésével értelmezem, tehát a kvantitatív elemzés a kutatásnak ezen a pontján kapcsolódik össze a kvalitatív elemzéssel. Az elemzés során bemutatom a feltárt témák társadalomtudományi használhatóságát és relevanciáját. Ezután megjelenítem a témák időbeliségét és pár példán bemutatom a módszer alkalmazhatóságát.

A topikok kulcsszavait a topic-keys.csv MALLET outputból kaptam meg, a topikok dokumentumhoz tartozását pedig a E.12 kód segítségével határoztam meg a doc-topics.csv MALLET outputból.[19] A kód behívja a doc-topics_1.csv fájlt és definiál egy olyan függvényt, amellyel a behívott fájlból kinyerhetők a topikonkénti dokumentum-topik hozzárendelések. A függvény meghívásával és a topikszám megadásával minden topikra lekértem a dokumentum-topik hozzárendeléseket és kiszámítottam a topikhozzárendelések arányát.

A modellek közül az aszimmetrikus α paraméterrel illesztett látens Dirichlet allokációs modell második láncát veszem alapul, ugyanis ennek a láncnak az illesztett eloszlásai tűntek a legstabilabbnak a konvergenciaellenőrzés során. A modellekhez tartozó kulcsszavak a Függelék L fejezetében olvashatók.

Hipotézisem szerint a rovatban a bűnözés topik lesz a mérvadó téma, amely a cikkekben a bűnözés kevésbé súlyos eseteitől kezdve a legsúlyosabbakig terjed. Ezek valószínűleg külön topikot fognak alkotni. Emellett feltételezem, hogy a nagy közfelháborodást kiváltó események, pl. a Cozma-gyilkosság vagy az olaszliszkai lincselés külön topikban fognak megjelenni. A topikoktól emellett azt várom, hogy a Bernáth és Messing romareprezentációs kutatásai során kialakított témastruktúrának nagyjából megfeleltethető topikstruktúrát fogok kapni. Utóbbi hipotézisemet a következő alfejezetben fogom vizsgálni.

Az L.3 táblázat alapján az első[20] topikhoz tartozó cikkek olyan esetekről számolnak be, amelyben európai, az Európai Unióhoz tartozó országok roma csoportok „kitoloncolását", „deportálását", „cigánytelepek felszámolását" kezdeményezték. Mindezt a kuruc.info-n helyes példaként mutatják be, az érintett országok bezzeg-

[19] A doc-topics.csv MALLET outputban akkor tartozik egy topik egy dokumentumhoz, ha aránya meghaladja a 10%-ot. Ennél természetesen lehetne szigorúbb küszöbértéket is szabni.

[20] A MALLET nullától számoz.

országokként jelennek meg. A roma probléma a cikkek szerint minden országban felüti a fejét, mert a cigányok „nem akarnak beilleszkedni a társadalomba", ezért nemkívánatos csoportnak számítanak Európán belül. Az EU-n belüli szabad közlekedés alkalmat kínál a romák ide-oda toloncolásának, a „bűnözők elsózásának", amit a „tökösebb" miniszterelnökök ki is használnak. Kiemelt helyet kap a cikkekben a bolgár és román romák Franciaországból való kitoloncolása, a franciaországi „illegális cigánytáborok" felszámolása, a szlovákok által véghezvitt szociális „segílykék" megvonása és a „hadsereg romák elleni bevetése". A topik a többi illesztett modellben is megjelenik. A téma a cikkek 6.4%-ban jelent meg.

A második topikhoz olyan cikkek tartoznak, amelyek főleg a Magyar Gárda, valamint a Jobbik, a Magyar Önvédelmi Mozgalom, a Nemzeti Őrsereg vagy a Magyar Nemzeti Front által szervezett felvonulásokról, megemlékezésekről, fórumokról, demonstrációkról számolnak be, vagy felhívást végeznek az eseményekre. Kitüntetett eseményként jelenik meg az olaszliszkai lincselés Szögi Lajos halálával, akinek „mártírhalála mérföldkő volt a magyarság öntudatra ébredésében". A rendezvények kiélezett témája a romák bűnöző életmódja és „a vidéki lakosságot fenyegetõ közbiztonsági problémák". Akik a rendezvényeket megakadályozzák, „bűnpártolónak" számítanak, a rendőrséget emiatt „ÁVH" címkével illetik. A cikkekben visszatérő téma a cigányok félelemkeltése, és az arra válaszul adott „mi nem félünk!" kijelentés. A szervezett rendezvényeket békés rendezvényekként jelenítik meg és tagadják a félelemkeltés vádját. A topik a korpusz 13%-ban jelent meg.

A harmadik topikhoz tartozó cikkek a roma önkormányzat és egyéb civil érdekvédelmi képviseletek eseményeiről számolnak be. Az egyik kiemelt eseménynek Kolompár Orbán, „fő cigánybűnöző" és társai ellen emelt vádak és büntetőeljárások számítanak.[21] A vádemelések és büntetőeljárások sorozata 2008-2013-ig tartott, amelyről a kuruc.info mindvégig beszámolt. A másik nagyobb mértékben megjelenő esemény az OCÖ új elnökének, Farkas Flóriánnak, a Lungo Drom vezetőjének megválasztása volt. A kuruc.info-n az elnökváltást „bűnözőcsereként" aposztrofálták. A topikot a cikkek 6.9%-a tartalmazta.

A negyedik topik lopásos bűncselekményekről számol be és az azokkal kapcsolatos rendőri eljárásokról. A hírekben általában nem jelenik meg expliciten, hogy roma származású az elkövető, csak sugalmazzák, például más forrásból vett idézőjelbe tett utalásokkal („fiatal", „fiú") vagy egyéb szófordulatokkal (pl. „Hiába takarták

[21] Kolompár Orbán az Országos Cigány Önkormányzat (OCÖ) volt elnöke, aki ellen jogosulatlan gazdasági előny megszerzésének bűntette és az Európai Közösségek pénzügyi érdekeinek megsértése miatt emeltek vádat.

ki K. István fejét"). A hírek Cigánybűnözés rovatban való szerepeltetése azonban egyértelművé teszik a szövegbeli célzásokat. A cikkek 28.5%-ban fordult elő ez a topik, amely a korpusz második legáltalánosabb topikja.

Az ötödik topik a romák és nem romák közötti általános társadalmi problémákról szól. A cikkek többsége a társadalom romaellenes attitűdjeiről, az etnikai konfliktusokat okozó társadalmi tényezőkről, a romák eltérő kulturális szokásairól és morális értékrendjéről világosítja fel az olvasókat. Általános problémakánt jelenik meg például a munkanélküliség, a bűnözés, az iskolai szocializáció visszautasítása, a társadalmi integráció visszautasítása vagy a szegénység. A szövegekben szépen kirajzolódnak a roma-magyar, a cigány-roma és a többség-kisebbség fogalmi ellentétek, és emellett a baloldali-jobboldali, a liberális-jobboldali[22] és a szegény-gazdag ellentétpárok is gyakran megjelennek. A topik a korpusz 18.4%-ban jelent meg.

A hatodik topik romákhoz kötött ingatlanügyi, lakhatással kapcsolatos gondokat jelöl. Olyan ügyek jelennek meg az ehhez a topikhoz tartozó cikkekben, mint a lakásmaffia, az áramlopások, az épületek leélése, „jogosulatlan" önkormányzati lakásépítési támogatások vagy a számlák nem fizetése miatt kilakoltatások. A cikkek között sok az olvasói levél, amelyekben az olvasók segítséget kérnek áramlopások miatt, vagy olyan körülményekről számolnak be, „amelyek közepette szinte már lehetetlen élni. Bűz, kosz, leszaggatott, életveszélyes vezetékek, csótányok és hangoskodás." Ez a téma a cikkek 8.9%-ában fordult elő.

A hetedik topikhoz tartozó cikkekben a politikai pártok intézkedéseiről, a politikusok tevékenykedéséről és nyilatkozatairól tájékoztatják az olvasókat. Gyakori Orbán Viktor, Gyurcsány Ferenc és egyéb „szocionista", illetve „zsideszes" politikusok tetteinek, nyilatkozatainak negatív színben való feltüntetése, például Mohácsi Viktória „lógásai" az Európai Parlament üléseiről. Vona Gábor és a Jobbik kezdeményezéseit, intézkedéseit, nyilatkozatait mindemellett pozitív hangnemben tárják az olvasók elé. Ezt a topikot az elemzett cikkek 8.7%-a tartalmazta.

A nyolcadik topik cikkei más médiumokban (hírlapok, TV-műsorok stb.) megjelent cikkekre, adásokra adott reakciókat tartalmaznak. Két általános osztálya létezik ezeknek a híreknek. Az egyikbe azok a cikkek tartoznak, amelyek a kuruc.info szerkesztői által tanulságosnak talált cikekkről, sztorikról, videókról számolnak be. A másik osztályba tartoznak azok a cikkek, amelyek a „cigányfajvédő", „bűnözővédő", „balzsidó" média (pl. „Zsindex", „Zsorigó", „168 tóra", „Heti Tetves") egy hírét, adását, megfogalmazását vagy egy ezekhez a médiumokhoz kapcsolódó eseményt tárgyalnak és saját szempontból láttatnak. A korpusz cikkeinek 30.2%-a tartalmazta ezt a top-

[22] ahol a liberális a kuruc.info értelmezésében a „zsidó, buzi vagy cigó" csoportokat jelenti

ikot, amely a leggyakoribb topik a korpuszban. Ez a téma látszólag nem igazán tartozik a romareprezentációs diskurzushoz, inkább az újságírás műfajából fakad, azonban a topikhoz tartozó cikkek megnyilvánulásai által kifejezett attitűdök és a kirajzolódó ellentétpárok könnyen beilleszthetők a romákról szóló diskurzusba.

A kilencedik topikban íródott cikkek helyszínei vidéki falvak, községek, ahol a helyi, gyakran idős gazdák nehéz helyzetéről számolnak be a romák állandó és megállíthatatlan terménylopásai, betörései és rongálásai miatt. Gyakori esetként jelenik meg, hogy a gazdák áramot vezetnek a portájuk kerítésébe, hogy azt védeni tudják az „orkok" elől. A cikkekben a közbiztonság alacsony szintjét hangsúlyozzák, ahol csak a polgárőrség nyújt némi védelmet. A topik a cikkek 9.7%-ában jelent meg.

A tizedik topik szintén a romák által elkövetett bűncselekményekről szól, azonban ez a topik a rendőri munka és jelenlét hiányosságait és az alacsony közbiztonságot hangsúlyozza. A cikkekben gyakran megjelenik, hogy az emberek a kezükbe szeretnék venni a saját védelmüket (pl. amikor Szepessy Zsolt monoki polgármester az önvédelmi törvény módosítására tett javaslatot), vagy hogy egyes településeken saját rendőrséget, polgárőrséget alakítanak az emberek, hogy megvédjék magukat. Ez a téma a cikkek 22.2%-ában fordul elő.

A tizenegyedik topik az egészségüggyel kapcsolatos tárgyköröket foglalja magába, a TBC és Hepatitis A vírusok romáknak tulajdonított elterjesztésétől kezdve, az újévi elsőszülött Rikárdón vagy a patkány marcangolta csecsemőn át az orvosnőt leköpő cigányasszonyig. Az egészségügyi topik a korpusz cikkeinek 2.2%-ához köthető.

A tizenkettedik topik szemantikailag lazább topiknak számít, ugyanis az ehhez a topikhoz tartozó cikkek legnagyobb része olvasói levél. Az olvasói leveleket a kuruc.info olvasói egészen vegyes témában írják, a húsvét romák általi koldusünneppé változtatásától kezdve a Blaha Lujza téren vagy a szoláriumstúdióban látottak leírásáig. A topik viszonylag gyakori a korpuszban, a cikkek 28.4%-ában fordulnak elő a topikhoz köthető szavak.

A tizenharmadik topik egy vegyes topik. A topikhoz tartozó cikkek egyrészt a romák fa- és fémlopásai okozta árvizeket tárgyalja. A cikkek másik fele a fegyvertartási engedélyről és az egyén önvédelméről szól, ahol az USA törvényeit szeretik felhozni példaként. A témát a cikkek 2.2%-a tartalmazza.

A tizennegyedik topikhoz tartozó cikkek a romák által elkövetett kábel- és fémlopásokról, illetve rongálásokról szólnak. A cikkek a lopások és rongálások okozta közlekedési fennakadásokról és értékbeli károkról is beszámolnak. A lopások külön kategóriája, mikor vallásos helyeken (templom, temető) történik a lopás vagy a ron-

gálás. A topik a korpusz cikkeinek 10.9%-ában jelenik meg.

A tizenötödik topikba került minden romák által elkövetett bűncselekmény bírósági eljárása, tárgyalása és kimenetele, a nagy port kavart olaszliszkai lincselés bírósági ügyének nyomon követésétől kezdve, a terhes nőt gyilkolt fiú elítélésén keresztül a „makói cigányállat" bírósági ítéletéig. A téma a cikkek 12.2%-ában fordul elő.

A tizenhatodik topik Marian Cozma, román kézilabdázó gyilkosságának témáját tartalmazza, akit 2009-ben egy veszprémi szórakozóhely előtt szúrtak szíven. A gyilkosság ügyében Raffael Sándort, Németh Győzőt és Sztojka Ivánt ítélték el. Az ügy hatalmas országos felháborodást keltett, ennek köszönhető, hogy az eset külön topikot alkot a „cigánybűnözés" korpuszon belül. A topikot a cikkek 4%-a tartalmazza.

A tizenhetedik topik közlekedési kihágásokat, baleseteket, bűncselekményeket foglal magában. A cikkekben az apróbb kihágásoktól kezdve (tilosban parkolás, záróvonal átlépése) az ittas vezetésen és kisebb balesetek okozásán keresztül a gázolásig és halálos balesetekig minden megjelenik. A cikkekben gyakran csak a BMW vagy a Lada autómárkák utalnak arra, hogy a történtek elkövetője roma. Ez a topik a korpusz cikkeinek 8.3%-ában jelenik meg.

A tizennyolcadik topik cikkei verekedésekről, késelésekről, erőszakos támadásokról számolnak be, amelyek kimenetele a könnyű testi sértéstől az életveszélyes állapotig terjed. A cikkek továbbá beszámolnak az ügyek rendőrségi folytatásáról (pl. feljelentés, őrizetbe vétel). A topik a cikkek negyedéhez köthető.

A tizenkilencedik topik egy a médiában és a közbeszédben szintén nagy port kavart esetet jelenít meg, Pásztor Albert miskolci rendőrkapitány ügyét, aki egy 2009. januári sajtótájékoztatójában kijelentette a cigánybűnözés tényét az előző két hónap rendőrségi ügyeire hivatkozva. Draskovics Tibor igazságügyi és rendészeti miniszter Pásztort nyilatkozata után más beosztásba helyezte. A kulcsszavak között megjelenik Lakatos Attila borsodi cigányvajda és Káli Sándor, miskolci polgármester neve is, akik kiálltak Pásztor mellett. A topik a korpusz 3.1%-ban fordul elő.

A huszadik topikhoz a kultúráról, roma vagy romákat szerepeltető filmekről, TV-műsorokról, zenészekről, színházi előadásokról szóló cikkek kerültek. A művészek és médiaszereplők sorában megjelenik Győzike az adócsalási ügyével, L.L. Junior „gádzóellenes uszítása" és barátnőjének kifosztása miatt, valamint Damu Roland színész, aki nőverési ügye miatt került a kuruc.info-ra.[23] A témát a cikkek 2%-a érinti.

[23] Ebbe a topikba került a néhány megmaradt nem magyar nyelvű szöveget tartalmazó cikk, amelyk ugyan nem zavartak be a topik legfontosabb kulcsszavaiba.

A huszonegyedik topikban főképp romák által elkövetett gyilkosságok, halálos áldozattal járó bűncselekmények jelennek meg. Kiemelt esemény Bándy Kata, pécsi lány gyilkossága és az ügy nyomozása. Emellett a házgyújtogatási esetek is ebben a topikban jönnek elő, például a tatárszentgyörgyi sorozatos gyújtogatás. A topik a cikkek 11.5%-ában jön elő.

A huszonkettedik topik két téma keveredését tartalmazza. Az egyik a kuruc.info-n nagy népszerűségnek örvendő A tücsök és a hangya, La Fontaine tanmese mindenféle átköltése hétköznapi esetekre. A mesében a tücsök szimbolizálja a cigányságot. A másik téma a finnországi cigánybűnözési eseteket írja le, ugyanis a rovatnak van egy finnül tudó állandó szerzője, aki előszeretettel fordít finn cikkeket és fórumbeszélgetéseket. A topikot a cikkek 2.5%-a tartalmazza.

A huszonharmadik topik a szociálpolitikai, foglalkoztatási, munkaerőpiaci ügyekkel foglalkozik. Ide tartozik a szociális segélyek és a közmunka kérdése, és kiemelt eseményként szerepel Szepessy Zsolt monoki polgármester döntése, aki az ingyenes szociális segélyt közmunkáért cserébe folyósította a roma lakosok számára. A cikkekben a romák naplopóként és lustának vannak feltüntetve és a polgármester döntését helyeslik. A topik a korpusz cikkeinek 8.6%-ában jelent meg.

A huszonnegyedik topikhoz tartozó cikkek a Jobbikhoz tartozó Szebb Jövőért Polgárőr Egyesület járőrözéseiről és demonstrációiról számolnak be, például Gyöngyöspatán, ahol az egyesület „hihetetlen sikert ért el", vagy Hajdúhadházán. A demonstrációk helyszínei tipikusan vidéki települések, ahol az egyesület az „etnikai konfliktus" ellen lépett fel. A téma a cikkek 2%-ában fordult elő.

A huszonötödik topik a romák európai bevándorlását foglalja magában. A cikkekben szereplő befogadó országokként fejlett nyugati jóléti államok jelennek meg, mint Finnország, Svédország, Németország vagy Dánia. A cikkek a roma bevándorlást a törökök, a muszlimok és a feketék bevándorlásával kötik össze, amely kulturális problémát jelentenek a felsorolt országokban. A finnek nyílt idegenellenessége és tettre készsége követendő példaként jelenik meg („a finnek nem gatyáznak, ha a cigányokkal, ill. a négerekkel kell leszámolniuk"). A topik a cikkek 3%-ában jött elő.

A huszonhatodik topik cikkei oktatási, iskolai konfliktusokat fogalmaznak meg. Ide tartoznak az iskolai erőszak fajtái, például a tanárverések, vagy a nem roma gyerekek bántalmazása, azaz a „cigányterror", és egyéb deviáns viselkedések (például ürülék szétkenése, lopás, teremből kiszökés). Kiemelt személyként jelenik meg a roma gyerekek iskolai szegregációja ellen küzdő Mohácsi Erzsébet, az Esélyt a Hátrányos Helyzetű Gyermekeknek Alapítvány elnöke, akinek nyilatkozatait előszeretettel

kritizálják az oldal szerzői.[24] Az oktatásügyi topik a korpusz 7.3%-ához köthető.

Az utolsó topik a cigány bevándorlók külföldi bűnözésének egy konkrétabb szeletét érinti, az emberkereskedelemmel, prostitúcióval vádolt cigány bűnszervezetek elleni nyomozásokat, rendőrségi eljárásokat elkövetett bűntényeket. Kiemelt ügyként kezelik a kanadai magyar cigányokból álló emberkereskedelemmel, szervkereskedelemmel, prostitúcióval, „fajtársaik csicskáztatásával" vádolt banda 2011-es esetét, amely „Kanada történetének legnagyobb emberkereskedelmi ügye". Egyéb országok is megjelennek az emberkereskedő, nőket és kiskorúakat futtató szervezetek székhelyéül, például Franciaország, Olaszország vagy Hollandia. A 2009. elején eltűnt svájci diáklány, Ophélie esetét is ebben a keretben helyezték el az oldal szerzői. Ezekről az ügyekről a kuruc.info szerint természetesen a „sajtó baloldali része mély kussban van". A topik a korpusz cikkeinek 4.5%-át érinti.

A többi modellben kisebb eltérésekkel a fent vázolt topikstruktúra rajzolódik ki, a szimmetrikus α paraméterrel illesztett modell topikjai is nagyjából megfeleltethetők a leírt topikoknak.

A korpuszban a hipotéziseket megerősítve a bűnözés téma van jelen a legnagyobb mértékben, amelynek mindenféle fokozata megjelenik a korpuszban. A romák által elkövetett bűncselekmények a lopásoktól kezdve (4., 9., 13., 14. topik), a verekedéseken, késeléseken, erőszakos támadásokon (18. topik) és a közlekedési bűncselekményeken át (17. topik), a gyilkosságokig (16., 21. topik) terjednek. Illetve a bűnözés témához tartozik a bűncselekményekhez köthető rendőrséghez (15., 19. topik) és a bírósághoz köthető folyamatok (18. topik). A bűnözés témakörhöz számíthatjuk a bűncselekményeket általánosabban érintő topikokat is és az emberek „cigánybűnözésre" adott reakcióit (10., 12., 13. topik). Emellett igazolódott az a hipotézisem is, hogy a legnagyobb közfelháborodást keltő események külön topikot alkotnak, mint a Cozma-gyilkosság vagy Pásztor Albert ügye. Az olaszliszkai lincselés azonban nem alkotott külön topikot, de a második topik fontos részeként jelenik meg.

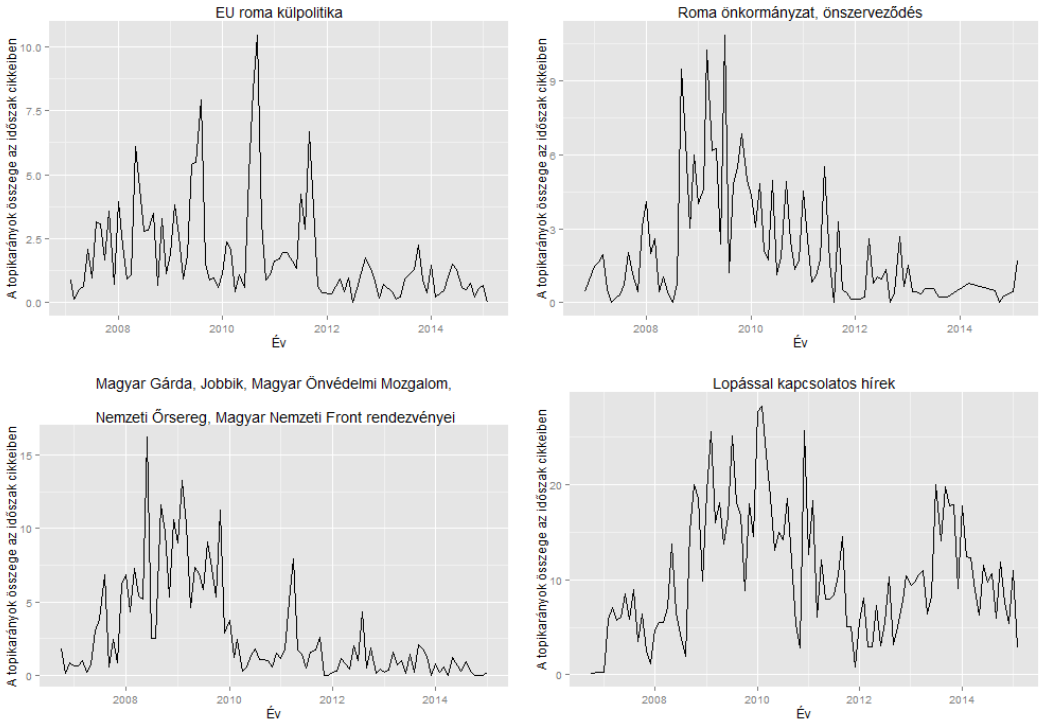
A bűnözés témája mellett a romák kivándorlása, a külföldi romák helyzete, bűncselekményei és a külpolitikai intézkedések témái is feltűnnek a korpuszban. Legtöbbször azonban ez a téma is besorolható a bűnözés témába, ugyanis a cikkek nagy részében a külföldre vándorló romák bűnöző életmódját emelik ki. Ehhez a témához tartozik az 1., a 22., a 25. és a 27. topik.

Különálló témaként jelenik meg továbbá a roma önkormányzatiság és önszerveződés (3. topik), a politika és a (szélső)jobboldalhoz kapcsolható szerveződések

[24] Olyannyira, hogy az iskolai problémákat okozó roma diákokat „Mohácsi Erzsébet-féle köztörvényes bűnözőknek" nevezik.

(2., 7., 24. topik), az egészségügy (11. topik), a kultúra és művészet (20. topik), a foglalkoztatás és a szociális helyzet (23. topik), az oktatás (26. topik), valamint az általános roma-nem roma konfliktusok, előítéletek (5., 6., 8. topik). Azonban a cikkek szinte az összes témát a bűnözés szempontjából világítják meg, a roma művészek, politikusok, diákok, aktivisták általában bűnözőként jelennek meg, ezért ezeket a topikokat is nehéz elválasztani a bűnözés témájától.

A topikok időbeli megjelenítését az E.13 függelékrészben lévő kóddal hoztam létre, amelyek közül a szakdolgozat keretein belül csak az első négy topik időbeli változásainak csúcspontjait elemzem. Az elkészült ábrák az M fejezetben találhatók, amelyek közül a 3.1 ábra topikváltozásait veszem szemügyre.



3.1. ábra. Topikok időbeli változása - 1-4. topik

Az első topik az EU államok roma külpolitikájával foglalkozik. Az ábra alapján 2010 második felében történt a téma szempontjából legkiemelkedőbb esemény, amely a 2010. augusztus-szeptemberi román és bolgár romák Sarkozy általi kiutasítása Franciaországból és a körülötte zajló EU-s viták.

A második topik a roma önkormányzatok, szervezetek témáját öleli fel. A topikban 2008 második felétől kezdve 2009 közepéig láthatók kicsúcsosodások, amelyek Kolompár Orbán büntetőeljárásainak ügyeit jelöli. 2008 őszén jogosulatlan gazdasági előny megszerzése miatt, 2009 júliusában sikkasztás és a számviteli rend megsértése miatt emeltek ellene vádat.

A harmadik topik a Magyar Gárdához és egyéb szélsőjobboldali és jobboldali szervezetekhez kapcsolódik. A topikot leginkább a Magyar Gárdával kapcsolatos hírek uralják. A topik időbeli alakulása követi a Gárda működésének idejét a 2007. augusztusi megalakulástól a 2009-es felbomlásig. A leginkább kiemelkedő csúcs 2008 első felében található, ez a csúcs 2008. júniusi pátkai eseményeket jelöli.

A negyedik topikhoz a lopással kapcsolatos bűncselekményeket tartalmazó cikkek tartoznak. A topik a korpusz egészében aktívnak mondható az eddigi három topikkal szemben. A 2010-es év eleji csúccsal kapcsolatban nem találtam semmilyen kirívó esetet, a jelenség oka az lehet, hogy januártól márciusig nagy számban jelentek meg a lopási ügyekkel foglalkozó cikkek a kuruc.info-n.

3.7. Az eredmények összehasonlítása kvalitatív kutatások eredményeivel

Az általam kapott topikstruktúrát a Bernáth és Messing által kivitelezett kutatások során használt témastruktúrával szeretném összehasonlítani. A célom, hogy megvizsgáljam, mennyire illeszthető bele az általam végzett kutatás az eddigi társadalomtudományi gyakorlatba.

Bernáth és Messing kutatásaikban humán annotálással végezték a cikkek kategóriához rendelését, azaz elolvastak minden cikket és besorolták a dokumentumokat a megfelelő osztály(ok)ba. Hogy a látens Dirichlet allokáció során kapott eredményeket össze tudjam vetni ezzel a témafelosztással, az elemzés során kapott topikokat a kategóriarendszer szerint osztom be, és a korpusz egy részét humán annotátorok segítségével osztályozom ugyanezen kategóriarendszer alapján. Az osztályozott cikkek véletlenszerűen lettek kisorsolva a D.13 kód segítségével. Összesen 622 cikk került annotálásra.

[Bernáth and Messing, 2012] tanulmány témastruktúrájába az LDA-val kapott topikokat a 3.1 táblázat szerint osztottam be. Az LDA-val kapott topikok viszonylag könnyen párosíthatók Bernáth és Messing témastruktúrájával, azonban voltak topikok, amiket több témakörbe is bekerültek. Például az első topik a külföldi romák és a külpolitika témakörbe is beleillik, a harmadik topik a politika, a roma önszerveződés és a bűnözés témakörökbe. Általánosan elmondható a topikokról, hogy a bűnözés, valamint a diszkrimináció, előítéletek, roma-többségi konfliktus témákhoz is besorolhatók. A kuruc.info cikkeiben tehát majdnem minden téma megtalálható, amely a magyar írott médiában általánosan, kivéve a jogvédelem, kisebbségi jogok, a gazdaság, vállalkozás, valamint a természeti katasztrófa témakörök. Utóbbi viszony-

lagos ritkasága miatt nem jelenhetett meg jellemző topikként, előbbi kettő pedig a kuruc.info értékbeállítódása miatt kaphattak kisebb hangsúlyt.

Az sem nyújt meglepetést, hogy a Cigánybűnözés rovat cikkei főként bűnözéssel kapcsolatos témákat tárgyalnak. Az LDA által feltárt topikok többsége a bűnözés szemszögéből láttatja az elsődleges témákat. A kuruc.info-n a bűnözésen kívül szívesen írnak előítéletes témákról, roma-nem roma konfliktusokról és a politikáról.

Téma sorszáma	[Bernáth and Messing, 2012] témái	LDA topikok
1	Politika, közpolitika (kormányzati és önkormányzati)	3., 5., 7., 19., 23., 24., 25.
2	Roma önszerveződés, önkormányzatiság	3.
3	Kivándorlás	25., 27.
4	Külföldi romák	1., 25., 27.
5	Jogvédelem, kisebbségi jogok	
6	Szegénység, szociális helyzet	5.
7	Oktatás	26.
8	Foglalkoztatás, munkapiac	23.
9	Gazdaság, vállalkozás	
10	Kultúra, művészet	20.
11	Egészség/ügy	11.
12	Diszkrimináció, előítéletek, roma-többségi konfliktus	2., 5., 6., 8., 9., 12., 16., 17., 21., 23., 24., 25.
13	Bűnözés	2., 3., 4., 6., 9., 10., 11., 12., 13., 14., 15., 16., 17., 18., 19., 20., 21., 22., 24., 26., 27.
14	Külpolitika, EU	1., 25.
15	Természeti katasztrófa	

3.1. táblázat: A [Bernáth and Messing, 2012] témastruktúrájának és az LDA-val kapott topikok összevetése

A hipotézisem, miszerint a kuruc.info cigányellenes cikkeiből kinyert topikok nagyjából megfeleltethetők lesznek a Bernáth és Messing által meghatározott témastruktúrának, tehát beigazolódott.

3.7.1. Kiértékelés

A következőkben a humán annotátorok által besorolt cikkek kategóriával hasonlítom össze a cikkekhez tartozó, látens Dirichlet allokációval kinyert topikokat. Ezzel tulajdonképpen a látens Dirichlet allokáció diskurzuselemzési hatékonyságát mérem fel.

A kiértékeléshez az információkinyerés területén az osztályozók teljesítményének mérésér gyakran használt két mérőszámot, a felidézést (recall, R) és a pontosságot (precision, P), illetve ezek harmonikus közepét, az F-mértéket (F-measure) fogom használni. A mérőszámokat [Tikk et al., 2007] irodalom alapján írom le. A metrikák meghatározásához először el kell különítenünk a valós pozitív (true positive, TP), a hamis negatív (false negative, FN) és a hamis pozitív (false positive, FP) eseteket az annotált és az LDA által meghatározott topikok alapján. *Valós pozitívnak* hívjuk azt az esetet, amikor a modell helyesen rendelte egy cikkhez a humán annotátorok által meghatározott kategóriát. *Hamis negatív* egy eset, amikor a modell nem rendelt hozzá egy cikkhez egy kategóriát a humán annotátorok által meghatározott kategóriák közül. Valamint *hamis pozitívnak* nevezzük azt az esetét, amikor a modell hozzárendelt egy cikket egy kategóriához, azonban a humán annotátorok alapján nem tartozik ahhoz a kategóriához.

A *felidézés* azt írja le, hogy a humán annotátorok által meghatározott kategóriákat milyen arányban kategorizálta be a modell helyesen. A felidézést a következő képlet szerint számoljuk:

$$R = \frac{TP}{TP + FN}.$$

A *pontosság* azt méri, hogy a modell által meghatározott kategóriák hány százaléka olyan, amelyet a humán annotátorok is bekategorizáltak. A pontosság tehát azt próbálja meghatározni, hogy a modell kategóriái mennyire relevánsak. A mérőszámot a következőképp számítjuk:

$$P = \frac{TP}{TP + FP}.$$

Az *F-mérték* a felidézés és a pontosság harmonikus közepe. Mivel a feladatban a felidézést és a pontosságot egyenlő súlyúnak tekintem, az F-mértéket a következőképp számítandó:

$$F = \frac{2PR}{P + R}$$

A Függelék E.3 részében található kód segítségével hasonlítottam össze az annotátorok és az LDA modell kategóriáit. Az annotation.csv tartalmazza az annotátorok kategóriát, a doc-topics_1.csv pedig a MALLET outputja a dokumentumok topikeloszlásáról. Miután kiszedtem utóbbi adatbázisból az annotált cikkeket sorszám alapján, a MALLET által adott topikszámokat átkategorizáltam a 3.1 táblázat szerint. Ezután kiszámoltam a hamis negatív, a valós pozitív és a hamis pozitív eseteket, valamint a felidézést és a pontosságot kategóriánként. Végül a kategóriák elemszámával arányosan összesítettem a felidézést és a pontosságot és meghatároztam az F-mértéket.

A felidézés 73.56%-os lett, azaz az LDA modellnek az esetek közel háromnegyedében sikerült helyesen meghatároznia a humán annotátorok által felismert kategóriákat. Ez az érték szemantikai célú feladatoknál jónak számít. A pontosság 54.64%-os lett, tehát a modell által meghatározott kategóriák 54.64%-át határozták meg a humán annotátorok is. Ez az érték azért lett alacsonyabb, mivel magas volt a hamis pozitív esetek aránya, tehát azoké az eseteké, amikor a modell hozzárendelt egy cikket egy kategóriához, azonban a humán annotálás alapján a cikk nem tartozott abba a kategóriába. Azt gondolom, hogy ez nem egyenlő azzal, hogy a modell által megjelölt egyéb kategóriák helytelenek lennének. Inkább arról lehet szó, hogy az LDA több kategóriával jelölte meg a cikkeket, mint a humán annotátorok, ugyanis a modell már akkor egy topikhoz rendelt egy cikket, ha a cikk szövegének 10%-a adott topikhoz tartozott. A két mérőszám alapján kiszámolt F-mérték 62.70%.

A kapott értékek azt mutatják, hogy a látens Dirichlet allokáció megfelelő támogatást nyújthat a diskurzuselemzéshez, a humán annotátorok által meghatározott kategóriákat ugyanis az esetek majdnem háromnegyedében helyesen határozta meg. A hamis pozitív esetek kérdését érdemes lenne egy másfajta vizsgálattal ellenőrizni.[25] Mindenesetre az elemzés során az LDA által a dokumentumokhoz rendelt topikok nagyfokú helytállóságát tapasztaltam. A látens Dirichlet allokációval támogatott diskurzuselemzés tehát beilleszthetőnek tűnik az eddigi társadalomtudományos gyakorlatba és alkalmasnak látszik az írott média romareprezentációs vizsgálatára.

[25] Érdemes lenne ezt a kérdést például magasabb topikarány küszöb esetén megvizsgálni.

4. MEGBESZÉLÉS

A látens Dirichlet allokáció és a topik modellek kutatása egy folyamatosan fejlődő terület, amelynek célja nagy méretű szöveggyűjtemények szemantikai, tematikus struktúrájának minél hatékonyabb feltárása. A topik modellek alkalmazása a társadalom- és egyéb humán tudományok szakértői számára eddig kiaknázatlan, de jelentős potenciállal rendelkezik, amelyet a szakdolgozatommal is bizonyítani szerettem volna.

A dolgozat első felében a látens Dirichlet allokáció matematikai és történeti hátterét mutattam be, amely során betekintést nyújtottam a látens Dirichlet allokáció tágabb statisztikai elméleti keretébe, a bayesi következtetéselméletbe, illetve a vegyes tagságú modellcsalád általános jellemzőibe, amely modellcsaládba a látens Dirichlet allokáció és a topik modellek tartoznak. Ezután ismertettem a látens Dirichlet allokáció közvetlen előzményeit, a látens szemantikus indexelést és a valószínűségi látens szemantikus indexelést, valamint a használatukhoz szükséges vektortérmodell nevű dokumentumreprezentációs modellt. Ezt követően a látens Dirichlet allokáció matematikai leírását részleteztem, majd leírtam a topik modellek aktuális kutatási területeit, például a modellilleszkedés ellenőrzését szolgáló módszerek kutatását. Ezután olyan kutatásokról számoltam be, amelyek során a látens Dirichlet allokációt és egyéb topik modelleket társadalomtudományi céllal használták fel.

A dolgozat második felében egy saját kutatás keretében szemléltettem a látens Diirchlet allokáció társadalomtudományi alkalmazhatóságát. A módszerrel a nemzeti radikális, szélsőjobboldali kuruc.info online hírportál Cigánybűnözés rovatának 10.304 cikkén végeztem elemzést, és a korpuszból a látens Dirichlet allokáció segítségével 27 társadalomtudományi szempontból releváns témát nyertem ki. Az elemzés során a hagyományos kvalitatív diskurzuselemzésnél a látens Dirichlet allokáció segítségével feltárt témákra támaszkodtam, azaz egy vegyes, a hagyományos kvalitatív diskurzuselemzést kvantitatív elemekkel ötvöző módszertant hajtottam végre. Az alkalmazott módszertan előnye, hogy segítségével nagy mennyiségű szöveges adatot elemezhetünk, az elemzés menetét a topikstruktúra automatikus kinyerése pedig gyorsabbá és egyszerűbbé teszi. Ezenkívül a folyamat kevesebb emberi és pénzbeli erőforrást igényel, valamint az alkalmazott módszerrel csökkenthető a szubjektivitás

és javítható a kutatás reprodukálhatósága.

A dolgozatban a látens Dirichlet allokáció által kinyert topikok kulcsszavainak, illetve a legjellemzőbb dokumentumainak segítségével értelmeztem a topikokat és a topikok időbeli változását is megjelenítettem a cikkek metaadatául szolgáló időbélyegek segítségével. Ezáltal elemezhetővé váltak a szélsőjobboldali médiadiskurzusban nagy visszhangot kapott események. A dolgozat keretein belül sajnos nem tudtam a topikokat és topikok időbeli változásait megfelelő mélységig elemezni, ezért a jövőben a dolgozat keretein kívül mindenképpen érdemes lenne egy bővebb elemzést készíteni.

Ezenkívül ajánlatos lenne a látens Dirichlet allokációnál szofisztikáltabb modellek alkalmazása, például a korrelált topik modellé, amelyhez az R topicmodels csomagja nyújt implementált modellillesztő függvényt. Ezen felül más metadatokkal bővített modelleket is érdemes lenne létrehozni, például a cikkek Cigánybűnözés rovaton kívüli egyéb rovatba tartozását is be lehetne vonni az elemzésbe. Utóbbi típusú elemzéshez a Roberts, Stewart és Tingley által fejlesztett stm csomag használatát lehetne igénybe venni.

Mindemellett az olyan kutatási területek kérdése is fontos, mint a modellek feltételeit ellenőrző eljárások fejlesztése. A szakdolgozatban a Blei és Mimno által kifejlesztett kölcsönös információt felhasználó posterior prediktív ellenőrzéshez írtam kódot, amely megfelelően kimutatta a szavak polinomiális feltételtől való szisztematikus eltérését.

A látens Dirichlet összességében megfelelő alapot nyújtott a kutatási téma szempontjából releváns információk kvalitatív elemzéséhez és értelmezéséhez, és a végrehajtott kutatás beilleszthetőnek bizonyult a magyar média romareprezentációs kutatásainak gyakorlatába. Mindez alátámasztotta, hogy a látens Dirichlet allokációval támogatott diskurzuselemzés hasznos eszköze lehet a társadalomtudományi célú szövegelemzéseknek.

IRODALOMJEGYZÉK

A. Agresti. *Categorical Data Analysis*. John Wiley & Sons Inc, Hoboken, New Jersey, second edition, 2002.

E. M. Airoldi, D. Blei, E. A. Erosheva, and S. E. Fienberg. Introduction to mixed membership models and methods. In E. M. Airoldi, D. Blei, E. A. Erosheva, and S. E. Fienberg, editors, *Handbook of Mixed Membership Models and Their Applications*, pages 3–15. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, 2014.

P. Baker, C. Gabrielatos, and T. McEnery. *Discourse Analysis and Media Attitudes: The Representation of Islam in the British Press*. Cambridge University Press, 2013.

A. Bernát, A. Juhász, P. Krekó, and C. Molnár. The roots of radicalism and anti-roma attitudes on the far right. TÁRKI, 2013.

G. Bernáth. Hozott anyagból. a magya média romaképe. *Beszélő*, június 2003.

G. Bernáth and V. Messing. Vágóképként csak némában. romák a magyarországi médiában. Nemzeti és Etnikai Kisebbségi Hivatal, Budapest, 1998.

G. Bernáth and V. Messing. Szélre tolva. *Médiakutató*, tavasz 2012.

D. Blei. Introduction to probabilistic topic models, 2011.

D. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

D. M. Blei and J. D. Lafferty. Dynamic topic models. *International Conference on Machine Learning, New York, NY, USA*, pages 113—120, 2006.

D. M. Blei and J. D. Lafferty. Topic models. In A. N. Srivastava and M. Sahami, editors, *Text mining: Classification, Clustering, and Applications*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, 2009.

M. Bolla and A. Krámli. *Statisztikai következtetések elmélete*. Typotex, Budapest, 2012.

M. K. Cowles and B. P. Carlin. Markov chain monte carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434): 883–904, 2006.

S. P. Crain, K. Zhou, and S.-H. Yang. Dimensionality reduction and topic modeling: From latent semantic idexing to latent dirichlet allocation and beyond. In C. C. Aggarwal and C. Zhai, editors, *Mining Text Data*, pages 3–15. Springer US, 2012.

P. DiMaggio, M. Nag, and D. Blei. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of u.s. government arts funding. *Poetics*, (41):570–606, 2013.

M. Feischmidt. Rasszizmus és média. In M. Bogdán, M. Feischmidt, and Á. Guld, editors, *„Csak másban". Romareprezentáció a magyar médiában*, pages 111–122. Gondolat Kiadó, PTE Kommunikáció- és Médiatudományi Tanszék, Budapest-Pécs, 2013.

A. Galyardt. Interpreting mixed membership models: Implications of erosheva's representation theorem. In E. M. Airoldi, D. Blei, E. A. Erosheva, and S. E. Fienberg, editors, *Handbook of Mixed Membership Models and Their Applications*, pages 39–65. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, 2014.

A. Gelman and X.-L. Meng. Model checking and model improvement. In W. Gilks, S. Richardson, and D. Spiegelhalter, editors, Practical Markov Chain Monte Carlo, pages 189–201. Chapman $ Hall, London, 2005.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analaysis*. Chapman & Hall/CRC Press, third edition, 2014.

R. Glózer. A „cigányok" mint ellenség diszkurzív konstrukciói a hazai online szélsőjobboldali médiában. In M. Bogdán, M. Feischmidt, and Á. Guld, editors, *„Csak másban". Romareprezentáció a magyar médiában*, pages 123–140. Gondolat Kiadó, PTE Kommunikáció- és Médiatudományi Tanszék, Budapest-Pécs, 2013.

T. Griffiths and M. M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, pages 5228–5235, 2004.

T. Griffiths, M. Steyvers, D. Blei, and J. Tenenbaum. Integrating topics and syntax. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 537—544. MIT Press, Cambridge, MA, 2005.

J. Grimmer. A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, (18):1–35, 2010.

G. Heinrich. Parameter estimation for text analysis. Technical report, 2005.

T. Hofmann. Probabilistic latent semantic indexing. http://ciir.cs.umass.edu/pubfiles/ir-464.pdf, 1999.

L. Hunyadi. Bayesi gondolkodás a statisztikában. *Statisztikai Szemle*, 89.(10-11.): 1150–1171, 2011.

D. Kehl and V. Várpalotai. A modern bayesi elemzések eszköztára és alkalmazása. *Statisztikai Szemle*, 91.(10.):971–992, 2013.

B. Kriza and Z. Vidra. A többség fogságában – kisebbségek médiareprezentációja. In M. Feischmidt, editor, *Etnicitás. Kisebbségteremtő társadalom*. Gondolat Kiadó - MTAKI, Budapest, 2010.

W. Li and A. McCallum. Pachinko allocation: Scalable mixture models of topic correlations. *Journal of Machine Learning Research*, 2008.

W.-H. Lin, E. P. Xing, and A. Hauptmann. A joint topic and perspective model for ideological discourse. In *ECML PKDD*, pages 17–32, 2008.

C. Lucas, R. Nielsen, M. E. Roberts, B. M. Stewart, A. Storer, and D. Tingley. Computer assisted text analysis for comparative politics, 2014.

C. D. Manning and H. Schütze. *Foundations of Satistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, London, England, 1999.

A. K. McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

D. Mimno and D. Blei. Bayesian checking for topic models. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011.

V. Munk. A romák reprezentációja a többségi média híreiben az 1960-as évektől napjainkig. *Médiakutató*, 14.(2.):89–100, 2013.

A. Pócsik. Romák a magyar médiaelemzésekben. In M. Bogdán, M. Feischmidt, and Á. Guld, editors, *„Csak másban". Romareprezentáció a magyar médiában*, pages 177–202. Gondolat Kiadó, PTE Kommunikáció- és Médiatudományi Tanszék, Budapest-Pécs, 2013.

M. Ponweiser. Latent dirichlet allocation in r. Master's thesis, Wirtschaftsuniversität Wien, 2012. URL `http://epub.wu.ac.at/3558/1/main.pdf`.

I. Porteous, A. Asuncion, D. Newman, P. Smyth, A. Ihler, and M. Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. In *In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 569–577, 2008.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org`.

D. Ramage, E. Rosen, J. Chuang, C. D. Manning, and D. A. McFarland. Topic modeling for the social sciences. In *NIPS 2009 Workshop on Applications for Topic Models: Text and Beyond*, Whistler, Canada, 2009.

M. E. Roberts and B. M. Stewart. Localization and coordination: How propaganda and censorship converge in chinese newspapers. 2014.

M. E. Roberts, B. M. Stewart, D. Tingley, and E. M. Airoldi. The structural topic model and applied social science. 2014a.

M. E. Roberts, B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson, and D. G. Rand. Structural topic models for open-ended survey responses. *American Journal of Political Science*, 2014b.

G. Rossum. Python reference manual. Technical report, Amsterdam, The Netherlands, 1995.

M. Steyvers and T. Griffiths. Probabilistic topic models. In T. Landauer, D. S. McNamara, D, and W. Kintsch, editors, *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum, 2006.

T. Terestyéni. A sajtó roma vonatkozású tartalmai a 2002-es parlamenti választások kontextusában. *Médiakutató*, nyár 2004.

D. Tikk, R. Farkas, Z. T. Kardkovács, L. Kovács, T. Répási, G. Szarvas, S. Szaszkó, and M. Vázsonyi. *Szövegbányászat*. Typotex, Budapest, 2007.

H. Wallach. Topic modeling: Beyond bag of words. *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

H. Wallach, D. Mimno, and A. McCallum. Rethinking lda: Why priors matter. NIPS, Vancouver, BCS, 2009a.

H. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. *Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09)*, pages 1105–1112, 2009b.

X. Wei and B. W. Croft. Lda-based document models for ad-hoc retrieval, 2006.

T. Yano, W. W. Cohen, and N. A. Smith. Predicting response to political blog posts with topic models. *HLT-NAACL 2009*, pages 477–485, 2009.

J. Zsibrita, V. Vincze, and R. Farkas. magyarlanc: A toolkit for morphological and dependency parsing of hungarian. In *Proceedings of RANLP 2013*, 2013.

FÜGGELÉK

A. A VEGYES TAGSÁGÚ MODELLEK ÁLTALÁNOS ALAKJA

Egy általános vegyes tagságú modell a feltételeknek négy szintjét tartalmazza: a populáció szintet, a szubjektum szintet, a látens változó szintet és a mintavételezési sémát.

A populációs szint feltételei leírják a populáció alapvető szerkezetét, ami minden szubjektumra érvényes. Feltételezzük, hogy a populáción belül K különböző alpopuláció létezik. Minden egyes k alpopuláció j függő változóra vonatkozó valószínűségeloszlását $f(x_j|\theta_{kj})$ írja le, ahol a θ_{kj} paraméterekből álló vektor. Emellett azt is feltételezzük, hogy egy alpopuláción belül a függő változó független a megfigyelt változóktól.

A szubjektum szint feltételezi, hogy minden egyén esetében a $\lambda = (\lambda_1, \ldots, \lambda_K)$ tagsági vektor határozza meg az egyén különböző alpopulációkhoz való tartozásának mértékét. A tagsági értékek általában ismeretlenek és ezért látens változóként kezeljük. Az egyes egyének x_j megfigyelt változóinak valószínűségeloszlását a következő feltételes valószínűség definiálja:

$$P(x_j|\lambda) = \sum_{k=1}^{K} \lambda_k f(x_j|\theta_{kj})$$

. Ez az egyenlet a generatív folyamat 2(a)i és 2(a)ii lépésének felel meg. Az x_j megfigyelt függő változók függetlenek egymástól a tagsági értékekre mint feltételre nézve, amelyet a generatív folyamat 2. lépése is definiál. Emellett a megfigyelt függő változók feltételese függetlenek az egyénektől a tagsági értékekre mint feltételre nézve.

A feltételek következő szintje a látens változók szintje, amely a generatív modell első lépésének felel meg. Ezen a szinten határozzuk meg, hogy a λ tagsági vektorokat ismeretlen fix konstansokként vagy egy bizonyos eloszlás random realizációiként kezeljük a modellben.

Ha a λ tagsági értékek fixek, de ismeretlenek, a az x_j megfigyelt változó feltételes eloszlása a θ paraméterű λ tagsági értékek feltétele mellett

$$P(x_j|\lambda;\theta) = \sum_{k=1}^{K} \lambda_k f(x_j|\theta_{kj})$$

.

Ha a λ tagsági értékek egy látens változó realizációi valamilyen D_α eloszlásból, az α vektorral paraméterezve az x_j megfigyelt változó feltételes eloszlása a paraméterekre mint feltételre nézve:

$$P(x_j|\alpha;\theta) = \int \left(\sum_{k=1}^{K} \lambda_k f(x_j|\theta_{kj}) \right) \mathrm{d}D_\alpha(\lambda)$$

A látens Dirichlet allokáció random valószínűségi változóként kezeli a topikokat, ezért az utóbbi esetre koncentrálunk.

A feltételek utolsó szintje a mintavételi séma szintje. Tegyük fel, hogy egy egyénhez J különböző jellemzőjű R független megfigyelés tartozik. Ha a tagsági értékeket egy D_α eloszlás realizációiként kezeljük, a feltételes valószínűség a következőképp alakul:

$$P\left(\{x_1^{(r)}, \ldots, x_J^{(r)}\})_{r=1}^{R}|\alpha;\theta\right) = \int \left(\prod_{j=1}^{J} \prod_{r=1}^{R} \sum_{k=1}^{K} \lambda_k f(x_j|\theta_{kj}) \right) \mathrm{d}D_\alpha(\lambda)$$

.

Ha a látens változók ismeretlen konstansok, a feltételes valószínűség:

$$P\left(\{x_1^{(r)}, \ldots, x_J^{(r)}\})_{r=1}^{R}|\alpha;\theta\right) = \prod_{j=1}^{J} \prod_{r=1}^{R} \sum_{k=1}^{K} \lambda_k f(x_j|\theta_{kj})$$

.

Általánosságban a megfigyelt J jellemzőknek nem kell ugyanannak lenniük minden egyénre, és az R megfigyelések számának sem.

B. A LÁTENS DIRICHLET ALLOKÁCIÓ POSTERIOR PREDIKTÍV ELLENŐRZÉSE

Jelen fejezetben *Mimno* és *Blei* tanulmányát szeretném bemutatni [Mimno and Blei, 2011], amelyben a látens Dirichlet allokáció azon feltételes függetlenségi feltételét ellenőrizték posterior prediktív ellenőrzés segítségével, miszerint egy $w_{d,n}$ szó független d dokumentum θ_d topikeloszlásától a $z_{d,n}$ topik hozzárendelésre, mint feltételre nézve. Azaz, ha tudjuk egy w szó z topikhozzárendelését, az, hogy melyik dokumentumban van, nem ad plusz információt a w szóra nézve. Tehát ha egy szó csak a topikhozzárendeléstől függ, akkor az egy topikhoz tartozó szavak egymástól függetlenül származnak ugyanabból a polinomiális eloszlásból. *Mimno*ék ezt a feltételt ellenőrzik azáltal, hogy kiszámolják a z topikhozzárendelés által meghatározott topikhoz tartozó szavak és az őket tartalmazó dokumentumok közötti kölcsönös információt (mutual information). Ha a feltétel igaz, a két változónak függetlennek kell lennie egymástól. Ezt az alacsony, 0-hoz közeli kölcsönös információ fejezi ki, míg a magas kölcsönös információ arra utal, hogy a megfigyelt korpusz nem illeszkedik a modellfeltételre.

A posterior prediktív ellenőrzések egyik újítása az ún. *realizált diszkrepancia függvény* (realized discrepancy function), amely a megfigyelt adatok és a modell rejtett változóinak függvénye. Ezeknek a függvényeknek a lényege, hogy a a rejtett változók kimarginalizálásával tudjuk kiértékelni a látens változókat érintő modellfeltételeket. A látens Dirichlet allokáció esetében a topikok jelentik a rejtett változókat, amelyeket úgy marginalizálunk a kölcsönös információ kiszámolása során, hogy az adott topikhoz tartozó szavak és dokumentumok adataival dolgozunk.

A kölcsönös információ egy információelméleti fogalom, amely az *entrópia* (*entropy*, H) fogalmára épül. A soron következő információelméleti definíciókat [Manning and Schütze, 1999] alapján dolgozom fel.

Az entrópia egy valószínűségi változó átlagos bizonytalanságát határozza meg, és a következőképp számolható ki X valószínűségi változó esetén:

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)}$$

Az entrópia egy valószínűségi változó információmennyiségét fejezi ki, amelyet

általában bitekben mérünk, ezért szerepel a képletben kettes alapú logaritmus.[1]

A kölcsönös információ fogalmához szükségünk van az együttes és a feltételes entrópia fogalmára is. A *kölcsönös entrópia* (*joint entropy*) két diszkrét valószínűségi változó, X és Y esetén annak az információnak az átlagos mennyisége, amelyre szükségünk van X és Y változó értékeinek meghatározásához.

$$H(X,Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{1}{p(x,y)}$$

A *feltételes entrópia* (*conditional entropy*) két diszkrét valószínűségi változó, X és Y esetében azt fejezi ki, hogy mennyi átlagos többletinformáció szükséges egyik változó ismerete mellett a másik változó ismeretéhez. Tehát ha az egyik valószínűségi változó ismert, mekkora a másik valószínűségi változó bizonytalanságának mértéke.

$$H(Y|X) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{1}{p(y|x)}$$

Két diszkrét valószínűségi változó, X és Y *kölcsönös információ*jának (*mutual information*, I) hívjuk azt az információmennyiséget, amellyel az egyik változó ismerete csökkenti a másik változó bizonytalanságát. A kölcsönös információ a két változó közös információjának szimmetrikus, nem-negatív mérőszáma. A kölcsönös információval tehát a két változó függetlenségét mérjük, tehát egyfajta asszociációs mérőszámként értelmezhető.

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(X) + H(Y) - H(X,Y) \\ &= \sum_{x\in X} p(x)\log_2\frac{1}{p(x)} + \sum_{y\in Y} p(y)\log_2\frac{1}{p(y)} + \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y) \\ &= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \end{aligned}$$

A *feltételes kölcsönös információ* két valószínűségi változó kölcsönös információja egy harmadik változóra mint feltételre nézve.

$$\begin{aligned} I(X;Y|Z) &= \sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z} p(z)p(x,y|z)\log\frac{p(x,y|z)}{p(x|z)p(y|z)} \\ &= \sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z} p(x,y,z)\log\frac{p(z)p(x,y,z)}{p(x,z)p(y,z)} \end{aligned}$$

[1] Az entrópia méréséhez tetszés szerint választhatunk alapot a logaritmushoz. A továbbiakban az alap megjelölése nélküli logaritmus kettes alapú logaritmusnak értendő.

Térjünk vissza Mimno és Blei posterior prediktv ellenőrzésére, amelyhez egy olyan diszkrepanciafüggvényt terveztek, ami ellenőrzi a szavak függetlenségét a dokumentumtól, amely tartalmazza a topik hozzárendelésre, mint feltételre nézve.

Képzeljük el egy korpusz LDA dekompozícióját, ami megjelöli az összes megfigyelt $w_{d,n}$ szót egy $z_{d,n}$ topikkal. Most korlátozzuk a figyelmünket a szavakra, amelyek hozzá vannak rendelve egy k topikhoz. Ez két valószínűségi változót formál: W-t, amely a topikhoz rendelt szavak gyűjteményét jelenti, és D-t, amely az adott topikhoz tartozó szavak dokumentumindexei. Ha az LDA feltétele igaz, akkor az, hogy tudjuk W-t, nem ad információt D-ről, mivel a szavak a topiktól függetlenül lettek kisorsolva. A függetlenséget ezért a W szavak és a D dokumentumindexek közötti feltételes kölcsönös információval mérjük k-ra, mint feltételre nézve.

$$\begin{aligned} MI(W;D|k) &= \sum_{w\in W}\sum_{d\in D} p(w,d|k)\log\frac{p(w,d|k)p(d|k)}{p(w|k)p(d|k)} \\ &= \sum_{w\in W}\sum_{d\in D}\frac{N(w,d,k)}{N(k)}\log\frac{N(w,d,k)N(k)}{N(d,k)N(w,k)} \end{aligned}$$

, ahol $N(w,d,k)$ a d dokumentumban lévő k topikhoz tartozó w szó száma, $N(w,k) = \sum_{d\in D} N(w,d,k)$ az összes dokumentumban a k topikhoz tartozó w szó száma, $N(d,k) = \sum_{w\in W} N(w,d,k)$ az összes szó száma a d dokumentumban, ami a k topikhoz tartozik, és $N(k) = \sum_{w\in W}\sum_{d\in D} N(w,d,k)$ az összes szó száma az összes dokumentumban, ami a k topikhoz tartozik. Ez a függvény megméri a szavak és dokumentumindexek feletti együttes eloszlás és a marginális eloszlások szorzata közötti különbséget. Végtelen minták esetében a független valószínűségi változók kölcsönös információja 0, és azt várjuk, hogy véges minták esetében nem-nulla értékeket kapunk ténylegesen független változók esetében is. Jegyezzük meg, hogy ez egy realizált diszkrepancia, a megfigyelt szavak topikokhoz való látens hozzárendelésétől függ.

Mimno és Blei úgy hajtották végre a posterior prediktív ellenőrzését, hogy új szavakat mintavételeztek minden egyes tokenre a topikok posterior eloszlásából a Gibbs-mintavételi állapotot figyelembe véve, és újraszámolták a rangját és a kölcsönös információját minden egyes szónak. A Gibbs-mintához tartozó állapot minden egyes szót megjelöl egy topikkal. Először megszámolták az eseteket, ahányszor egy szó az egyes topikokhoz lett rendelve, ezeket a $N(w,k)$ jelöli. Aztán minden $w_{d,n}$ szóra a korpuszban mintavételeztek egy új megfigyelt szót, $w^{rep}_{d,n}$-t, ahol $P(w) \propto N(w|z_{d,n})$. Végezetül újraszámolták a kölcsönös információt minden egyes topik minden egyes szavára. Az ezekből kirajzolódó intervallum egy referenciaeloszlást képez, ami a kölcsönös információ várható értékét adja minden rangra a polinomiális feltétel mellett.

Ha a megfigyelt adathoz tartozó kölcsönös információk eltérnek ettől az eloszlástól, azaz az eloszlás 2.5 illetve 97.5%-án kívül esnek, a modellfeltétel nem teljesül.

C. ESZKÖZÖK A LÁTENS DIRICHLET ALLOKÁCIÓ ILLESZTÉSÉHEZ

A legnépszerűbb programozási nyelvek és az adatelemzéshez használt programozási nyelvek (pl. Java, C, C++, Python, Matlab, R) nagy része lehetőséget biztosít a látens Dirichlet allokáció illesztéséhez. Ezek közül az eszközök közül a MALLET Java könyvtárat, az R három csomagját, a topicmodels és az lda csomagokat, valamint a MALLET R wrapperjét, a mallet csomagot néztem meg közelebbről.

Az R csomagok közül a topicmodels csomag kezelése volt a legkényelmesebb. A topicmodelst Bettina Grün és Kurt Hornik írták. Az LDA és CTM (Correlated Topic Model) modellek illesztéséhez írt kódokhoz a David Blei és társai által C-ben írt nyílt forráskódú implementációkat használták fel. A posterior közelítéshez két algoritmust implementáltak, a variational expectation-maximization (VEM) algoritmust és a Gibbs-mintavételt a nyílt GibbsLDA++ forráskód alapján. A topicmodels csomag előnye, hogy a szövegbányászathoz használt tm csomagra épül, ezért a szó-dokumentum mátrixot a megszokott módon hozhattam létre. Ezenkívül a csomaghoz tartozó dokumentáció is könnyűvé teszi az alkalmazást.[1] A csomag hátránya, hogy a modellillesztés relatíve sok időbe telik, valamint a modellillesztésnél nem lehetséges aszimmetrikus α illesztése. Az adatbehívás és a modellillesztés kipróbálásához használt kód a Függelék E.4 részében található. [2]

Az lda csomag, melyet David Blei tanítványa, Jonathan Chang fejlesztett, az LDA és a RTM (Related Topic Model) implementációját tartalmazza, amelyekhez ő is Bleiék C kódjait használták fel. A posterior közelítéshez a Gibbs-mintavétel C implementációját dolgozta át. Emellett az expectation maximization (EM) algoritmus is implementálva van a csomagban. Az lda csomaghoz nem tartozik a topicmodelshez hasonló dokumentáció és a korpuszbeolvasás is a megszokottól eltérő módon történik, ezért a csomag kezelését nem találtam megfelelően kényelmesnek. Emellett a modellillesztésnél nem lehetséges aszimmetrikus α illesztése.[3]

A mallet csomag a MALLET Java-alapú könyvtár wrapperje, amelyet David

[1] http://cran.r-project.org/web/packages/topicmodels/vignettes/topicmodels.pdf
[2] A leírás a topicmodels package dokumentációja alapján készült.
[3] http://cran.r-project.org/web/packages/lda/lda.pdf

Mimno írt. A csomag nem biztosít minden funkciót, amit a MALLET biztosít, azok egy része csak körülményesen vagy sehogyan sem érhetők el. Emellett a csomag szövegbeolvasáshoz kínált függvénye a magyar korpusz karakterkódolását nem tudta kezelni, ezért a tm csomaggal kellett hegeszteni, ami meghosszabbította a szövegbeolvasás futási idejét. A modellillesztés azonban gyorsabb, mint az előző két csomag esetén. A korpuszbeolvasáshoz és a modellillesztéshez használt kód a Függelék E.5 része alatt található.

A MALLET Java-alapú könyvtár bizonyult a legkönnyebben kezelhető és leggyorsabb futási idejű megoldásnak. A korpuszbeolvasás és a modellillesztés csupán néhány egyszerű parancs használatát igényli és az alap beállításokkal 3-6 perc a parancsok futási ideje.

D. PYTHON KÓDOK

D.1. collect_coloumn.py

```
#!/usr/bin/env python
import codecs
from urlparse import urljoin
from boilerpipe.extract import Extractor
from bs4 import BeautifulSoup

base = 'https://kuruc.info/to/35/'
kuruc = 'https://kuruc.info/'
urls = []
i = 10
while i != 11810:
    url = base + str(i) + '/'
    urls.append(url)
    i += 10


article_urls = []
for url in urls:
    try:
        extractor = Extractor(extractor='ArticleExtractor',
           url=url)
        extracted_html = extractor.getHTML()
        soup = BeautifulSoup(extracted_html)
        links = soup('a')
        for link in links:
            if ('href' in dict(link.attrs)):
                purl = urljoin(kuruc, link['href'])
                article_urls.append(purl)
    except:
        continue

article_urls = list(set(article_urls))

rovat = codecs.open('rovat.tsv', 'w', 'utf-8')
n = 1
for e in article_urls:
    rovat.write(str(n) + '\t' + e + '\n')
rovat.close()
```

D.2. harvest.py

```
#!/usr/bin/env python
import codecs
import urllib2
from os.path import join
from time import sleep, time

out_path = '/home/kitti/Desktop/ciganybunzes/html/'

uf = codecs.open('rovat.tsv', 'r', 'utf-8')
urls = {}
for l in uf:
    l = l.strip()
    c, url = l.split('\t')
    urls[url] = c

i = 0
h = 0


for url in urls.keys():
    try:
        t = time()
        wait = 120
        f = urllib2.urlopen(url)
        if time() > t + wait:
            continue
        else:
            print "open"
            html = f.read()
            print "read"
            fname = urls[url] + '.html'
            print "fname", fname
            of = open(join(out_path, fname), 'w')
            of.write(html)
            print "wrote"
            of.close()
            i += 1
            print i
            sleep(1)
    except:
        print "ERROR"
        h += 1
        continue
```

D.3. hate_ids.py

```
#!/usr/bin/env python
import codecs
import shutil
from os import listdir
from os.path import isfile, join

in_path = "/home/kitti/Desktop/ciganybunzes/cb/"
html_path = "/home/kitti/Desktop/ciganybunzes/html/"
hate_path = "/home/kitti/Desktop/ciganybunzes/hate_html/"
of = codecs.open('articles.tsv', 'w', 'utf-8')

txt_files = [f for f in listdir(in_path) if isfile(join(
   in_path, f))]
for f in txt_files:
    o = f + '\n'
    of.write(o)
of.close()

inf = codecs.open('articles.tsv', 'r', 'utf-8')
articles = []
for l in inf:
    l = l.strip()
    articles.append(l)

html_files = [f for f in listdir(html_path) if isfile(join(
   html_path, f))]
for f in html_files:
    if f in articles:
        shutil.copy2(join(html_path, f), join(hate_path, f))
```

D.4. *get_text.py*

```
#!/usr/bin/env python
import codecs
from os import listdir
from os.path import isfile, join
from boilerpipe.extract import Extractor

hate_path = "/home/kitti/Desktop/ciganybunzes/hate_html/"
plain_path = "/home/kitti/Desktop/ciganybunzes/hate_text/"

html_files = [f for f in listdir(hate_path) if isfile(join(
   hate_path, f))]

for f in html_files:
    inf = open(join(hate_path, f), 'r')
    txt = inf.read()
```

```
    txt = txt.decode('iso-8859-2')
    txt = txt.encode('utf-8')
    inf.close()
    try:
        extractor = Extractor(extractor='ArticleExtractor',
           html = txt)
        extracted_text = extractor.getText()
        outf = codecs.open(join(plain_path, f[:-5] + '.txt'),
           'w', 'utf-8')
        outf.write(extracted_text)
        outf.close()
    except:
        print 'failed'
        continue

print "I'm DONE!"
```

D.5. run_magyarlanc.py

```
#!/usr/bin/env python
import subprocess
from os.path import isfile, join

in_path = "/home/kitti/Desktop/ciganybunzes/hate_text/"
out_path = "/home/kitti/Desktop/ciganybunzes/processed/"


# running magyarlanc on raw corpora
i = 1
while i <= 10477:
    f = str(i) + '.txt'
    if isfile(join(in_path, f)):
        subprocess.call(
        "java -Xmx1G -jar magyarlanc-2.0.jar -mode morphparse
           -input "
        + join(in_path, f) + " -output " + join(out_path, f) +
            ".out",
        stderr=subprocess.STDOUT, shell=True)
    i += 1
print 'DONE'
```

D.6. datefinder.py

```
#!/usr/bin/env python
import codecs
```

```
import re
import dateutil.parser as dparser
from os import listdir
from os.path import isfile, join

data_path = "hate_text/"

txts = [f for f in listdir(data_path) if isfile(join(data_path
    , f))]

months = {}
f = codecs.open('months.txt', 'r', 'utf-8')
for l in f:
    l = l.strip()
    l = l.split('\t')
    m, d = l[0], l[1]
    months[m] = d


of = codecs.open('code_date.tsv', 'w', 'utf-8')

for f in txts:
    try:
        h = codecs.open(join(data_path, f), 'r', 'utf-8')
        txt = h.read()
        m = re.search(":: ((19|20)[0-9]{2}\. \D*
            ([0-9]|[0-9]{2})\.?)", txt)
        if m != None:
                n = m.group(1)
                n = n.replace('.', '')
                n = n.split(' ')
                if len(n[2]) == 1:
                    n[2] = '0' + n[2]
                if len(n) == 3:
                    h1 = n[1]
                    h2 = months[h1]
                    n[1] = h2
                    n = "-".join(n)
                    dt = dparser.parse(n, fuzzy = True)
                    ft = dt.strftime("%Y-%m-%d %H:%M:%S")
                    ft = ft.split(' ')
                    st = ft[0]
                    fid = f[:-4]
                    o = fid + '\t' + st + '\n'
                    of.write(o)
    except:
        continue
```

D.7. which_modified.py

```
#!/usr/bin/env python

from os import listdir
from os.path import isfile, join

in_path_ang = "D:/ciganybunzes/modified_out_txt/angol"
in_path_fin = "D:/ciganybunzes/modified_out_txt/finn"
in_path_fra = "D:/ciganybunzes/modified_out_txt/francia"
in_path_nem = "D:/ciganybunzes/modified_out_txt/nemet"
in_path_ol = "D:/ciganybunzes/modified_out_txt/olasz"
in_path_rom = "D:/ciganybunzes/modified_out_txt/roman"
in_path_sp = "D:/ciganybunzes/modified_out_txt/spanyol"
in_path_sv = "D:/ciganybunzes/modified_out_txt/sved"
in_path_szrb = "D:/ciganybunzes/modified_out_txt/szerb"

def get_idlist(in_path):
    idlist = []
    in_filenames = [f for f in listdir(in_path) if isfile(join
       (in_path, f))]
    for filename in in_filenames:
        fileid = int(filename[:(-8)])
        idlist.append(fileid)
    print idlist

get_idlist(in_path_ang)
get_idlist(in_path_fin)
get_idlist(in_path_fra)
get_idlist(in_path_nem)
get_idlist(in_path_ol)
get_idlist(in_path_rom)
get_idlist(in_path_sp)
get_idlist(in_path_sv)
get_idlist(in_path_szrb)
```

D.8. stem_filter.py

```
#!/usr/bin/env python
import codecs
from os import listdir
from os.path import isfile, join

out_path = "/home/kitti/Desktop/ciganybunzes/
   stemmed_filtered_4_mod/"
in_path =  "/home/kitti/Desktop/ciganybunzes/processed/"
```

```
sf = codecs.open("/home/kitti/Desktop/ciganybunzes/
    stoplist_stemmed/stoplist_stemmed_2.tsv", 'r', 'utf-8')
fd = codecs.open("/home/kitti/Desktop/ciganybunzes/freq_delete
    .csv", 'r', 'utf-8')


stoplist = ['van', 'lesz', 'kell', 'volna']
for l in sf:
    w = l.strip()
    stoplist.append(w)

for k in sf:
    w = k.strip()
    stoplist.append(w)


# normalizing text and saving results
ml_files = [f for f in listdir(in_path) if isfile(join(in_path
    , f))]
for f in ml_files:
    lf = codecs.open(join(in_path, f), 'r', 'utf-8')
    f = f[:-4]
    of = codecs.open(join(out_path, f), 'w', 'utf-8')
    stems = [] # second row, lower case
    for l in lf:
        wd = l.split("\t")
        if len(wd) > 3:
            pos = wd[2]
            pos = pos[0]
            w = wd[1]
            w = w.lower()
            if w.isalpha() and pos != 'R' and pos != 'T' and
                pos != 'C'
            and pos != 'P' and pos != 'S' and pos != 'M' and
                pos != 'I'
            and pos != 'O' and pos != 'V' and pos != 'Z'
            and w not in stoplist:
                stems.append(w)
    stems = " ".join(stems)
    of.write(stems)

# 4_mod: fonev, melleknev, ismeretlen; kiszurt: nemmagyar,
    stoplist_2, 5-nel kevesbe gyakori
```

D.9. wrong_chars.py

```
# -*- coding: utf-8 -*-
```

```
from os import listdir
from os.path import isfile, join
import codecs

in_path =  "D:/ciganybunzes/stemmed_filtered_4_mod/"
out_path = "D:/ciganybunzes/stemmed_filtered_4_mod_2/"

ts_files = [f for f in listdir(in_path) if isfile(join(in_path
    , f))]
for f in ts_files:
    text = codecs.open(join(in_path, f), 'r', 'utf-8').read()
    text = text.replace(u"õ", u"ő")
    text = text.replace(u"û", u"ű")
    of = codecs.open(join(out_path, f), 'w', 'utf-8')
    of.write(text)
    of.close()
```

D.10. textstats.py

```
#!/usr/bin/env python
# generating text stats
import codecs
import nltk
from os import listdir
from os.path import isfile, join

# globs for data path
text_path = 'D:/ciganybunzes/hate_text/'
stem_path = 'D:/ciganybunzes/stemmed_filtered_4_mod_2/'

# making a list from file ids having a timpe stamp
df = codecs.open('D:/ciganybunzes/date/code_date.tsv', 'r', '
    utf-8')
fids = []
for l in df:
    l = l.strip()
    l = l.split('\t')
    fids.append(l[0])

tt = codecs.open('D:/ciganybunzes/textstat.tsv', 'w', 'utf-8')

sent_detector = nltk.data.load('tokenizers/punkt/english.
    pickle')

# going through raw texts
txts = [f for f in listdir(text_path) if isfile(join(text_path
    , f))]
for t in txts:
```

```
    fid = t[:-4]
    # check if we have a time stamp for the file id
    if fid in fids:
        txt = codecs.open(join(text_path, t), 'r', 'utf-8').
            read()
        stem = codecs.open(join(stem_path, t), 'r', 'utf-8').
            read()
        char_no = len(txt)
        normalized_txt = [c.lower() for c in txt if c.isalpha
            ()
        or c == ' ']
        normalized_txt = ''.join(normalized_txt)
        normalized_txt = ' '.join(normalized_txt.split())
        normalized_txt_no = len(normalized_txt)
        tokens_no = len(normalized_txt.split(' '))
        types_no = len(set(normalized_txt.split(' ')))
        sent_tokens_no = len(sent_detector.tokenize(txt))
        char_stem_no = len(stem)
        tokens_stem_no = len(stem.split(' '))
        types_stem_no = len(set(stem.split(' ')))
        stats = fid + '\t' + str(char_no) + '\t' + str(
            normalized_txt_no)
        + '\t' + str(tokens_no) + '\t' + str(types_no) + '\t'
        + str(sent_tokens_no) + '\t' + str(char_stem_no) + '\t
            '
        + str(tokens_stem_no) + '\t' + '\n'
        tt.write(stats)
```

D.11. params_sym.py

```
# -*- coding: utf-8 -*-

from os import listdir
from os.path import isfile, join
import codecs

in_path =  ".../ciganybunzes/lda/Mallet/kuruc_mallet_4_1_sym/
topic_states_unzipped/"

alpha = {}
beta = {}

ts_files = [f for f in listdir(in_path) if
                                    isfile(join(in_path, f))]
for f in ts_files:
    text = codecs.open(join(in_path, f), 'r', 'utf-8')
    fname = f.split('.')
    fid = fname[2]
```

```
    i = 3
    p = '#doc source pos typeindex type topic\n'
    for l in text:
        if l == p:
            continue
        elif l.startswith('#alpha'):
            l = l.split(' ')
            alpha[fid] = l[2]
        else:
            l = l.split(' ')
            beta[fid] = l[2]
        i -= 1
        if i == 1:
            break

of = codecs.open('param.csv', 'w', 'utf-8')
for k in alpha.keys():
    o = k + '\t' + alpha[k] + '\t' + beta[k]
    of.write(o)
```

D.12. params_asym.py

```
# -*- coding: utf-8 -*-

from os import listdir
from os.path import isfile, join
import codecs

in_path =
"../ciganybunzes/lda/mallet-2.0.7/kuruc_mallet_4_1_asym/
topic_states_unzipped/"

alpha = {}
beta = {}

ts_files = [f for f in listdir(in_path) if
                                    isfile(join(in_path, f))]
for f in ts_files:
    text = codecs.open(join(in_path, f), 'r', 'utf-8')
    fname = f.split('.')
    fid = fname[2]
    i = 3
    p = '#doc source pos typeindex type topic\n'
    for l in text:
        if l == p:
            continue
        elif l.startswith('#alpha'):
            l = l.split(' ')
```

```
                alpha[fid] = l[2] + '\t' + l[3] + '\t' + l[4] +
                '\t' + l[5] + '\t' + l[6] + '\t' + l[7] + '\t' +
                l[8] + '\t' + l[9] + '\t' + l[10] + '\t' + l[11] +
                '\t' + l[12] + '\t' + l[13]+ '\t' + l[14]+ '\t' +
                l[15] + '\t' + l[16] + '\t' + l[17] + '\t' + l[18]
                + '\t' + l[19] + '\t' + l[20] + '\t' + l[21] + '\t
                    '
                + l[22] + '\t' + l[23] + '\t' + l[24] + '\t' +
                l[25] + '\t' + l[26] + '\t' + l[27] + '\t' + l[28]
                + '\t' + l[29]
            else:
                l = l.split(' ')
                beta[fid] = l[2]
            i -= 1
            if i == 1:
                break

of = codecs.open('param_4_1_asym.csv', 'w', 'utf-8')
for k in alpha.keys():
    o = k + '\t' + alpha[k] + '\t' + beta[k]
    of.write(o)
```

D.13. eval_sampling.py

```
#!/usr/bin/env python
import random
import shutil
from os import listdir
from os.path import isfile, join

from_path = "D:/ciganybunzes/hate_text/"
to_path = "D:/ciganybunzes/sample_eval/"

random.seed(251673)
ids = []
for i in range(1, 650):
    n = random.randint(1, 10476)
    ids.append(n)

ids = sorted(ids)

txt_files = [f for f in listdir(from_path) if isfile(join(
   from_path, f))]

for f in txt_files:
    if int(f[:-4]) in ids:
        shutil.copy2(join(from_path, f), join(to_path, f))
```

```
# deleted because of having no timestamp: 89, 400, 780, 966,
   1262,
 2626, 3331, 4320, 8139

# getting ids
new_files = [f for f in listdir(to_path) if isfile(join(
   to_path, f))]

new_ids = []
for f in new_files:
    new_ids.append(int(f[:-4]))

print new_ids
```

E. R KÓDOK

E.1. textstats_sum.R

```
setwd("D:/ciganybunzes")

textstat <- read.csv("textstat.tsv", header = F, sep = "\t",
   dec = ",",
                     stringsAsFactors = F, encoding = "UTF-8")

colnames(textstat) <- c("id", "charno", "normtextno", "
   tokensno",
                        "typesno", "senttokensno", "charstemno
                           ",
                        "tokenstemno", "typestemno")

textstat <- textstat[order(textstat$textid), ]

# textstats for normalized text
sum(textstat$charno, na.rm = T) # 36753335
sum(textstat$normtextno, na.rm = T) # 34853393
sum(textstat$tokensno, na.rm = T) # 4772766
sum(textstat$typesno, na.rm = T) # 3052765
sum(textstat$senttokensno, na.rm = T) # 313560

summary(textstat$charno, na.rm = T)
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 135    1290    2307    3569    4226   90300
summary(textstat$normtextno, na.rm = T)
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
# 108    1211    2183    3384    4015   84440       1
summary(textstat$tokensno, na.rm = T)
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
# 11.0   157.0   292.0   463.5   553.0 12000.0       1
summary(textstat$typesno, na.rm = T)
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
# 10.0   120.0   210.0   296.4   365.0  4950.0       1
summary(textstat$senttokensno, na.rm = T)
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
# 2.00   10.00   19.00   30.45   36.00 1100.00       1

sd(textstat$charno, na.rm = T) # 4053.192
sd(textstat$normtextno, na.rm = T) # 3865.34
sd(textstat$tokensno, na.rm = T) # 543.3783
```

```
sd(textstat$typesno, na.rm = T) # 285.2733
sd(textstat$senttokensno, na.rm = T) # 37.59377


# textstats for stemmed, filtered texts
sum(textstat$charstemno, na.rm = T) # 13673147
sum(textstat$tokenstemno, na.rm = T) # 1594606
sum(textstat$typestemno, na.rm = T) # 1142896

summary(textstat$charstemno, na.rm = T)
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
# 6     479     856    1328    1583   34200       1
summary(textstat$tokenstemno, na.rm = T)
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
# 1.0    56.0   100.0   154.8   185.0  4171.0       1
summary(textstat$typestemno, na.rm = T)
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
# 1       45      76     111     135    1987       1

sd(textstat$charstemno, na.rm = T) # 1554.166
sd(textstat$tokenstemno, na.rm = T) # 179.5022
sd(textstat$typestemno, na.rm = T) # 112.6567
```

E.2. freq_delete.R

```
require(tm)

setwd(".../ciganybunzes/")

# create a TermDocumentMatrix from the corpus
a  <-Corpus(DirSource("stemmed_filtered", encoding = "UTF-8"),
            readerControl = list(language="hun"))
tdma <-TermDocumentMatrix(a)

# get terms with highest frequency = 5
frqa <- findFreqTerms(tdma, highfreq = 5)

# open a file with utf8 encoding and write content out
con <- file("freq_delete.csv", encoding = "utf8")
write.csv(frq, file = con, quote = F, row.names = F)
```

E.3. filter_non-hungarian.R

```
setwd("/.../ciganybunzes/lda/mallet-2.0.7/kuruc_mallet_4/
   topicstate")
```

```
topic_state4 <- read.csv("topic-state_4.csv", header = T, sep
   = "", dec = ",",
                         stringsAsFactors = F, encoding = "UTF
                            -8")

docid_1 <- topic_state4$X.doc[which(topic_state4$topic == 1)]
typeid_1 <- topic_state4$typeindex[which(topic_state4$topic ==
    1)]
sourceid_1 <- topic_state4$source[which(topic_state4$topic ==
   1)]

df1 <- data.frame(docid = docid_1, typeid = typeid_1, sourceid
    = sourceid_1,
                  num = rep(1, length(docid_1)))
agr1 <- aggregate(df1$num, list(docid = df1$docid, typeid =
   df1$typeid, sourceid = sourceid_1),
                  FUN = sum)
agr1 <- agr1[order(agr1$docid), ]

unique(agr1[agr1$x > 20, ]$sourceid)


docid_16 <- topic_state4$X.doc[which(topic_state4$topic == 16)
   ]
typeid_16 <- topic_state4$typeindex[which(topic_state4$topic
   == 16)]
sourceid_16 <- topic_state4$source[which(topic_state4$topic ==
    16)]

df16 <- data.frame(docid = docid_16, typeid = typeid_16,
   sourceid = sourceid_16,
                   num = rep(1, length(docid_16)))
agr16 <- aggregate(df16$num, list(docid = df16$docid, typeid =
    df16$typeid, sourceid = sourceid_16),
                   FUN = sum)
agr16 <- agr16[order(agr16$docid), ]

unique(agr16[agr16$x > 20, ]$sourceid)


docid_20 <- topic_state4$X.doc[which(topic_state4$topic == 20)
   ]
typeid_20 <- topic_state4$typeindex[which(topic_state4$topic
   == 20)]
sourceid_20 <- topic_state4$source[which(topic_state4$topic ==
    20)]

df20 <- data.frame(docid = docid_20, typeid = typeid_20,
   sourceid = sourceid_20,
                   num = rep(1, length(docid_20)))
```

```
agr20 <- aggregate(df20$num, list(docid = df20$docid, typeid =
    df20$typeid, sourceid = sourceid_20),
                   FUN = sum)
agr20 <- agr20[order(agr20$docid), ]

unique(agr20[agr20$x > 20, ]$sourceid)


setwd("/.../ciganybunzes/lda/mallet-2.0.7/kuruc_mallet_4_mod/
   topicstate")

topic_state4 <- read.csv("topic-state_4_mod.csv", header = T,
   sep = "", dec = ",",
                            stringsAsFactors = F, encoding = "UTF
                               -8")

docid_14 <- topic_state4$X.doc[which(topic_state4$topic == 14)
   ]
typeid_14 <- topic_state4$typeindex[which(topic_state4$topic
   == 14)]
sourceid_14 <- topic_state4$source[which(topic_state4$topic ==
    14)]

df14 <- data.frame(docid = docid_14, typeid = typeid_14,
   sourceid = sourceid_14,
                   num = rep(1, length(docid_14)))
agr14 <- aggregate(df14$num, list(docid = df14$docid, typeid =
    df14$typeid, sourceid = df14$sourceid),
                   FUN = sum)
agr14 <- agr14[order(agr14$docid), ]
agr14 <- aggregate(agr14$x, list(docid = agr14$docid, sourceid
    = agr14$sourceid), FUN = sum)

agr14[agr14$x > 10, ]$sourceid
```

E.4. topicmodels.R

```
require(topicmodels)
require(tm)

setwd(".../ciganybunzes/")

# create a TermDocumentMatrix from the corpus
ptm <- proc.time()
a  <- Corpus(DirSource("stemmed_filtered_4_mod_2"))
dtm <- DocumentTermMatrix(a)
proc.time() - ptm
#    user  system elapsed
```

```
# 31.50    2.57   53.01

k = 30
burnin = 200
iter = 1000
keep = 50

ptm2 <- proc.time()
fitted <- LDA(dtm, k = k, method = "Gibbs",control = list(
   burnin = burnin, iter = iter, keep = keep) )
proc.time() - ptm2
#    user  system elapsed
# 1426.09    0.28 1428.66
```

E.5. mallet.R

```
Sys.setenv(JAVA_HOME="C:\\Program Files\\Java\\jre1.8.0_25")

require(mallet)

setwd(".../ciganybunzes")

a  <- Corpus(DirSource("stemmed_filtered_4_mod_2", encoding =
   "UTF-8"),
            readerControl = list(language="hun"))

docs <- data.frame(id = as.character(c(1:length(a))),
                   text = unlist(sapply(a, `[`, "content")),
                   stringsAsFactors = F)

mallet.instances <- mallet.import(docs$id,
                                  docs$text,
                                  stoplist.file =
                                    "stoplist_stemmed.tsv",
                                  token.regexp =
                                    "\\p{L}[\\p{L}\\p{P}]+\\p{
                                       L}")
ptm <- proc.time()
topic.model <- MalletLDA(num.topics = 30)
proc.time() - ptm
# user  system elapsed
# 0.31    0.13    1.86
```

E.6. fit_no_topics.R

A kód forrása: http://stackoverflow.com/questions/21355156/topic-models-cross-validation-with-loglikelihood-or-perplexity

```
library(topicmodels)
library(tm)

setwd("../ciganybunzes/")

# create a TermDocumentMatrix from the corpus
a  <-Corpus(DirSource("sample"))
dtm <-DocumentTermMatrix(a)

#### determining the right # of topics
harmonicMean <- function(logLikelihoods, precision=2000L) {
    library("Rmpfr")
    llMed <- median(logLikelihoods)
    as.double(llMed - log(mean(exp(-mpfr(logLikelihoods,
                                         prec = precision) +
                                    llMed))))
}

burnin = 200
iter = 1000
keep = 50

sequ <- seq(25, 30, 1)
fitted_many <- lapply(sequ, function(k)
  LDA(dtm, k = k, method = "Gibbs",
      control = list(burnin = burnin,iter = iter, keep = keep)
         ))

logLiks_many <- lapply(fitted_many, function(L)
                                    L@logLiks[-c(1:(burnin/
                                       keep))])
hm_many <- sapply(logLiks_many, function(h) harmonicMean(h))

png('topic_nums_50.png')
plot(sequ, hm_many, type = "l")
dev.off()

sequ[which.max(hm_many)] # 26
```

E.7. multiplot.R

A kód forrása: http://www.cookbook-r.com/Graphs/Multiple_graphs_on_one_page_%28ggplot

```
require(ggplot2)

# Multiple plot function

multiplot <- function(..., plotlist=NULL, file, cols=1, layout
   =NULL) {
  library(grid)

  # Make a list from the ... arguments and plotlist
  plots <- c(list(...), plotlist)

  numPlots = length(plots)

  # If layout is NULL, then use 'cols' to determine layout
  if (is.null(layout)) {
    # Make the panel
    # ncol: Number of columns of plots
    # nrow: Number of rows needed, calculated from # of cols
    layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
                     ncol = cols, nrow = ceiling(numPlots/cols
                        ))
  }

  if (numPlots==1) {
    print(plots[[1]])

  } else {
    # Set up the page
    grid.newpage()
    pushViewport(viewport(layout = grid.layout(nrow(layout),
    ncol(layout))))

    # Make each plot, in the correct location
    for (i in 1:numPlots) {
      # Get the i,j matrix positions of the regions that
      contain this subplot
      matchidx <- as.data.frame(which(layout == i, arr.ind =
         TRUE))

      print(plots[[i]], vp = viewport(layout.pos.row =
         matchidx$row,
                                      layout.pos.col =
                                         matchidx$col))
    }
  }
}
```

E.8. param_diagnostics.R

```
require(ggplot2)

setwd("../ciganybunzes")

params1 <- read.csv("param_4_2_asym.csv", header = F, sep =
"\t", dec = ".", stringsAsFactors = F, encoding = "UTF-8")

params1[seq(1, 1000, 2), 29] <- params1[seq(2, 1000, 2), 2]
params1 <- params1[-(seq(2, 1000, 2)), ]

colnames(params1) <- c("iter", "a1", "a2", "a3", "a4", "a5",
                       "a6", "a7", "a8","a9", "a10", "a11",
                       "a12", "a13", "a14", "a15", "a16",
                       "a17", "a18", "a19", "a20", "a21",
                       "a22", "a23", "a24", "a25", "a26",
                       "a27", "b")
params1 <- params1[order(params1$iter), ]
rownames(params1) <- params1$iter
params1_iter <- params1[21:500, ]


p1 <- ggplot(params1_iter, aes(x = a1)) +
  geom_histogram(binwidth=.001, colour="black", fill="white")
     +
  xlab("alpha - topik 1") + ylab("Gyakoriság") +
  ggtitle("Az 1. topikhoz tartozó alfa paraméter \n értékeinek
          hisztogramja")

p2 <- ggplot(params1_iter, aes(x = a2)) +
  geom_histogram(binwidth=.001, colour="black", fill="white")
     +
  xlab("alpha - topik 2") + ylab("Gyakoriság") +
  ggtitle("A 2. topikhoz tartozó alfa paraméter \n értékeinek
          hisztogramja")

p3 <- ggplot(params1_iter, aes(x = a3)) +
  geom_histogram(binwidth=.001, colour="black", fill="white")
     +
  xlab("alpha - topik 3") + ylab("Gyakoriság") +
  ggtitle("A 3. topikhoz tartozó alfa paraméter \n értékeinek
          hisztogramja")

p4 <- ggplot(params1_iter, aes(x = a4)) +
  geom_histogram(binwidth=.001, colour="black", fill="white")
     +
  xlab("alpha - topik 4") + ylab("Gyakoriság") +
  ggtitle("A 4. topikhoz tartozó alfa paraméter \n értékeinek
          hisztogramja")
```

```
p5 <- ggplot(params1_iter, aes(x = a5)) +
  geom_histogram(binwidth=.001, colour="black", fill="white")
     +
  xlab("alpha - topik 5") + ylab("Gyakoriság") +
  ggtitle("Az 5. topikhoz tartozó alfa paraméter \n értékeinek
          hisztogramja")

p6 <- ggplot(params1_iter, aes(x = a6)) +
  geom_histogram(binwidth=.001, colour="black", fill="white")
     +
  xlab("alpha - topik 6") + ylab("Gyakoriság") +
  ggtitle("A 6. topikhoz tartozó alfa paraméter \n értékeinek
          hisztogramja")

p7 <- ggplot(params1_iter, aes(x = a7)) +
  geom_histogram(binwidth=.001, colour="black", fill="white")
     +
  xlab("alpha - topik 7") + ylab("Gyakoriság") +
  ggtitle("A 7. topikhoz tartozó alfa paraméter \n értékeinek
          hisztogramja")

p8 <- ggplot(params1_iter, aes(x = a8)) +
  geom_histogram(binwidth=.001, colour="black", fill="white")
     +
  xlab("alpha - topik 8") + ylab("Gyakoriság") +
  ggtitle("A 8. topikhoz tartozó alfa paraméter \n értékeinek
          hisztogramja")

p9 <- ggplot(params1_iter, aes(x = a9)) +
  geom_histogram(binwidth=.001, colour="black", fill="white")
     +
  xlab("alpha - topik 9") + ylab("Gyakoriság") +
  ggtitle("A 9. topikhoz tartozó alfa paraméter \n értékeinek
          hisztogramja")

p10 <- ggplot(params1_iter, aes(x = a10)) +
  geom_histogram(binwidth=.001, colour="black", fill="white")
     +
  xlab("alpha - topik 10") + ylab("Gyakoriság") +
  ggtitle("A 10. topikhoz tartozó alfa paraméter \n értékeinek
          hisztogramja")

p11 <- ggplot(params1_iter, aes(x = a11)) +
  geom_histogram(binwidth=.001, colour="black", fill="white")
     +
  xlab("alpha - topik 11") + ylab("Gyakoriság") +
  ggtitle("A 11. topikhoz tartozó alfa paraméter \n értékeinek
          hisztogramja")

p12 <- ggplot(params1_iter, aes(x = a12)) +
```

```
  geom_histogram(binwidth=.001, colour="black", fill="white")
     +
  xlab("alpha - topik 12") + ylab("Gyakoriság") +
  ggtitle("A 12. topikhoz tartozó alfa paraméter \n értékeinek
          hisztogramja")

p13 <- ggplot(params1_iter, aes(x = a13)) +
  geom_histogram(binwidth=.001, colour="black", fill="white")
     +
  xlab("alpha - topik 13") + ylab("Gyakoriság") +
  ggtitle("A 13. topikhoz tartozó alfa paraméter \n értékeinek
          hisztogramja")

p14 <- ggplot(params1_iter, aes(x = a14)) +
  geom_histogram(binwidth=.001, colour="black", fill="white")
     +
  xlab("alpha - topik 14") + ylab("Gyakoriság") +
  ggtitle("A 14. topikhoz tartozó alfa paraméter \n értékeinek
          hisztogramja")

p15 <- ggplot(params1_iter, aes(x = a15)) +
  geom_histogram(binwidth=.001, colour="black", fill="white")
     +
  xlab("alpha - topik 15") + ylab("Gyakoriság") +
  ggtitle("A 15. topikhoz tartozó alfa paraméter \n értékeinek
          hisztogramja")

p16 <- ggplot(params1_iter, aes(x = a16)) +
  geom_histogram(binwidth=.001, colour="black", fill="white")
     +
  xlab("alpha - topik 16") + ylab("Gyakoriság") +
  ggtitle("A 16. topikhoz tartozó alfa paraméter \n értékeinek
          hisztogramja")

p17 <- ggplot(params1_iter, aes(x = a17)) +
  geom_histogram(binwidth=.001, colour="black", fill="white")
     +
  xlab("alpha - topik 17") + ylab("Gyakoriság") +
  ggtitle("A 17. topikhoz tartozó alfa paraméter \n értékeinek
          hisztogramja")

p18 <- ggplot(params1_iter, aes(x = a18)) +
  geom_histogram(binwidth=.001, colour="black", fill="white")
     +
  xlab("alpha - topik 18") + ylab("Gyakoriság") +
  ggtitle("A 18. topikhoz tartozó alfa paraméter \n értékeinek
          hisztogramja")

p19 <- ggplot(params1_iter, aes(x = a19)) +
  geom_histogram(binwidth=.001, colour="black", fill="white")
     +
```

```
  xlab("alpha - topik 19") + ylab("Gyakoriság") +
  ggtitle("A 19. topikhoz tartozó alfa paraméter \n értékeinek
          hisztogramja")

p20 <- ggplot(params1_iter, aes(x = a20)) +
  geom_histogram(binwidth=.001, colour="black", fill="white")
     +
  xlab("alpha - topik 20") + ylab("Gyakoriság") +
  ggtitle("A 20. topikhoz tartozó alfa paraméter \n értékeinek
          hisztogramja")

p21 <- ggplot(params1_iter, aes(x = a21)) +
  geom_histogram(binwidth=.001, colour="black", fill="white")
     +
  xlab("alpha - topik 21") + ylab("Gyakoriság") +
  ggtitle("A 21. topikhoz tartozó alfa paraméter \n értékeinek
          hisztogramja")

p22 <- ggplot(params1_iter, aes(x = a22)) +
  geom_histogram(binwidth=.001, colour="black", fill="white")
     +
  xlab("alpha - topik 22") + ylab("Gyakoriság") +
  ggtitle("A 22. topikhoz tartozó alfa paraméter \n értékeinek
          hisztogramja")

p23 <- ggplot(params1_iter, aes(x = a23)) +
  geom_histogram(binwidth=.001, colour="black", fill="white")
     +
  xlab("alpha - topik 23") + ylab("Gyakoriság") +
  ggtitle("A 23. topikhoz tartozó alfa paraméter \n értékeinek
          hisztogramja")

p24 <- ggplot(params1_iter, aes(x = a24)) +
  geom_histogram(binwidth=.001, colour="black", fill="white")
     +
  xlab("alpha - topik 24") + ylab("Gyakoriság") +
  ggtitle("A 24. topikhoz tartozó alfa paraméter \n értékeinek
          hisztogramja")

p25 <- ggplot(params1_iter, aes(x = a25)) +
  geom_histogram(binwidth=.001, colour="black", fill="white")
     +
  xlab("alpha - topik 25") + ylab("Gyakoriság") +
  ggtitle("A 25. topikhoz tartozó alfa paraméter \n értékeinek
          hisztogramja")

p26 <- ggplot(params1_iter, aes(x = a26)) +
  geom_histogram(binwidth=.001, colour="black", fill="white")
     +
  xlab("alpha - topik 26") + ylab("Gyakoriság") +
  ggtitle("A 26. topikhoz tartozó alfa paraméter \n értékeinek
```

```
          hisztogramja")

p27 <- ggplot(params1_iter, aes(x = a27)) +
  geom_histogram(binwidth=.001, colour="black", fill="white")
     +
  xlab("alpha - topik 27") + ylab("Gyakoriság") +
  ggtitle("A 27. topikhoz tartozó alfa paraméter \n értékeinek
          hisztogramja")

p28 <- ggplot(params1_iter, aes(x = b)) +
  geom_histogram(binwidth=.0001, colour="black", fill="white")
      +
  xlab("beta") + ylab("Gyakoriság") +
  ggtitle("A béta paraméter \n értékeinek hisztogramja")

png("param_hist_a1_a9_2.png", width = 1920, height = 960)
multiplot(p1, p2, p3, p4, p5, p6, p7, p8, p9, cols = 3)
dev.off()

png("param_hist_a10_a18_2.png", width = 1920, height = 960)
multiplot(p10, p11, p12, p13, p14, p15, p16, p17, p18, cols =
   3)
dev.off()

png("param_hist_a19_a27_2.png", width = 1920, height = 960)
multiplot(p19, p20, p21, p22, p23, p24, p25, p26, p27, cols =
   3)
dev.off()

png("param_hist_b_2.png")
p28
dev.off()


g1 <- ggplot(params1, aes(x = iter, y = a1)) +
  geom_line() +
  xlab("Iteráció száma") + ylab("Alfa paraméter értéke") +
  ggtitle("Az 1. topik alfa paraméterének konvergenciája 10
          iterációnként")

g2 <- ggplot(params1, aes(x = iter, y = a2)) +
  geom_line() +
  xlab("Iteráció száma") + ylab("Alfa paraméter értéke") +
  ggtitle("A 2. topik alfa paraméterének konvergenciája 10
          iterációnként")

g3 <- ggplot(params1, aes(x = iter, y = a3)) +
  geom_line() +
  xlab("Iteráció száma") + ylab("Alfa paraméter értéke") +
  ggtitle("A 3. topik alfa paraméterének konvergenciája 10
          iterációnként")
```

```
g4 <- ggplot(params1, aes(x = iter, y = a4)) +
  geom_line() +
  xlab("Iteráció száma") + ylab("Alfa paraméter értéke") +
  ggtitle("A 4. topik alfa paraméterének konvergenciája 10
          iterációnként")

g5 <- ggplot(params1, aes(x = iter, y = a5)) +
  geom_line() +
  xlab("Iteráció száma") + ylab("Alfa paraméter értéke") +
  ggtitle("Az 5. topik alfa paraméterének konvergenciája 10
          iterációnként")

g6 <- ggplot(params1, aes(x = iter, y = a6)) +
  geom_line() +
  xlab("Iteráció száma") + ylab("Alfa paraméter értéke") +
  ggtitle("A 6. topik alfa paraméterének konvergenciája 10
          iterációnként")

g7 <- ggplot(params1, aes(x = iter, y = a7)) +
  geom_line() +
  xlab("Iteráció száma") + ylab("Alfa paraméter értéke") +
  ggtitle("A 7. topik alfa paraméterének konvergenciája 10
          iterációnként")

g8 <- ggplot(params1, aes(x = iter, y = a8)) +
  geom_line() +
  xlab("Iteráció száma") + ylab("Alfa paraméter értéke") +
  ggtitle("A 8. topik alfa paraméterének konvergenciája 10
          iterációnként")

g9 <- ggplot(params1, aes(x = iter, y = a9)) +
  geom_line() +
  xlab("Iteráció száma") + ylab("Alfa paraméter értéke") +
  ggtitle("A 9. topik alfa paraméterének konvergenciája 10
          iterációnként")

g10 <- ggplot(params1, aes(x = iter, y = a10)) +
  geom_line() +
  xlab("Iteráció száma") + ylab("Alfa paraméter értéke") +
  ggtitle("A 10. topik alfa paraméterének konvergenciája 10
          iterációnként")

g11 <- ggplot(params1, aes(x = iter, y = a11)) +
  geom_line() +
  xlab("Iteráció száma") + ylab("Alfa paraméter értéke") +
  ggtitle("A 11. topik alfa paraméterének konvergenciája 10
          iterációnként")

g12 <- ggplot(params1, aes(x = iter, y = a12)) +
  geom_line() +
```

```
  xlab("Iteráció száma") + ylab("Alfa paraméter értéke") +
  ggtitle("A 12. topik alfa paraméterének konvergenciája 10
          iterációnként")

g13 <- ggplot(params1, aes(x = iter, y = a13)) +
  geom_line() +
  xlab("Iteráció száma") + ylab("Alfa paraméter értéke") +
  ggtitle("A 13. topik alfa paraméterének konvergenciája 10
          iterációnként")

g14 <- ggplot(params1, aes(x = iter, y = a14)) +
  geom_line() +
  xlab("Iteráció száma") + ylab("Alfa paraméter értéke") +
  ggtitle("A 14. topik alfa paraméterének konvergenciája 10
          iterációnként")

g15 <- ggplot(params1, aes(x = iter, y = a15)) +
  geom_line() +
  xlab("Iteráció száma") + ylab("Alfa paraméter értéke") +
  ggtitle("A 15. topik alfa paraméterének konvergenciája 10
          iterációnként")

g16 <- ggplot(params1, aes(x = iter, y = a16)) +
  geom_line() +
  xlab("Iteráció száma") + ylab("Alfa paraméter értéke") +
  ggtitle("A 16. topik alfa paraméterének konvergenciája 10
          iterációnként")

g17 <- ggplot(params1, aes(x = iter, y = a17)) +
  geom_line() +
  xlab("Iteráció száma") + ylab("Alfa paraméter értéke") +
  ggtitle("A 17. topik alfa paraméterének konvergenciája 10
          iterációnként")

g18 <- ggplot(params1, aes(x = iter, y = a18)) +
  geom_line() +
  xlab("Iteráció száma") + ylab("Alfa paraméter értéke") +
  ggtitle("A 18. topik alfa paraméterének konvergenciája 10
          iterációnként")

g19 <- ggplot(params1, aes(x = iter, y = a19)) +
  geom_line() +
  xlab("Iteráció száma") + ylab("Alfa paraméter értéke") +
  ggtitle("A 19. topik alfa paraméterének konvergenciája 10
          iterációnként")

g20 <- ggplot(params1, aes(x = iter, y = a20)) +
  geom_line() +
  xlab("Iteráció száma") + ylab("Alfa paraméter értéke") +
  ggtitle("A 20. topik alfa paraméterének konvergenciája 10
          iterációnként")
```

```
g21 <- ggplot(params1, aes(x = iter, y = a21)) +
  geom_line() +
  xlab("Iteráció száma") + ylab("Alfa paraméter értéke") +
  ggtitle("A 21. topik alfa paraméterének konvergenciája 10
          iterációnként")

g22 <- ggplot(params1, aes(x = iter, y = a22)) +
  geom_line() +
  xlab("Iteráció száma") + ylab("Alfa paraméter értéke") +
  ggtitle("A 22. topik alfa paraméterének konvergenciája 10
          iterációnként")

g23 <- ggplot(params1, aes(x = iter, y = a23)) +
  geom_line() +
  xlab("Iteráció száma") + ylab("Alfa paraméter értéke") +
  ggtitle("A 23. topik alfa paraméterének konvergenciája 10
          iterációnként")

g24 <- ggplot(params1, aes(x = iter, y = a24)) +
  geom_line() +
  xlab("Iteráció száma") + ylab("Alfa paraméter értéke") +
  ggtitle("A 24. topik alfa paraméterének konvergenciája 10
          iterációnként")

g25 <- ggplot(params1, aes(x = iter, y = a25)) +
  geom_line() +
  xlab("Iteráció száma") + ylab("Alfa paraméter értéke") +
  ggtitle("A 25. topik alfa paraméterének konvergenciája 10
          iterációnként")

g26 <- ggplot(params1, aes(x = iter, y = a26)) +
  geom_line() +
  xlab("Iteráció száma") + ylab("Alfa paraméter értéke") +
  ggtitle("A 26. topik alfa paraméterének konvergenciája 10
          iterációnként")

g27 <- ggplot(params1, aes(x = iter, y = a27)) +
  geom_line() +
  xlab("Iteráció száma") + ylab("Alfa paraméter értéke") +
  ggtitle("A 27. topik alfa paraméterének konvergenciája 10
          iterációnként")

g28 <- ggplot(params1, aes(x = iter, y = b)) +
  geom_line() +
  xlab("Iteráció száma") + ylab("Béta paraméter értéke") +
  ggtitle("A béta paraméter konvergenciája 10 iterációnként")

# call the multiplot function in multiplot.R before using

png("param_trace_a1_a9_2.png", width = 1920, height = 960)
```

```
multiplot(g1, g2, g3, g4, g5, g6, g7, g8, g9, cols = 3)
dev.off()

png("param_trace_a10_a18_2.png", width = 1920, height = 960)
multiplot(g10, g11, g12, g13, g14, g15, g16, g17, g18, cols =
   3)
dev.off()

png("param_trace_a19_a27_2.png", width = 1920, height = 960)
multiplot(g19, g20, g21, g22, g23, g24, g25, g26, g27, cols =
   3)
dev.off()

png("param_trace_b_2.png")
g28
dev.off()
```

E.9. ppcheck.R

```
require(parallel)

create_dfs <- function(topic_number, rep_number)
{
  # extract documents and wordtypes which belong to topic
     topic_number
  docid <- topic_state$X.doc[which(topic_state$topic == topic_
     number)]
  typeid <- topic_state$typeindex[which(topic_state$topic ==
     topic_number)]

  # replicate samples from typeid, N = rep_number
  reps <- replicate(rep_number, sample(typeid, length(docid),
     replace = T))

  # create a dataframe from docid, original typeid and
     replicated typeids
  df <- data.frame(docid = docid, typeid = typeid, num = rep
     (1, length(docid)))
  df <- aggregate(df$num, list(docid = df$docid, typeid = df$
     typeid), FUN = sum)
  df <- df[order(df$docid), ]
  df_reps <- apply(reps, 2, function(x) data.frame(docid =
     docid, typeid_rep = x, num = rep(1, length(docid))))

  for (i in 1:rep_number){
    df_reps[[i]] <- aggregate(df_reps[[i]]$num, list(docid =
       df_reps[[i]]$docid, typeid_rep = df_reps[[i]]$typeid_
       rep), FUN = sum)
    df_reps[[i]] <- df_reps[[i]][order(df_reps[[i]]$docid), ]}
```

```
  dfs <- c(list(df), df_reps)

  # return df
  dfs
}


calculate_mi <- function(df)
{
  # calculating mutual information for all document-word pairs
  n_k <- sum(df$x)
  cl <- makeCluster(detectCores())
  n_docs <- parSapply(cl, df$docid, function(y) sum(df$x[which
     (df$docid == y)]))
  n_words <- parSapply(cl, df$typeid, function(y) sum(df$x[
     which(df$typeid == y)]))
  df <- data.frame(df, list(n_docs = n_docs, n_words = n_words
     ))
  mi <- parRapply(cl, df, function(y) y[3] / n_k * log2((y[3]
     * n_k) / (y[4] * y[5])))
  stopCluster(cl)

  # concetanate mutual information to dataframe
  df <- data.frame(df, list(mi = mi))

  # returning dataframe
  df
}

calculate_mis <- function(dfs)
{
  lapply(dfs, calculate_mi)
}


getting_key_ids <- function(topic_number, key_number)
{
  keys <- unlist(strsplit(topic_keys[topic_number + 1, 3], " "
     ))[1:key_number]
  key_ids <- topic_state$typeindex[sapply(keys, function(x)
     which(topic_state$type == x)[1])]
  key_ids
}
```

E.10. ppcheck_4_1_sym.R

```
require(ggplot2)
```

```
setwd("D:/ciganybunzes/lda/mallet-2.0.7/kuruc_mallet_4_1_sym")

topic_state <- read.csv("topic-state_1_sym.txt", header = T,
                        sep = " ", dec = ",",
                        stringsAsFactors = F, encoding = "UTF
                           -8")

topic_keys <- read.csv("topic-keys_1.csv", header = F,
                       sep = "\t", dec = ",",
                       stringsAsFactors = F, encoding = "UTF-8
                          ")


set.seed(2817)
sample(c(0:30), 6) # 8  0  3 25 13 16

# calculating real mutual information for topic 0 with 10
# replicates
df0 <- create_dfs(0, 100)
mi0 <- calculate_mis(df0)

# summing for documents --> getting mutual informations for
# uinque wordtypes
mi_uniquetype0 <- lapply(mi0, function(x)
  aggregate(x, list(typeid = x$typeid), sum))
# summing for words --> getting mutual information for the
# whole topic


df3 <- create_dfs(3, 100)
mi3 <- calculate_mis(df3)
mi_uniquetype3 <- lapply(mi3, function(x)
  aggregate(x, list(typeid = x$typeid), sum))


df8 <- create_dfs(8, 100)
mi8 <- calculate_mis(df8)
mi_uniquetype8 <- lapply(mi8, function(x)
  aggregate(x, list(typeid = x$typeid), sum))


df13 <- create_dfs(13, 100)
mi13 <- calculate_mis(df13)
mi_uniquetype13 <- lapply(mi13, function(x)
  aggregate(x, list(typeid = x$typeid), sum))


df16 <- create_dfs(16, 100)
mi16 <- calculate_mis(df16)
mi_uniquetype16 <- lapply(mi16, function(x)
  aggregate(x, list(typeid = x$typeid), sum))
```

```
df25 <- create_dfs(25, 100)
mi25 <- calculate_mis(df25)
mi_uniquetype25 <- lapply(mi25, function(x)
  aggregate(x, list(typeid = x$typeid), sum))


mi_uniquetype <- mi_uniquetype25

key_ids <- getting_key_ids(25, 10)

t_1 <- sapply(mi_uniquetype, function(y)
  y[which(y[1] == key_ids[1]), 7])
t_2 <- sapply(mi_uniquetype, function(y)
  y[which(y[1] == key_ids[2]), 7])
t_3 <- sapply(mi_uniquetype, function(y)
  y[which(y[1] == key_ids[3]), 7])
t_4 <- sapply(mi_uniquetype, function(y)
  y[which(y[1] == key_ids[4]), 7])
t_5 <- sapply(mi_uniquetype, function(y)
  y[which(y[1] == key_ids[5]), 7])
t_6 <- sapply(mi_uniquetype, function(y)
  y[which(y[1] == key_ids[6]), 7])
t_7 <- sapply(mi_uniquetype, function(y)
  y[which(y[1] == key_ids[7]), 7])
t_8 <- sapply(mi_uniquetype, function(y)
  y[which(y[1] == key_ids[8]), 7])
t_9 <- sapply(mi_uniquetype, function(y)
  y[which(y[1] == key_ids[9]), 7])
t_10 <- sapply(mi_uniquetype, function(y)
  y[which(y[1] == key_ids[10]), 7])


p1 <- qplot(t_1[2:101]) +
  geom_histogram(binwidth = 0.001, colour="black", fill="white
     ") +
  xlab("Kölcsönös információ értéke") + ylab("Gyakoriság") +
  ggtitle("A 25. topikhoz tartozó 1. kulcsszó \n posterior
          pediktív ellenőrzése") +
  geom_vline(xintercept = t_1[1], linetype="dashed", size=1,
             colour="red")

p2 <- qplot(t_2[2:101]) +
  geom_histogram(binwidth = 0.001, colour="black", fill="white
     ") +
  xlab("Kölcsönös információ értéke") + ylab("Gyakoriság") +
  ggtitle("A 25. topikhoz tartozó 2. kulcsszó \n posterior
          pediktív ellenőrzése") +
  geom_vline(xintercept = t_2[1], linetype="dashed", size=1,
             colour="red")
```

```
p3 <- qplot(t_3[2:101]) +
  geom_histogram(binwidth = 0.001, colour="black", fill="white
     ") +
  xlab("Kölcsönös információ értéke") + ylab("Gyakoriság") +
  ggtitle("A 25. topikhoz tartozó 3. kulcsszó \n pposterior
          pediktív ellenőrzése") +
  geom_vline(xintercept = t_3[1], linetype="dashed", size=1,
             colour="red")

p4 <- qplot(t_4[2:101]) +
  geom_histogram(binwidth = 0.001, colour="black", fill="white
     ") +
  xlab("Kölcsönös információ értéke") + ylab("Gyakoriság") +
  ggtitle("A 25. topikhoz tartozó 4. kulcsszó \n posterior
          pediktív ellenőrzése") +
  geom_vline(xintercept = t_4[1], linetype="dashed", size=1,
             colour="red")

p5 <- qplot(t_5[2:101]) +
  geom_histogram(binwidth = 0.001, colour="black", fill="white
     ") +
  xlab("Kölcsönös információ értéke") + ylab("Gyakoriság") +
  ggtitle("A 25. topikhoz tartozó 5. kulcsszó \n posterior
          pediktív ellenőrzése") +
  geom_vline(xintercept = t_5[1], linetype="dashed", size=1,
             colour="red")

p6 <- qplot(t_6[2:101]) +
  geom_histogram(binwidth = 0.001, colour="black", fill="white
     ") +
  xlab("Kölcsönös információ értéke") + ylab("Gyakoriság") +
  ggtitle("A 25. topikhoz tartozó 6. kulcsszó \n posterior
          pediktív ellenőrzése") +
  geom_vline(xintercept = t_6[1], linetype="dashed", size=1,
             colour="red")

p7 <- qplot(t_7[2:101]) +
  geom_histogram(binwidth = 0.001, colour="black", fill="white
     ") +
  xlab("Kölcsönös információ értéke") + ylab("Gyakoriság") +
  ggtitle("A 25. topikhoz tartozó 7. kulcsszó \n posterior
          pediktív ellenőrzése") +
  geom_vline(xintercept = t_7[1], linetype="dashed", size=1,
             colour="red")

p8 <- qplot(t_8[2:101]) +
  geom_histogram(binwidth = 0.001, colour="black", fill="white
     ") +
  xlab("Kölcsönös információ értéke") + ylab("Gyakoriság") +
  ggtitle("A 25. topikhoz tartozó 8. kulcsszó \n posterior
```

```
          pediktív ellenőrzése") +
  geom_vline(xintercept = t_8[1], linetype="dashed", size=1,
             colour="red")

p9 <- qplot(t_9[2:101]) +
  geom_histogram(binwidth = 0.001, colour="black", fill="white
     ") +
  xlab("Kölcsönös információ értéke") + ylab("Gyakoriság") +
  ggtitle("A 25. topikhoz tartozó 9. kulcsszó \n posterior
          pediktív ellenőrzése") +
  geom_vline(xintercept = t_9[1], linetype="dashed", size=1,
             colour="red")

p10 <- qplot(t_10[2:101]) +
  geom_histogram(binwidth = 0.001, colour="black", fill="white
     ") +
  xlab("Kölcsönös információ értéke") + ylab("Gyakoriság") +
  ggtitle("A 25. topikhoz tartozó 10. kulcsszó \n posterior
          pediktív ellenőrzése") +
  geom_vline(xintercept = t_10[1], linetype="dashed", size=1,
             colour="red")

# call the multiplot function in multiplot.R before using

png("ppc_sym_1_t25.png", width = 1920, height = 960)
multiplot(p1, p2, p3, p4, p5, p6, p7, p8, p9, p10, cols = 5)
dev.off()
```

E.11. ppcheck_4_1_asym.R

```
require(ggplot2)

setwd(".../ciganybunzes/lda/mallet-2.0.7/kuruc_mallet_4_1_asym
   ")

topic_state <- read.csv("topic-state_1_asym.txt", header = T,
                        sep = " ", dec = ",",
                        stringsAsFactors = F, encoding = "UTF
                           -8")

topic_keys <- read.csv("topic-keys_1.csv", header = F,
                       sep = "\t", dec = ",",
                       stringsAsFactors = F, encoding = "UTF-8
                          ")


set.seed(23086)
sample(c(0:27), 6) #3 12 15 14  2 16
```

```
df2 <- create_dfs(2, 100)
mi2 <- calculate_mis(df2)
mi_uniquetype2 <- lapply(mi2, function(x)
  aggregate(x, list(typeid = x$typeid), sum))


df3 <- create_dfs(3, 100)
mi3 <- calculate_mis(df3)
mi_uniquetype3 <- lapply(mi3, function(x)
  aggregate(x, list(typeid = x$typeid), sum))


df12 <- create_dfs(12, 100)
mi12 <- calculate_mis(df12)
mi_uniquetype12 <- lapply(mi12, function(x)
  aggregate(x, list(typeid = x$typeid), sum))


df14 <- create_dfs(14, 100)
mi14 <- calculate_mis(df14)
mi_uniquetype14 <- lapply(mi14, function(x)
  aggregate(x, list(typeid = x$typeid), sum))


df15 <- create_dfs(15, 100)
mi15 <- calculate_mis(df15)
mi_uniquetype15 <- lapply(mi15, function(x)
  aggregate(x, list(typeid = x$typeid), sum))


df16 <- create_dfs(16, 100)
mi16 <- calculate_mis(df16)
mi_uniquetype16 <- lapply(mi16, function(x)
  aggregate(x, list(typeid = x$typeid), sum))


mi_uniquetype <- mi_uniquetype16

key_ids <- getting_key_ids(16, 10)

t_1 <- sapply(mi_uniquetype, function(y) y[which(y[1] ==
key_ids[1]), 7])
t_2 <- sapply(mi_uniquetype, function(y) y[which(y[1] ==
key_ids[2]), 7])
t_3 <- sapply(mi_uniquetype, function(y) y[which(y[1] ==
key_ids[3]), 7])
t_4 <- sapply(mi_uniquetype, function(y) y[which(y[1] ==
key_ids[4]), 7])
t_5 <- sapply(mi_uniquetype, function(y) y[which(y[1] ==
```

```
key_ids[5]), 7])
t_6 <- sapply(mi_uniquetype, function(y) y[which(y[1] ==
key_ids[6]), 7])
t_7 <- sapply(mi_uniquetype, function(y) y[which(y[1] ==
key_ids[7]), 7])
t_8 <- sapply(mi_uniquetype, function(y) y[which(y[1] ==
key_ids[8]), 7])
t_9 <- sapply(mi_uniquetype, function(y) y[which(y[1] ==
key_ids[9]), 7])
t_10 <- sapply(mi_uniquetype, function(y) y[which(y[1] ==
key_ids[10]), 7])


p1 <- qplot(t_1[2:101]) +
  geom_histogram(binwidth = 0.001, colour="black", fill="white
     ") +
  xlab("Kölcsönös információ értéke") + ylab("Gyakoriság") +
  ggtitle("A 16. topikhoz tartozó 1. kulcsszó \n posterior
          pediktív ellenőrzése") +
  geom_vline(xintercept = t_1[1], linetype="dashed", size=1,
             colour="red")

p2 <- qplot(t_2[2:101]) +
  geom_histogram(binwidth = 0.001, colour="black", fill="white
     ") +
  xlab("Kölcsönös információ értéke") + ylab("Gyakoriság") +
  ggtitle("A 16. topikhoz tartozó 2. kulcsszó \n posterior
          pediktív ellenőrzése") +
  geom_vline(xintercept = t_2[1], linetype="dashed", size=1,
             colour="red")

p3 <- qplot(t_3[2:101]) +
  geom_histogram(binwidth = 0.001, colour="black", fill="white
     ") +
  xlab("Kölcsönös információ értéke") + ylab("Gyakoriság") +
  ggtitle("A 16. topikhoz tartozó 3. kulcsszó \n pposterior
          pediktív ellenőrzése") +
  geom_vline(xintercept = t_3[1], linetype="dashed", size=1,
             colour="red")

p4 <- qplot(t_4[2:101]) +
  geom_histogram(binwidth = 0.001, colour="black", fill="white
     ") +
  xlab("Kölcsönös információ értéke") + ylab("Gyakoriság") +
  ggtitle("A 16. topikhoz tartozó 4. kulcsszó \n posterior
          pediktív ellenőrzése") +
  geom_vline(xintercept = t_4[1], linetype="dashed", size=1,
             colour="red")

p5 <- qplot(t_5[2:101]) +
```

```
  geom_histogram(binwidth = 0.001, colour="black", fill="white
     ") +
  xlab("Kölcsönös információ értéke") + ylab("Gyakoriság") +
  ggtitle("A 16. topikhoz tartozó 5. kulcsszó \n posterior
          pediktív ellenőrzése") +
  geom_vline(xintercept = t_5[1], linetype="dashed", size=1,
             colour="red")

p6 <- qplot(t_6[2:101]) +
  geom_histogram(binwidth = 0.001, colour="black", fill="white
     ") +
  xlab("Kölcsönös információ értéke") + ylab("Gyakoriság") +
  ggtitle("A 16. topikhoz tartozó 6. kulcsszó \n posterior
          pediktív ellenőrzése") +
  geom_vline(xintercept = t_6[1], linetype="dashed", size=1,
             colour="red")

p7 <- qplot(t_7[2:101]) +
  geom_histogram(binwidth = 0.001, colour="black", fill="white
     ") +
  xlab("Kölcsönös információ értéke") + ylab("Gyakoriság") +
  ggtitle("A 16. topikhoz tartozó 7. kulcsszó \n posterior
          pediktív ellenőrzése") +
  geom_vline(xintercept = t_7[1], linetype="dashed", size=1,
             colour="red")

p8 <- qplot(t_8[2:101]) +
  geom_histogram(binwidth = 0.001, colour="black", fill="white
     ") +
  xlab("Kölcsönös információ értéke") + ylab("Gyakoriság") +
  ggtitle("A 16. topikhoz tartozó 8. kulcsszó \n posterior
          pediktív ellenőrzése") +
  geom_vline(xintercept = t_8[1], linetype="dashed", size=1,
             colour="red")

p9 <- qplot(t_9[2:101]) +
  geom_histogram(binwidth = 0.001, colour="black", fill="white
     ") +
  xlab("Kölcsönös információ értéke") + ylab("Gyakoriság") +
  ggtitle("A 16. topikhoz tartozó 9. kulcsszó \n posterior
          pediktív ellenőrzése") +
  geom_vline(xintercept = t_9[1], linetype="dashed", size=1,
             colour="red")

p10 <- qplot(t_10[2:101]) +
  geom_histogram(binwidth = 0.001, colour="black", fill="white
     ") +
  xlab("Kölcsönös információ értéke") + ylab("Gyakoriság") +
  ggtitle("A 16. topikhoz tartozó 10. kulcsszó \n posterior
          pediktív ellenőrzése") +
  geom_vline(xintercept = t_10[1], linetype="dashed", size=1,
```

```
              colour="red")

# call the multiplot function in multiplot.R before using

png("ppc_asym_1_t16.png", width = 1920, height = 960)
multiplot(p1, p2, p3, p4, p5, p6, p7, p8, p9, p10, cols = 5)
dev.off()
```

E.12. doctopics_analysis.R

```
setwd(".../ciganybunzes/lda/mallet-2.0.7/kuruc_mallet_4_2_asym
   ")

doctop <- read.csv("doc-topics_1.csv", header = F, sep = "\t",
    dec = ".", stringsAsFactors = F, encoding = "UTF-8")

doctop_topic <- function(topicnumber){
  id1 <-which(doctop$V3 == topicnumber)
  id2 <-which(doctop$V5 == topicnumber)
  id3 <-which(doctop$V7 == topicnumber)
  id4 <-which(doctop$V9 == topicnumber)
  id5 <-which(doctop$V11 == topicnumber)

  doctop_t2 <- doctop[id2, c(1,2,5,6)]
  colnames(doctop_t2) <- c("V1", "V2", "V3", "V4")
  doctop_t3 <- doctop[id3, c(1,2,7,8)]
  colnames(doctop_t3) <- c("V1", "V2", "V3", "V4")
  doctop_t4 <- doctop[id4, c(1,2,9,10)]
  colnames(doctop_t4) <- c("V1", "V2", "V3", "V4")
  doctop_t5 <- doctop[id5, c(1,2,11,12)]
  colnames(doctop_t5) <- c("V1", "V2", "V3", "V4")

  doctop_t <- rbind(doctop[id1, 1:4], doctop_t2, doctop_t3,
     doctop_t4, doctop_t5)

  doctop_t <- doctop_t[order(doctop_t$V4, decreasing = T), ]

  doctop_t
}


doctop_0 <-doctop_topic(0)
nrow(doctop_0)/nrow(doctop)  # 0.06457209

doctop_1 <-doctop_topic(1)
nrow(doctop_1)/nrow(doctop) # 0.1305898

doctop_2 <-doctop_topic(2)
```

```
nrow(doctop_2)/nrow(doctop) # 0.06919815

doctop_3 <-doctop_topic(3)
nrow(doctop_3)/nrow(doctop)  # 0.2854665

doctop_4 <-doctop_topic(4)
nrow(doctop_4)/nrow(doctop) # 0.1838859

doctop_5 <-doctop_topic(5)
nrow(doctop_5)/nrow(doctop)  # 0.08914803

doctop_6 <-doctop_topic(6)
nrow(doctop_6)/nrow(doctop) # 0.08741326

doctop_7 <-doctop_topic(7)
nrow(doctop_7)/nrow(doctop) # 0.3022359

doctop_8 <-doctop_topic(8)
nrow(doctop_8)/nrow(doctop) # 0.09724364

doctop_9 <-doctop_topic(9)
nrow(doctop_9)/nrow(doctop) # 0.22234

doctop_10 <-doctop_topic(10)
nrow(doctop_10)/nrow(doctop) # 0.06476484

doctop_11 <-doctop_topic(11)
nrow(doctop_11)/nrow(doctop) # 0.2843099

doctop_12 <-doctop_topic(12)
nrow(doctop_12)/nrow(doctop) # 0.02178103

doctop_13 <-doctop_topic(13)
nrow(doctop_13)/nrow(doctop) # 0.1087124

doctop_14 <-doctop_topic(14)
nrow(doctop_14)/nrow(doctop) # 0.1222051

doctop_15 <-doctop_topic(15)
nrow(doctop_15)/nrow(doctop) # 0.0401889

doctop_16 <-doctop_topic(16)
nrow(doctop_16)/nrow(doctop) # 0.08259445

doctop_17 <-doctop_topic(17)
nrow(doctop_17)/nrow(doctop) # 0.2499036

doctop_18 <-doctop_topic(18)
nrow(doctop_18)/nrow(doctop) # 0.03132228

doctop_19 <-doctop_topic(19)
```

```
nrow(doctop_19)/nrow(doctop) # 0.02091365

doctop_20 <-doctop_topic(20)
nrow(doctop_20)/nrow(doctop) # 0.1153624

doctop_21 <-doctop_topic(21)
nrow(doctop_21)/nrow(doctop) # 0.02505783

doctop_22 <-doctop_topic(22)
nrow(doctop_22)/nrow(doctop) # 0.08577487

doctop_23 <-doctop_topic(23)
nrow(doctop_23)/nrow(doctop) # 0.02081727

doctop_24 <-doctop_topic(24)
nrow(doctop_24)/nrow(doctop) # 0.03016577

doctop_25 <-doctop_topic(25)
nrow(doctop_25)/nrow(doctop) # 0.07314958

doctop_26 <-doctop_topic(26)
nrow(doctop_26)/nrow(doctop) # 0.04481496
```

E.13. topics_in_time.R

```
require(ggplot2)


setwd("../ciganybunzes/lda/mallet-2.0.7/kuruc_mallet_4_2_asym
   ")

doctop <- read.csv("doc-topics_1.csv", header = F, sep = "\t",
    dec = ".", stringsAsFactors = F, encoding = "UTF-8")
doctop <- doctop[- which(doctop[, 1] < 1)[-1], ]


setwd("../ciganybunzes/metadata")

date <- read.csv("code_date.tsv", header = F, sep = "\t", dec
   = ".", stringsAsFactors = F, encoding = "UTF-8")

id <- as.numeric(apply(doctop, 1,  function(x)
    unlist(strsplit(strsplit(x[2], "/")[[1]][7], ".txt"))))

doctop <- cbind(doctop, id)

matching <- match(doctop$id, date[, 1])
```

```
doctop <- cbind(doctop, date = date[matching, 2])
doctop <- doctop[-(which(is.na(doctop$date))), ]


doctop$date <- as.character(doctop$date)
doctop$date <- apply(doctop, 1, function(x) unlist(strsplit(x
   [15], "'"))[2])
doctop$date <- as.Date(doctop$date)

date2 <- apply(doctop, 1, function(x) substring(as.character(x
   [15]), 1, 7))


monthly_topic <- function(topicnumber){
  id1 <-which(doctop$V3 == topicnumber)
  id2 <-which(doctop$V5 == topicnumber)
  id3 <-which(doctop$V7 == topicnumber)
  id4 <-which(doctop$V9 == topicnumber)
  id5 <-which(doctop$V11 == topicnumber)

  doctop_t1 <- doctop[id1, c(3, 4, 14, 15)]
  colnames(doctop_t1) <- c("V3", "V4", "id", "date")
  doctop_t2 <- doctop[id2, c(5, 6, 14, 15)]
  colnames(doctop_t2) <- c("V3", "V4", "id", "date")
  doctop_t3 <- doctop[id3, c(7, 8, 14, 15)]
  colnames(doctop_t3) <- c("V3", "V4", "id", "date")
  doctop_t3 <- doctop[id3, c(9, 10, 14, 15)]
  colnames(doctop_t3) <- c("V3", "V4", "id", "date")
  doctop_t3 <- doctop[id3, c(11, 12, 14, 15)]
  colnames(doctop_t3) <- c("V3", "V4", "id", "date")

  doctop_t <- rbind(doctop_t1, doctop_t2, doctop_t3)

  doctop_t <- doctop_t[order(doctop_t$date, decreasing = F), ]
  monthly_t <- as.data.frame(aggregate(doctop_t$V4, list(
     format(doctop_t$date,"%Y-%m")), FUN = sum, na.rm = T))
  monthly_t[, 1] <- as.Date(paste0(monthly_t[, 1], "-01"))

  monthly_t
}


doctop_0 <- monthly_topic(0)
g1 <- ggplot(doctop_0, aes(x=Group.1, y=x)) + geom_line() +
  xlab("Év") + ylab("A topikarányok összege az időszak
     cikkeiben") +
  ggtitle("EU roma külpolitika")

doctop_1 <- monthly_topic(1)
g2 <- ggplot(doctop_1, aes(x=Group.1, y=x)) + geom_line() +
```

```
  xlab("Év") + ylab("A topikarányok összege az időszak
     cikkeiben") +
  ggtitle("Magyar Gárda, Jobbik, Magyar Önvédelmi Mozgalom,
          \n Nemzeti Őrsereg, Magyar Nemzeti Front rendezvé
             nyei")

doctop_2 <- monthly_topic(2)
g3 <- ggplot(doctop_2, aes(x=Group.1, y=x)) + geom_line() +
  xlab("Év") + ylab("A topikarányok összege az időszak
     cikkeiben") +
  ggtitle("Roma önkormányzat, önszerveződés")

doctop_3 <- monthly_topic(3)
g4 <- ggplot(doctop_3, aes(x=Group.1, y=x)) + geom_line() +
  xlab("Év") + ylab("A topikarányok összege az időszak
     cikkeiben") +
  ggtitle("Lopással kapcsolatos hírek")

doctop_4 <- monthly_topic(4)
g5 <- ggplot(doctop_4, aes(x=Group.1, y=x)) + geom_line() +
  xlab("Év") + ylab("A topikarányok összege az időszak
     cikkeiben") +
  ggtitle("Roma-nem roma társadalmi problémák, előítéletesség"
     )

doctop_5 <- monthly_topic(5)
g6 <- ggplot(doctop_5, aes(x=Group.1, y=x)) + geom_line() +
  xlab("Év") + ylab("A topikarányok összege az időszak
     cikkeiben") +
  ggtitle("Ingatlanüggyel, lakhatással kapcsolatos problémák,
          bűntények")

doctop_6 <- monthly_topic(6)
g7 <- ggplot(doctop_6, aes(x=Group.1, y=x)) + geom_line() +
  xlab("Év") + ylab("A topikarányok összege az időszak
     cikkeiben") +
  ggtitle("Politika, pártok, politikusok")

doctop_7 <- monthly_topic(7)
g8 <- ggplot(doctop_7, aes(x=Group.1, y=x)) + geom_line() +
  xlab("Év") + ylab("A topikarányok összege az időszak
     cikkeiben") +
  ggtitle("Máshonnan átvett tartalmak kritikája, értékelése")

doctop_8 <- monthly_topic(8)
g9 <- ggplot(doctop_8, aes(x=Group.1, y=x)) + geom_line() +
  xlab("Év") + ylab("A topikarányok összege az időszak
     cikkeiben") +
  ggtitle("Vidéki települések roma-többségi konfliktusai")

doctop_9 <- monthly_topic(9)
```

```
g10 <- ggplot(doctop_9, aes(x=Group.1, y=x)) + geom_line() +
  xlab("Év") + ylab("A topikarányok összege az időszak
     cikkeiben") +
  ggtitle("Közbiztonság, önvédelem, polgárőrség")

doctop_10 <- monthly_topic(10)
g11 <- ggplot(doctop_10, aes(x=Group.1, y=x)) + geom_line() +
  xlab("Év") + ylab("A topikarányok összege az időszak
     cikkeiben") +
  ggtitle("Egészségügy")

doctop_11 <- monthly_topic(11)
g12 <- ggplot(doctop_11, aes(x=Group.1, y=x)) + geom_line() +
  xlab("Év") + ylab("A topikarányok összege az időszak
     cikkeiben") +
  ggtitle("Olvasói történetek")


doctop_12 <- monthly_topic(12)
g13 <- ggplot(doctop_12, aes(x=Group.1, y=x)) + geom_line() +
  xlab("Év") + ylab("A topikarányok összege az időszak
     cikkeiben") +
  ggtitle("Fa- és fémlopás okozta árvizek, önvédelem")


doctop_13 <- monthly_topic(13)
g14 <- ggplot(doctop_13, aes(x=Group.1, y=x)) + geom_line() +
  xlab("Év") + ylab("A topikarányok összege az időszak
     cikkeiben") +
  ggtitle("Lopások, rongálások okozta közlekedési fennakadások
     ,
          \n lopások templomból, temetőből")


doctop_14 <- monthly_topic(14)
g15 <- ggplot(doctop_14, aes(x=Group.1, y=x)) + geom_line() +
  xlab("Év") + ylab("A topikarányok összege az időszak
     cikkeiben") +
  ggtitle("Bírósági eljárások, tárgyalások")


doctop_15 <- monthly_topic(15)
g16 <- ggplot(doctop_15, aes(x=Group.1, y=x)) + geom_line() +
  xlab("Év") + ylab("A topikarányok összege az időszak
     cikkeiben") +
  ggtitle("Cozma-gyilkosság")

doctop_16 <- monthly_topic(16)
g17 <- ggplot(doctop_16, aes(x=Group.1, y=x)) + geom_line() +
  xlab("Év") + ylab("A topikarányok összege az időszak
     cikkeiben") +
```

```
  ggtitle("Közlekedéssel kapcsolatos kihágások, bűncselekmé
     nyek")

doctop_17 <- monthly_topic(17)
g18 <- ggplot(doctop_17, aes(x=Group.1, y=x)) + geom_line() +
  xlab("Év") + ylab("A topikarányok összege az időszak
     cikkeiben") +
  ggtitle("Verekedés, késelés, támadás")

doctop_18 <- monthly_topic(18)
g19 <- ggplot(doctop_18, aes(x=Group.1, y=x)) + geom_line() +
  xlab("Év") + ylab("A topikarányok összege az időszak
     cikkeiben") +
  ggtitle("Pásztor Albert és a cigánybűnözés")

doctop_19 <- monthly_topic(19)
g20 <- ggplot(doctop_19, aes(x=Group.1, y=x)) + geom_line() +
  xlab("Év") + ylab("A topikarányok összege az időszak
     cikkeiben") +
  ggtitle("Roma művészek, celebek bűncselekményei")

doctop_20 <- monthly_topic(20)
g21 <- ggplot(doctop_20, aes(x=Group.1, y=x)) + geom_line() +
  xlab("Év") + ylab("A topikarányok összege az időszak
     cikkeiben") +
  ggtitle("Gyilkosságok, gyújtogatások")

doctop_21 <- monthly_topic(21)
g22 <- ggplot(doctop_21, aes(x=Group.1, y=x)) + geom_line() +
  xlab("Év") + ylab("A topikarányok összege az időszak
     cikkeiben") +
  ggtitle("Tücsök és a hangya, finnországi roma bűncselekmé
     nyek")

doctop_22 <- monthly_topic(22)
g23 <- ggplot(doctop_22, aes(x=Group.1, y=x)) + geom_line() +
  xlab("Év") + ylab("A topikarányok összege az időszak
     cikkeiben") +
  ggtitle("Szociális segély, közmunka")

doctop_23 <- monthly_topic(23)
g24 <- ggplot(doctop_23, aes(x=Group.1, y=x)) + geom_line() +
  xlab("Év") + ylab("A topikarányok összege az időszak
     cikkeiben") +
  ggtitle("Szebb Jövőért Polgárőr Egyesület és Gyöngyöspata")

doctop_24 <- monthly_topic(24)
g25 <- ggplot(doctop_24, aes(x=Group.1, y=x)) + geom_line() +
  xlab("Év") + ylab("A topikarányok összege az időszak
     cikkeiben") +
```

```
  ggtitle("Roma, török, muszlim, fekete bevándorlók Európában"
     )

doctop_25 <- monthly_topic(25)
g26 <- ggplot(doctop_25, aes(x=Group.1, y=x)) + geom_line() +
  xlab("Év") + ylab("A topikarányok összege az időszak
     cikkeiben") +
  ggtitle("Oktatási, iskolai problémák")

doctop_26 <- monthly_topic(26)
g27 <- ggplot(doctop_26, aes(x=Group.1, y=x)) + geom_line() +
  xlab("Év") + ylab("A topikarányok összege az időszak
     cikkeiben") +
  ggtitle("Emberkereskedő, nőket futtató magyar roma bű
     nszervezetek
          külföldön")

setwd(".../ciganybunzes")

# call the multiplot function in multiplot.R before using

png("topic_in_time_1_4.png", width = 1000, height = 700)
multiplot(g1, g2, g3, g4, cols = 2)
dev.off()

png("topic_in_time_5_8.png", width = 1000, height = 700)
multiplot(g5, g6, g7, g8, cols = 2)
dev.off()

png("topic_in_time_9_12.png", width = 1000, height = 700)
multiplot(g9, g10, g11, g12, cols = 2)
dev.off()

png("topic_in_time_13_16.png", width = 1000, height = 700)
multiplot(g13, g14, g15, g16, cols = 2)
dev.off()

png("topic_in_time_17_20.png", width = 1000, height = 700)
multiplot(g17, g18, g19, g20, cols = 2)
dev.off()

png("topic_in_time_21_24.png", width = 1000, height = 700)
multiplot(g21, g22, g23, g24, cols = 2)
dev.off()

png("topic_in_time_25_27.png", width = 1000, height = 700)
multiplot(g25, g26, g27, cols = 2)
dev.off()
```

E.14. evaluation.R

```
# read the annotated database with document-topic assignments

setwd("../ciganybunzes")

annot <- read.csv("annotation.csv", header = T, sep = ";", dec
    = ".", stringsAsFactors = F, encoding = "UTF-8")

# read the MALLET document-topic output

setwd("../ciganybunzes/lda/mallet-2.0.7/kuruc_mallet_4_2_asym
   ")

doctop <- read.csv("doc-topics_1.csv", header = F, sep = "\t",
    dec = ".",  stringsAsFactors = F, encoding = "UTF-8")

# getting the annotation id's for doctop
id <- as.numeric(apply(doctop, 1,  function(x)
  unlist(strsplit(strsplit(x[2], "/")[[1]][7], ".txt"))))
doctop <- cbind(doctop, id)
matching <- match(doctop$id, annot$text_id)
doctop <- cbind(doctop, matching)
doctop_eval <- doctop[-(which(is.na(doctop$matching))), ]


# recategorization of the LDA topic assignments

doctop_eval$V4[which(doctop_eval$V3 == 0)] <- 1
doctop_eval$V3[which(doctop_eval$V3 == 0)] <- 14
doctop_eval$V4[which(doctop_eval$V3 == 1)] <- 12
doctop_eval$V3[which(doctop_eval$V3 == 1)] <- 13
doctop_eval$V4[which(doctop_eval$V3 == 2)] <- 1
doctop_eval$V6[which(doctop_eval$V3 == 2)] <- 2
doctop_eval$V3[which(doctop_eval$V3 == 2)] <- 13
doctop_eval$V3[which(doctop_eval$V3 == 3)] <- 13
doctop_eval$V4[which(doctop_eval$V3 == 4)] <- 1
doctop_eval$V6[which(doctop_eval$V3 == 4)] <- 6
doctop_eval$V3[which(doctop_eval$V3 == 4)] <- 12
doctop_eval$V4[which(doctop_eval$V3 == 5)] <- 13
doctop_eval$V3[which(doctop_eval$V3 == 5)] <- 12
doctop_eval$V3[which(doctop_eval$V3 == 6)] <- 1
doctop_eval$V3[which(doctop_eval$V3 == 7)] <- 12
doctop_eval$V4[which(doctop_eval$V3 == 8)] <- 12
doctop_eval$V3[which(doctop_eval$V3 == 8)] <- 13
doctop_eval$V3[which(doctop_eval$V3 == 9)] <- 13
doctop_eval$V4[which(doctop_eval$V3 == 10)] <- 13
doctop_eval$V3[which(doctop_eval$V3 == 10)] <- 11
doctop_eval$V4[which(doctop_eval$V3 == 11)] <- 12
doctop_eval$V3[which(doctop_eval$V3 == 11)] <- 13
```

```
doctop_eval$V3[which(doctop_eval$V3 == 12)] <- 13
doctop_eval$V3[which(doctop_eval$V3 == 13)] <- 13
doctop_eval$V3[which(doctop_eval$V3 == 14)] <- 13
doctop_eval$V4[which(doctop_eval$V3 == 15)] <- 12
doctop_eval$V3[which(doctop_eval$V3 == 15)] <- 13
doctop_eval$V4[which(doctop_eval$V3 == 16)] <- 12
doctop_eval$V3[which(doctop_eval$V3 == 16)] <- 13
doctop_eval$V3[which(doctop_eval$V3 == 17)] <- 13
doctop_eval$V4[which(doctop_eval$V3 == 18)] <- 1
doctop_eval$V3[which(doctop_eval$V3 == 18)] <- 13
doctop_eval$V4[which(doctop_eval$V3 == 19)] <- 10
doctop_eval$V3[which(doctop_eval$V3 == 19)] <- 13
doctop_eval$V4[which(doctop_eval$V3 == 20)] <- 12
doctop_eval$V3[which(doctop_eval$V3 == 20)] <- 13
doctop_eval$V3[which(doctop_eval$V3 == 21)] <- 13
doctop_eval$V4[which(doctop_eval$V3 == 22)] <- 1
doctop_eval$V6[which(doctop_eval$V3 == 22)] <- 6
doctop_eval$V8[which(doctop_eval$V3 == 22)] <- 8
doctop_eval$V3[which(doctop_eval$V3 == 22)] <- 12
doctop_eval$V4[which(doctop_eval$V3 == 23)] <- 1
doctop_eval$V6[which(doctop_eval$V3 == 23)] <- 12
doctop_eval$V3[which(doctop_eval$V3 == 23)] <- 13
doctop_eval$V4[which(doctop_eval$V3 == 24)] <- 1
doctop_eval$V6[which(doctop_eval$V3 == 24)] <- 4
doctop_eval$V8[which(doctop_eval$V3 == 24)] <- 12
doctop_eval$V3[which(doctop_eval$V3 == 24)] <- 14
doctop_eval$V3[which(doctop_eval$V3 == 25)] <- 7
doctop_eval$V4[which(doctop_eval$V3 == 26)] <- 3
doctop_eval$V6[which(doctop_eval$V3 == 26)] <- 4
doctop_eval$V3[which(doctop_eval$V3 == 26)] <- 13


doctop_eval <- cbind(doctop_eval, V14 = rep(NA, 622), V15 =
   rep(NA, 622),
                V16 = rep(NA, 622), V17 = rep(NA, 622),
                V18 = rep(NA, 622), V19 = rep(NA, 622),
                V20 = rep(NA, 622), V21 = rep(NA, 622),
                V22 = rep(NA, 622), V23 = rep(NA, 622))

doctop_eval$V10[which(doctop_eval$V5 == 0)] <- 1
doctop_eval$V5[which(doctop_eval$V5 == 0)] <- 14
doctop_eval$V10[which(doctop_eval$V5 == 1)] <- 12
doctop_eval$V5[which(doctop_eval$V5 == 1)] <- 13
doctop_eval$V10[which(doctop_eval$V5 == 2)] <- 1
doctop_eval$V12[which(doctop_eval$V5 == 2)] <- 2
doctop_eval$V5[which(doctop_eval$V5 == 2)] <- 13
doctop_eval$V5[which(doctop_eval$V5 == 3)] <- 13
doctop_eval$V10[which(doctop_eval$V5 == 4)] <- 1
doctop_eval$V12[which(doctop_eval$V5 == 4)] <- 6
doctop_eval$V5[which(doctop_eval$V5 == 4)] <- 12
doctop_eval$V10[which(doctop_eval$V5 == 5)] <- 13
```

```
doctop_eval$V5[which(doctop_eval$V5 == 5)] <- 12
doctop_eval$V5[which(doctop_eval$V5 == 6)] <- 1
doctop_eval$V5[which(doctop_eval$V5 == 7)] <- 12
doctop_eval$V10[which(doctop_eval$V5 == 8)] <- 12
doctop_eval$V5[which(doctop_eval$V5 == 8)] <- 13
doctop_eval$V5[which(doctop_eval$V5 == 9)] <- 13
doctop_eval$V10[which(doctop_eval$V5 == 10)] <- 13
doctop_eval$V5[which(doctop_eval$V5 == 10)] <- 11
doctop_eval$V10[which(doctop_eval$V5 == 11)] <- 12
doctop_eval$V5[which(doctop_eval$V5 == 11)] <- 13
doctop_eval$V5[which(doctop_eval$V5 == 12)] <- 13
doctop_eval$V5[which(doctop_eval$V5 == 13)] <- 13
doctop_eval$V5[which(doctop_eval$V5 == 14)] <- 13
doctop_eval$V10[which(doctop_eval$V5 == 15)] <- 12
doctop_eval$V5[which(doctop_eval$V5 == 15)] <- 13
doctop_eval$V10[which(doctop_eval$V5 == 16)] <- 12
doctop_eval$V5[which(doctop_eval$V5 == 16)] <- 13
doctop_eval$V5[which(doctop_eval$V5 == 17)] <- 13
doctop_eval$V10[which(doctop_eval$V5 == 18)] <- 1
doctop_eval$V5[which(doctop_eval$V5 == 18)] <- 13
doctop_eval$V10[which(doctop_eval$V5 == 19)] <- 10
doctop_eval$V5[which(doctop_eval$V5 == 19)] <- 13
doctop_eval$V10[which(doctop_eval$V5 == 20)] <- 12
doctop_eval$V5[which(doctop_eval$V5 == 20)] <- 13
doctop_eval$V5[which(doctop_eval$V5 == 21)] <- 13
doctop_eval$V10[which(doctop_eval$V5 == 22)] <- 1
doctop_eval$V12[which(doctop_eval$V5 == 22)] <- 6
doctop_eval$V14[which(doctop_eval$V5 == 22)] <- 8
doctop_eval$V5[which(doctop_eval$V5 == 22)] <- 12
doctop_eval$V10[which(doctop_eval$V5 == 23)] <- 1
doctop_eval$V12[which(doctop_eval$V5 == 23)] <- 12
doctop_eval$V5[which(doctop_eval$V5 == 23)] <- 13
doctop_eval$V10[which(doctop_eval$V5 == 24)] <- 1
doctop_eval$V12[which(doctop_eval$V5 == 24)] <- 4
doctop_eval$V14[which(doctop_eval$V5 == 24)] <- 12
doctop_eval$V5[which(doctop_eval$V5 == 24)] <- 14
doctop_eval$V5[which(doctop_eval$V5 == 25)] <- 7
doctop_eval$V10[which(doctop_eval$V5 == 26)] <- 3
doctop_eval$V12[which(doctop_eval$V5 == 26)] <- 4
doctop_eval$V5[which(doctop_eval$V5 == 26)] <- 13


doctop_eval$V15[which(doctop_eval$V7 == 0)] <- 1
doctop_eval$V7[which(doctop_eval$V7 == 0)] <- 14
doctop_eval$V15[which(doctop_eval$V7 == 1)] <- 12
doctop_eval$V7[which(doctop_eval$V7 == 1)] <- 13
doctop_eval$V15[which(doctop_eval$V7 == 2)] <- 1
doctop_eval$V16[which(doctop_eval$V7 == 2)] <- 2
doctop_eval$V7[which(doctop_eval$V7 == 2)] <- 13
doctop_eval$V7[which(doctop_eval$V7 == 3)] <- 13
doctop_eval$V15[which(doctop_eval$V7 == 4)] <- 1
```

```
doctop_eval$V16[which(doctop_eval$V7 == 4)] <- 6
doctop_eval$V7[which(doctop_eval$V7 == 4)] <- 12
doctop_eval$V15[which(doctop_eval$V7 == 5)] <- 13
doctop_eval$V7[which(doctop_eval$V7 == 5)] <- 12
doctop_eval$V7[which(doctop_eval$V7 == 6)] <- 1
doctop_eval$V7[which(doctop_eval$V7 == 7)] <- 12
doctop_eval$V15[which(doctop_eval$V7 == 8)] <- 12
doctop_eval$V7[which(doctop_eval$V7 == 8)] <- 13
doctop_eval$V7[which(doctop_eval$V7 == 9)] <- 13
doctop_eval$V15[which(doctop_eval$V7 == 10)] <- 13
doctop_eval$V7[which(doctop_eval$V7 == 10)] <- 11
doctop_eval$V15[which(doctop_eval$V7 == 11)] <- 12
doctop_eval$V7[which(doctop_eval$V7 == 11)] <- 13
doctop_eval$V7[which(doctop_eval$V7 == 12)] <- 13
doctop_eval$V7[which(doctop_eval$V7 == 13)] <- 13
doctop_eval$V7[which(doctop_eval$V7 == 14)] <- 13
doctop_eval$V15[which(doctop_eval$V7 == 15)] <- 12
doctop_eval$V7[which(doctop_eval$V7 == 15)] <- 13
doctop_eval$V15[which(doctop_eval$V7 == 16)] <- 12
doctop_eval$V7[which(doctop_eval$V7 == 16)] <- 13
doctop_eval$V7[which(doctop_eval$V7 == 17)] <- 13
doctop_eval$V15[which(doctop_eval$V7 == 18)] <- 1
doctop_eval$V7[which(doctop_eval$V7 == 18)] <- 13
doctop_eval$V15[which(doctop_eval$V7 == 19)] <- 10
doctop_eval$V7[which(doctop_eval$V7 == 19)] <- 13
doctop_eval$V15[which(doctop_eval$V7 == 20)] <- 12
doctop_eval$V7[which(doctop_eval$V7 == 20)] <- 13
doctop_eval$V7[which(doctop_eval$V7 == 21)] <- 13
doctop_eval$V15[which(doctop_eval$V7 == 22)] <- 1
doctop_eval$V16[which(doctop_eval$V7 == 22)] <- 6
doctop_eval$V17[which(doctop_eval$V7 == 22)] <- 8
doctop_eval$V7[which(doctop_eval$V7 == 22)] <- 12
doctop_eval$V15[which(doctop_eval$V7 == 23)] <- 1
doctop_eval$V16[which(doctop_eval$V7 == 23)] <- 12
doctop_eval$V7[which(doctop_eval$V7 == 23)] <- 13
doctop_eval$V15[which(doctop_eval$V7 == 24)] <- 1
doctop_eval$V16[which(doctop_eval$V7 == 24)] <- 4
doctop_eval$V17[which(doctop_eval$V7 == 24)] <- 12
doctop_eval$V7[which(doctop_eval$V7 == 24)] <- 14
doctop_eval$V7[which(doctop_eval$V7 == 25)] <- 7
doctop_eval$V15[which(doctop_eval$V7 == 26)] <- 3
doctop_eval$V16[which(doctop_eval$V7 == 26)] <- 4
doctop_eval$V7[which(doctop_eval$V7 == 26)] <- 13


doctop_eval$V18[which(doctop_eval$V9 == 0)] <- 1
doctop_eval$V9[which(doctop_eval$V9 == 0)] <- 14
doctop_eval$V18[which(doctop_eval$V9 == 1)] <- 12
doctop_eval$V9[which(doctop_eval$V9 == 1)] <- 13
doctop_eval$V18[which(doctop_eval$V9 == 2)] <- 1
doctop_eval$V19[which(doctop_eval$V9 == 2)] <- 2
```

```
doctop_eval$V9[which(doctop_eval$V9 == 2)] <- 13
doctop_eval$V9[which(doctop_eval$V9 == 3)] <- 13
doctop_eval$V18[which(doctop_eval$V9 == 4)] <- 1
doctop_eval$V19[which(doctop_eval$V9 == 4)] <- 6
doctop_eval$V9[which(doctop_eval$V9 == 4)] <- 12
doctop_eval$V18[which(doctop_eval$V9 == 5)] <- 13
doctop_eval$V9[which(doctop_eval$V9 == 5)] <- 12
doctop_eval$V9[which(doctop_eval$V9 == 6)] <- 1
doctop_eval$V9[which(doctop_eval$V9 == 7)] <- 12
doctop_eval$V18[which(doctop_eval$V9 == 8)] <- 12
doctop_eval$V9[which(doctop_eval$V9 == 8)] <- 13
doctop_eval$V9[which(doctop_eval$V9 == 9)] <- 13
doctop_eval$V18[which(doctop_eval$V9 == 10)] <- 13
doctop_eval$V9[which(doctop_eval$V9 == 10)] <- 11
doctop_eval$V18[which(doctop_eval$V9 == 11)] <- 12
doctop_eval$V9[which(doctop_eval$V9 == 11)] <- 13
doctop_eval$V9[which(doctop_eval$V9 == 12)] <- 13
doctop_eval$V9[which(doctop_eval$V9 == 13)] <- 13
doctop_eval$V9[which(doctop_eval$V9 == 14)] <- 13
doctop_eval$V18[which(doctop_eval$V9 == 15)] <- 12
doctop_eval$V9[which(doctop_eval$V9 == 15)] <- 13
doctop_eval$V18[which(doctop_eval$V9 == 16)] <- 12
doctop_eval$V9[which(doctop_eval$V9 == 16)] <- 13
doctop_eval$V9[which(doctop_eval$V9 == 17)] <- 13
doctop_eval$V18[which(doctop_eval$V9 == 18)] <- 1
doctop_eval$V9[which(doctop_eval$V9 == 18)] <- 13
doctop_eval$V18[which(doctop_eval$V9 == 19)] <- 10
doctop_eval$V9[which(doctop_eval$V9 == 19)] <- 13
doctop_eval$V18[which(doctop_eval$V9 == 20)] <- 12
doctop_eval$V9[which(doctop_eval$V9 == 20)] <- 13
doctop_eval$V9[which(doctop_eval$V9 == 21)] <- 13
doctop_eval$V18[which(doctop_eval$V9 == 22)] <- 1
doctop_eval$V19[which(doctop_eval$V9 == 22)] <- 6
doctop_eval$V20[which(doctop_eval$V9 == 22)] <- 8
doctop_eval$V9[which(doctop_eval$V9 == 22)] <- 12
doctop_eval$V18[which(doctop_eval$V9 == 23)] <- 1
doctop_eval$V19[which(doctop_eval$V9 == 23)] <- 12
doctop_eval$V9[which(doctop_eval$V9 == 23)] <- 13
doctop_eval$V18[which(doctop_eval$V9 == 24)] <- 1
doctop_eval$V19[which(doctop_eval$V9 == 24)] <- 4
doctop_eval$V20[which(doctop_eval$V9 == 24)] <- 12
doctop_eval$V9[which(doctop_eval$V9 == 24)] <- 14
doctop_eval$V9[which(doctop_eval$V9 == 25)] <- 7
doctop_eval$V18[which(doctop_eval$V9 == 26)] <- 3
doctop_eval$V19[which(doctop_eval$V9 == 26)] <- 4
doctop_eval$V9[which(doctop_eval$V9 == 26)] <- 13


doctop_eval$V21[which(doctop_eval$V11 == 0)] <- 1
doctop_eval$V11[which(doctop_eval$V11 == 0)] <- 14
doctop_eval$V21[which(doctop_eval$V11 == 1)] <- 12
```

```
doctop_eval$V11[which(doctop_eval$V11 == 1)] <- 13
doctop_eval$V21[which(doctop_eval$V11 == 2)] <- 1
doctop_eval$V22[which(doctop_eval$V11 == 2)] <- 2
doctop_eval$V11[which(doctop_eval$V11 == 2)] <- 13
doctop_eval$V11[which(doctop_eval$V11 == 3)] <- 13
doctop_eval$V21[which(doctop_eval$V11 == 4)] <- 1
doctop_eval$V22[which(doctop_eval$V11 == 4)] <- 6
doctop_eval$V11[which(doctop_eval$V11 == 4)] <- 12
doctop_eval$V21[which(doctop_eval$V11 == 5)] <- 13
doctop_eval$V11[which(doctop_eval$V11 == 5)] <- 12
doctop_eval$V11[which(doctop_eval$V11 == 6)] <- 1
doctop_eval$V11[which(doctop_eval$V11 == 7)] <- 12
doctop_eval$V21[which(doctop_eval$V11 == 8)] <- 12
doctop_eval$V11[which(doctop_eval$V11 == 8)] <- 13
doctop_eval$V11[which(doctop_eval$V11 == 9)] <- 13
doctop_eval$V21[which(doctop_eval$V11 == 10)] <- 13
doctop_eval$V11[which(doctop_eval$V11 == 10)] <- 11
doctop_eval$V21[which(doctop_eval$V11 == 11)] <- 12
doctop_eval$V11[which(doctop_eval$V11 == 11)] <- 13
doctop_eval$V11[which(doctop_eval$V11 == 12)] <- 13
doctop_eval$V11[which(doctop_eval$V11 == 13)] <- 13
doctop_eval$V11[which(doctop_eval$V11 == 14)] <- 13
doctop_eval$V21[which(doctop_eval$V11 == 15)] <- 12
doctop_eval$V11[which(doctop_eval$V11 == 15)] <- 13
doctop_eval$V21[which(doctop_eval$V11 == 16)] <- 12
doctop_eval$V11[which(doctop_eval$V11 == 16)] <- 13
doctop_eval$V11[which(doctop_eval$V11 == 17)] <- 13
doctop_eval$V21[which(doctop_eval$V11 == 18)] <- 1
doctop_eval$V11[which(doctop_eval$V11 == 18)] <- 13
doctop_eval$V21[which(doctop_eval$V11 == 19)] <- 10
doctop_eval$V11[which(doctop_eval$V11 == 19)] <- 13
doctop_eval$V21[which(doctop_eval$V11 == 20)] <- 12
doctop_eval$V11[which(doctop_eval$V11 == 20)] <- 13
doctop_eval$V11[which(doctop_eval$V11 == 21)] <- 13
doctop_eval$V21[which(doctop_eval$V11 == 22)] <- 1
doctop_eval$V22[which(doctop_eval$V11 == 22)] <- 6
doctop_eval$V23[which(doctop_eval$V11 == 22)] <- 8
doctop_eval$V11[which(doctop_eval$V11 == 22)] <- 12
doctop_eval$V21[which(doctop_eval$V11 == 23)] <- 1
doctop_eval$V22[which(doctop_eval$V11 == 23)] <- 12
doctop_eval$V11[which(doctop_eval$V11 == 23)] <- 13
doctop_eval$V21[which(doctop_eval$V11 == 24)] <- 1
doctop_eval$V22[which(doctop_eval$V11 == 24)] <- 4
doctop_eval$V23[which(doctop_eval$V11 == 24)] <- 12
doctop_eval$V11[which(doctop_eval$V11 == 24)] <- 14
doctop_eval$V11[which(doctop_eval$V11 == 25)] <- 7
doctop_eval$V21[which(doctop_eval$V11 == 26)] <- 3
doctop_eval$V22[which(doctop_eval$V11 == 26)] <- 4
doctop_eval$V11[which(doctop_eval$V11 == 26)] <- 13
```

```
doctop_eval <- doctop_eval[order(doctop_eval$id), ]

# defining doctop_eval_topic function for getting the LDA
   topic assignments for a topic
doctop_eval_topic <- function(topicnumber){
  id1 <-which(doctop_eval$V3 == topicnumber)
  id2 <-which(doctop_eval$V4 == topicnumber)
  id3 <-which(doctop_eval$V5 == topicnumber)
  id4 <-which(doctop_eval$V6 == topicnumber)
  id5 <-which(doctop_eval$V7 == topicnumber)
  id6 <-which(doctop_eval$V8 == topicnumber)
  id7 <-which(doctop_eval$V9 == topicnumber)
  id8 <-which(doctop_eval$V10 == topicnumber)
  id9 <-which(doctop_eval$V11 == topicnumber)
  id10 <-which(doctop_eval$V12 == topicnumber)
  id11 <-which(doctop_eval$V14 == topicnumber)
  id12 <-which(doctop_eval$V15 == topicnumber)
  id13 <-which(doctop_eval$V16 == topicnumber)
  id14 <-which(doctop_eval$V17 == topicnumber)
  id15 <-which(doctop_eval$V18 == topicnumber)
  id16 <-which(doctop_eval$V19 == topicnumber)
  id17 <-which(doctop_eval$V20 == topicnumber)
  id18 <-which(doctop_eval$V21 == topicnumber)
  id19 <-which(doctop_eval$V22 == topicnumber)
  id20 <-which(doctop_eval$V23 == topicnumber)

  unique_id_topic <- sort(unique(c(id1, id2, id3, id4, id5,
                                   id6, id7, id8, id9, id10,
                                   id11, id12, id13, id14,
                                   id15, id16, id17, id18,
                                   id19, id20)))

  unique_id_topic
}

# defining annot_topic function for getting the annotated
   topic assignments for a topic
annot_topic <- function(topicnumber){
  id1 <-which(annot$category1 == topicnumber)
  id2 <-which(annot$category2 == topicnumber)
  id3 <-which(annot$category3 == topicnumber)
  id4 <-which(annot$category4 == topicnumber)
  id5 <-which(annot$category5 == topicnumber)

  unique_id_topic <- sort(c(id1, id2, id3, id4, id5))

  unique_id_topic
}

# calculating false negatives (FN), true positives (TP), false
    positives (FP), recall (R) and precision (P) for a topic
```

```
FN_1 <- length(which(is.na(match(annot_topic(1), doctop_eval_
   topic(1)))))
TP_1 <- length(match(annot_topic(1), doctop_eval_topic(1))) -
   FN_1
FP_1 <- length(which(is.na(match(doctop_eval_topic(1), annot_
   topic(1)))))

R_1 <- TP_1/(TP_1 + FN_1)
P_1 <- TP_1/(TP_1 + FP_1)

FN_2 <- length(which(is.na(match(annot_topic(2), doctop_eval_
   topic(2)))))
TP_2 <- length(match(annot_topic(2), doctop_eval_topic(2))) -
   FN_2
FP_2 <- length(which(is.na(match(doctop_eval_topic(2), annot_
   topic(2)))))

R_2 <- TP_2/(TP_2 + FN_2)
P_2 <- TP_2/(TP_2 + FP_2)

FN_3 <- length(which(is.na(match(annot_topic(3), doctop_eval_
   topic(3)))))
TP_3 <- length(match(annot_topic(3), doctop_eval_topic(3))) -
   FN_3
FP_3 <- length(which(is.na(match(doctop_eval_topic(3), annot_
   topic(3)))))

R_3 <- TP_3/(TP_3 + FN_3)
P_3 <- TP_3/(TP_3 + FP_3)

FN_4 <- length(which(is.na(match(annot_topic(4), doctop_eval_
   topic(4)))))
TP_4 <- length(match(annot_topic(4), doctop_eval_topic(4))) -
   FN_4
FP_4 <- length(which(is.na(match(doctop_eval_topic(4), annot_
   topic(4)))))

R_4 <- TP_4/(TP_4 + FN_4)
P_4 <- TP_4/(TP_4 + FP_4)

FN_5 <- length(which(is.na(match(annot_topic(5), doctop_eval_
   topic(5)))))
TP_5 <- length(match(annot_topic(5), doctop_eval_topic(5))) -
   FN_5
FP_5 <- length(which(is.na(match(doctop_eval_topic(5), annot_
   topic(5)))))

R_5 <- TP_5/(TP_5 + FN_5)
P_5 <- TP_5/(TP_5 + FP_5)
```

```
FN_6 <- length(which(is.na(match(annot_topic(6), doctop_eval_
   topic(6)))))
TP_6 <- length(match(annot_topic(6), doctop_eval_topic(6))) -
   FN_6
FP_6 <- length(which(is.na(match(doctop_eval_topic(6), annot_
   topic(6)))))

R_6 <- TP_6/(TP_6 + FN_6)
P_6 <- TP_6/(TP_6 + FP_6)


FN_7 <- length(which(is.na(match(annot_topic(7), doctop_eval_
   topic(7)))))
TP_7 <- length(match(annot_topic(7), doctop_eval_topic(7))) -
   FN_7
FP_7 <- length(which(is.na(match(doctop_eval_topic(7), annot_
   topic(7)))))

R_7 <- TP_7/(TP_7 + FN_7)
P_7 <- TP_7/(TP_7 + FP_7)


FN_8 <- length(which(is.na(match(annot_topic(8), doctop_eval_
   topic(8)))))
TP_8 <- length(match(annot_topic(8), doctop_eval_topic(8))) -
   FN_8
FP_8 <- length(which(is.na(match(doctop_eval_topic(8), annot_
   topic(8)))))

R_8 <- TP_8/(TP_8 + FN_8)
P_8 <- TP_8/(TP_8 + FP_8)

FN_9 <- length(which(is.na(match(annot_topic(9), doctop_eval_
   topic(9)))))
TP_9 <- length(match(annot_topic(9), doctop_eval_topic(9))) -
   FN_9
FP_9 <- length(which(is.na(match(doctop_eval_topic(9), annot_
   topic(9)))))

R_9 <- TP_9/(TP_9 + FN_9)
P_9 <- TP_9/(TP_9 + FP_9)

FN_10 <- length(which(is.na(match(annot_topic(10), doctop_eval
   _topic(10)))))
TP_10 <- length(match(annot_topic(10), doctop_eval_topic(10)))
    - FN_10
FP_10 <- length(which(is.na(match(doctop_eval_topic(10), annot
   _topic(10)))))

R_10 <- TP_10/(TP_10 + FN_10)
P_10 <- TP_10/(TP_10 + FP_10)
```

```
FN_11 <- length(which(is.na(match(annot_topic(11), doctop_eval
   _topic(11)))))
TP_11 <- length(match(annot_topic(11), doctop_eval_topic(11)))
    - FN_11
FP_11 <- length(which(is.na(match(doctop_eval_topic(11), annot
   _topic(11)))))

R_11 <- TP_11/(TP_11 + FN_11)
P_11 <- TP_11/(TP_11 + FP_11)

FN_12 <- length(which(is.na(match(annot_topic(12), doctop_eval
   _topic(12)))))
TP_12 <- length(match(annot_topic(12), doctop_eval_topic(12)))
    - FN_12
FP_12 <- length(which(is.na(match(doctop_eval_topic(12), annot
   _topic(12)))))

R_12 <- TP_12/(TP_12 + FN_12)
P_12 <- TP_12/(TP_12 + FP_12)


FN_13 <- length(which(is.na(match(annot_topic(13), doctop_eval
   _topic(13)))))
TP_13 <- length(match(annot_topic(13), doctop_eval_topic(13)))
    - FN_13
FP_13 <- length(which(is.na(match(doctop_eval_topic(13), annot
   _topic(13)))))

R_13 <- TP_13/(TP_13 + FN_13)
P_13 <- TP_13/(TP_13 + FP_13)


FN_14 <- length(which(is.na(match(annot_topic(14), doctop_eval
   _topic(14)))))
TP_14 <- length(match(annot_topic(14), doctop_eval_topic(14)))
    - FN_14
FP_14 <- length(which(is.na(match(doctop_eval_topic(14), annot
   _topic(14)))))

R_14 <- TP_14/(TP_14 + FN_14)
P_14 <- TP_14/(TP_14 + FP_14)

FN_15 <- length(which(is.na(match(annot_topic(15), doctop_eval
   _topic(15)))))
TP_15 <- length(match(annot_topic(15), doctop_eval_topic(15)))
    - FN_15
FP_15 <- length(which(is.na(match(doctop_eval_topic(15), annot
   _topic(15)))))

R_15 <- TP_15/(TP_15 + FN_15)
```

```
P_15 <- TP_15/(TP_15 + FP_15)


# summing all annotated topic assignments

annot_all <- length(annot_topic(1)) + length(annot_topic(2)) +
  length(annot_topic(3)) + length(annot_topic(4)) +
  length(annot_topic(5)) + length(annot_topic(6)) +
  length(annot_topic(7)) + length(annot_topic(8)) +
  length(annot_topic(9)) + length(annot_topic(10)) +
  length(annot_topic(11)) + length(annot_topic(12)) +
  length(annot_topic(13)) + length(annot_topic(14)) +
  length(annot_topic(15))

# calculating total recall and precision proportional to size

R <- (length(annot_topic(1))/annot_all * R_1 +
        length(annot_topic(2))/annot_all * R_2 +
        length(annot_topic(3))/annot_all * R_3 +
        length(annot_topic(4))/annot_all * R_4 +
        length(annot_topic(5))/annot_all * R_5 +
        length(annot_topic(6))/annot_all * R_6 +
        length(annot_topic(7))/annot_all * R_7 +
        length(annot_topic(8))/annot_all * R_8 +
        length(annot_topic(9))/annot_all * R_9 +
        length(annot_topic(10))/annot_all * R_10 +
        length(annot_topic(11))/annot_all * R_11 +
        length(annot_topic(12))/annot_all * R_12 +
        length(annot_topic(13))/annot_all * R_13 +
        length(annot_topic(14))/annot_all * R_14 +
        length(annot_topic(15))/annot_all * R_15)

P <- (length(annot_topic(1))/annot_all * P_1 +
        length(annot_topic(2))/annot_all * P_2 +
        length(annot_topic(3))/annot_all * P_3 +
        length(annot_topic(4))/annot_all * P_4 +
        length(annot_topic(6))/annot_all * P_6 +
        length(annot_topic(7))/annot_all * P_7 +
        length(annot_topic(8))/annot_all * P_8 +
        length(annot_topic(10))/annot_all * P_10 +
        length(annot_topic(12))/annot_all * P_12 +
        length(annot_topic(13))/annot_all * P_13 +
        length(annot_topic(14))/annot_all * P_14)

# calcualting F1-measure
F1 <- (2*P*R)/(P+R)
```

F. MALLET KÓDOK

F.1. import_dir

```
cd /.../ciganybunzes/lda/Mallet

bin/mallet import-dir --input /home/kitti/Desktop/ciganybunzes
   /stemmed_filtered_4_mod_2/ --output kuruc_4.mallet --token-
   regex '[\p{L}\p{M}]+' --keep-sequence --stoplist-file
   stoplists/stoplist_stemmed_2.txt
```

F.2. choose_lda_sym

```
### fitting LDA for kuruc_4.mallet
# iteration no: 1000, symmetric alpha, topic no: 20-30


bin/mallet train-topics --input kuruc_4.mallet --num-topics 20
    --num-iterations 1000 --optimize-interval 10 --num-top-
   words 30 --random-seed 103352 --use-symmetric-alpha

## LL/token
##  -9.30273

bin/mallet train-topics --input kuruc_4.mallet --num-topics 21
    --num-iterations 1000 --optimize-interval 10 --num-top-
   words 30 --random-seed 103352 --use-symmetric-alpha


## LL/token
##  -9.30468

bin/mallet train-topics --input kuruc_4.mallet --num-topics 22
    --num-iterations 1000 --optimize-interval 10 --num-top-
   words 30 --random-seed 103352 --use-symmetric-alpha


## LL/token
##  -9.30176
```

```
bin/mallet train-topics --input kuruc_4.mallet --num-topics 23
    --num-iterations 1000 --optimize-interval 10 --num-top-
   words 30 --random-seed 103352 --use-symmetric-alpha


## LL/token
## -9.29597

bin/mallet train-topics --input kuruc_4.mallet --num-topics 24
    --num-iterations 1000 --optimize-interval 10 --num-top-
   words 30 --random-seed 103352 --use-symmetric-alpha


## LL/token
## -9.29712

bin/mallet train-topics --input kuruc_4.mallet --num-topics 25
    --num-iterations 1000 --optimize-interval 10 --num-top-
   words 30 --random-seed 103352 --use-symmetric-alpha

## LL/token
## -9.29706

bin/mallet train-topics --input kuruc_4.mallet --num-topics 26
    --num-iterations 1000 --optimize-interval 10 --num-top-
   words 30 --random-seed 103352 --use-symmetric-alpha


## LL/token
## -9.29934

bin/mallet train-topics --input kuruc_4.mallet --num-topics 27
    --num-iterations 1000 --optimize-interval 10 --num-top-
   words 30 --random-seed 103352 --use-symmetric-alpha


## LL/token
## -9.29982

bin/mallet train-topics --input kuruc_4.mallet --num-topics 28
    --num-iterations 1000 --optimize-interval 10 --num-top-
   words 30 --random-seed 103352 --use-symmetric-alpha


## LL/token
## -9.29321

bin/mallet train-topics --input kuruc_4.mallet --num-topics 29
    --num-iterations 1000 --optimize-interval 10 --num-top-
   words 30 --random-seed 103352 --use-symmetric-alpha
```

```
## LL/token
##  -9.29414


bin/mallet train-topics --input kuruc_4.mallet --num-topics 30
    \
--num-iterations 1000 --optimize-interval 10 --num-top-words
   30 \
--random-seed 103352 --use-symmetric-alpha


## LL/token
##  -9.28881
```

F.3. train_lda_sym

```
### fitting LDA for kuruc_4.mallet
# iteration no: 5000, symmetric alpha, topic no:

bin/mallet train-topics --input kuruc_4.mallet --num-topics 30
    --num-iterations 5000 --optimize-interval 10 --num-top-
   words 30 --word-topic-counts-file word-topic-counts_1.csv
   --output-state topic-state_1.gz --output-state-interval 10
   --output-doc-topics doc-topics_1.csv --doc-topics-threshold
    0.1 --output-topic-keys topic-keys_1.csv --diagnostics-
   file diagnostics_1.csv --random-seed 103352 --use-symmetric
   -alpha true --alpha 5.0 --beta 0.01
```

F.4. choose_lda_asym

```
### fitting LDA for kuruc_4.mallet
# iteration no: 1000, asymmetric alpha, topic no: 25-30

bin/mallet train-topics --input kuruc_4.mallet --num-topics 25
    --num-iterations 1000 --optimize-interval 10 --num-top-
   words 30 --random-seed 452

## LL/token
##  -9.28559

bin/mallet train-topics --input kuruc_4.mallet --num-topics 26
    --num-iterations 1000 --optimize-interval 10 --num-top-
   words 30 --random-seed 452
```

```
## LL/token
## -9.30074

bin/mallet train-topics --input kuruc_4.mallet --num-topics 27
    --num-iterations 1000 --optimize-interval 10 --num-top-
   words 30 --random-seed 452

## LL/token
## -9.28196

bin/mallet train-topics --input kuruc_4.mallet --num-topics 28
    --num-iterations 1000 --optimize-interval 10 --num-top-
   words 30 --random-seed 452

## LL/token
## -9.28983

bin/mallet train-topics --input kuruc_4.mallet --num-topics 29
    --num-iterations 1000 --optimize-interval 10 --num-top-
   words 30 --random-seed 452

## LL/token
## -9.28983

bin/mallet train-topics --input kuruc_4.mallet --num-topics 30
    --num-iterations 1000 --optimize-interval 10 --num-top-
   words 30 --random-seed 452

## LL/token
## -9.2842
```

F.5. train_lda_asym

```
### fitting LDA for kuruc_4.mallet
# iteration no: 5000, asymmetric alpha, topic no: 27

bin/mallet train-topics --input kuruc_4.mallet --num-topics 27
    --num-iterations 5000 --optimize-interval 10 --num-top-
   words 30 --word-topic-counts-file word-topic-counts_1.csv
   --output-state topic-state_1.gz --output-state-interval 10
   --output-doc-topics doc-topics_1.csv --doc-topics-threshold
    0.1 --output-topic-keys topic-keys_1.csv --diagnostics-
   file diagnostics_1.csv --random-seed 452 --alpha 5.0 --beta
    0.01

bin/mallet train-topics --input kuruc_4.mallet --num-topics 27
    --num-iterations 5000 --optimize-interval 10 --num-top-
   words 30 --word-topic-counts-file word-topic-counts_1.csv
```

```
    --output-state topic-state_1.gz --output-state-interval 10
    --output-doc-topics doc-topics_1.csv --doc-topics-threshold
     0.1 --output-topic-keys topic-keys_1.csv  --diagnostics-
    file diagnostics_1.csv --random-seed 893515 --alpha 9.13 --
    beta 0.15

bin/mallet train-topics --input kuruc_4.mallet --num-topics 27
     --num-iterations 5000 --optimize-interval 10 --num-top-
    words 30 --word-topic-counts-file word-topic-counts_1.csv
    --output-state topic-state_1.gz --output-state-interval 10
    --output-doc-topics doc-topics_1.csv --doc-topics-threshold
     0.1 --output-topic-keys topic-keys_1.csv --diagnostics-
    file diagnostics_1.csv --random-seed 2385467 --alpha 0.65
    --beta 0.0032
```

G. EGYÉB KÓDOK

G.1. recode.sh

```
1. iconv -f ISO-8859-2 -t UTF-8 *
2. find . -name "*.html" -exec iconv -f ISO-8859-2 -t UTF-8 {}
    -o ../recoded/{} \;

grep -ilR '<p class=\"cikkdatum\"><.>Cigánybűnözés</a>'
grep -ilR '<p class=\"cikkheader\">Cigánybűnözés</p>'
find . -type f -exec grep -ilR '<p class=\"cikkdatum\"><.*>\
Cigánybűnözés</a>' {} \; | xargs -I % cp % ../cb/

find . -type f -exec grep -ilR '<p class=\"cikkheader\">\
Cigánybűnözés</p>' {} \; | xargs -I % cp % ../cb/
```

G.2. rename_as_gz

```
rename s/'$'/'.gz'/ *
```

H. STOPSZÓ LISTA

ahogy a ablak ad ahogyan ahol ajtó akar akár aki akkora alap alaposan alapvetően albert alig aligha áll állandóan állat állít állítólag álló általában általános amely amelyik ami amiért amikor amilyen amúgy and andrás anélkül angol annyira anya anyag apa április ár are arppi artúr as asszony at átlagosan attila augusztus autós az azért azon azonnal azóta aztán azután balogh barát bárki bármely bármilyen been belőle benne berendezés beszél bizony bizonyára biztos biztosan blikk borsodi budapesti by cég cél cigánybűnözés cikk cím című csaba csak csakis csaknem csapat csatorna csinál csupán csütörtökön dávid december délelőtt délután dolgoz dolog édesanya egész egészen egy egyáltalán egyaránt egyben egyben egyéb egyébként egyedül egyelőre egyenesen egyértelmű egyértelműen egyes egyidejűleg egyik egymás egyre egyrészt egyszer egyszerre egyszerű egyszerűen együtt egyúttal éjjel él elég élet eleve elkezd elleni elmegy elmond elmúlt élő előbb elolvas előre először előtt első elsősorban ember emiatt én enged épp éppen ér érdekes érdemes erdő eredeti eredetileg érez érkezik érkező erősen ért es esemény eset esetleg este ész eszik európa év évente éves ez ezáltal ezelőtt ezután ezúttal fa fal február fej feltehetően feltétlenül fent fenti ferenc férfi finnország fiú flórián fő fog fogalmaz foglalkozik főként fokozatosan föld főleg folyamatos folyamatosan fontos for franciaország friss frissítés függelék függetlenül gábor gáspár gondol gyakorlatilag gyakran gyermek gyermekek györgy gyorsan győző had hagy hall hamar hamarosan hány három has hasonló hasonlóan hat hát határozottan have helsingin hely helyi helyzet hét heves hiába hihetetlen hír hisz hivatalosan hivatkozás hogyan hol holnap holott hónap honnan hoz hozzászólás I idén idéz idő időnként igaz igazán igazgatónő igen igencsak igenis így ii illetően ily ilyen ilyenkor immár in indított indul info információ inkább innen intézmény ír irány írás is ismer ismét istván it itt itthon ivan iván jakab János január jár jelen jelenleg jelent jelentősen jó jobb jobban jól jön jövő jövőre józsef július június jut kanadai kap kapcsolat kapcsolódó károly kártya kedves kép képes kér kérdés kérdez kerítés kerül kerületi később készül két kettő kevésbé kéz kezd ki kiba kicsi kicsit kifejezetten kijelentés kilenc kis kissé kivéve kizárólag ko kommentár komolyan könnyen könyv kor kör korábban körülbelül köszön köszönhetően következő követően közben közel közöl közösen közvetlenül külföld külön különben különösen kuruc ku-

rucok kutya lajos lak lakó lap lassan lászló lát látható legalább legalábbis legfeljebb leginkább lehet lényegesen lép levél levelek lévő link ma maga magyarul mai majd majdnem május már marad március marian máris más másik másként másrészt még megfelelő megfelelően megint mégis mégiscsak meglehetősen mégsem megy megye megyei mely melyik mennyi mennyire mentős mer mesél messze méter mi mielőtt miért míg miként miklós mikor miközben milyen milyen mind minden mindenekelőtt mindenesetre mindenképpen mindenki mindenütt mindez mindig mindjárt mindkét mindössze minél mintegy mire miután mivel mód mokka molnár mond most mti nagy nagyjából nagyon nap naponta ne négy néha néhány nehezen nem nemcsak németh némi nemrég netán név nevez new néz néző nincs nő noha november nyilván nyilvánvalóan nyolc nyugodtan ő of ok ők őket oktatás október olaszország oldal oli olvas olvasható olvasó olvasói olvasottabb oly olyan olykor on önkormányzati önmaga onnan or óra ország országos összes összesen oszkár osztály öt ott párhuzamosan pedagógus példa például perc péter police pontosabban pontosan próbál pusztán rá ráadás raffael régen reggel régóta reklám rendkívül rendszeresen rész részben ritkán rögtön románia romániai rossz rosszul s said sajnos sándor se segít sehol sem soha sohasem sokáig sokan sokszor sor sosem szabad szabadon száll szám számomra számukra számunkra szem személy személyesen szép szépen szeptember szeret szerkesztőség szint szinte szintén szívesen szó szól szóval szóvivő szükség tag talál található talán tanárnő tanulmány társ társaság tart tavaly távol te tegnap település teljes teljesen téma tényleg természetes természetesen terület tesz tetszik tett that the their they this ti tibor tímea tisztelt to többé többen többi többnyire többször topic tör történet történik történt tovább továbbá továbbra tud túl tulajdonképpen tulajdonos túlságosan tűz üdv udvar üdvözlet úgy úgy ügy ugyan ugyanaz ugyancsak ugyanolyan ugyanúgy ugye úgyhogy új újra ül úr út utána utca utóbbi vág valaha valahogy valahol valaki valamely valamelyik valami válasz válik való valóban valójában valószínűleg változatlanul vár várhatóan város vég végre végül vélemény véletlenül vendég vesz vezető videó világ visz viszonylag vonatkozóan was were with zoltán zsidóbűnözés

I. NEM MAGYAR SZÖVEGRÉSZLETEKET TARTALMAZÓ CIKKEK LISTÁJA

Angol szövegrészleteket tartalmazó cikkek azonosítója: 10149, 1032, 10363, 10420, 1122, 1196, 1224, 1357, 1451, 1510, 181, 1812, 1831, 2172, 2479, 2558, 256, 2591, 2750, 2785, 2815, 2884, 2913, 2915, 2941, 3004, 3033, 3304, 3352, 3357, 3384, 3435, 3506, 3759, 3856, 4148, 4383, 4560, 4627, 4713, 4850, 4906, 5433, 5548, 5653, 5878, 6042, 6061, 6096, 6124, 6269, 6436, 6441, 6507, 6767, 7016, 7072, 7104, 713, 724, 7368, 746, 7528, 7531, 7665, 7787, 7846, 809, 8815, 8904, 8944, 9006, 9079, 9251, 9377, 9884, 9896, 9931, 9965, 9967

Finn szövegrészleteket tartalmazó cikkek azonosítója: 10078, 1334, 1361, 1427, 155, 1569, 1575, 2195, 2318, 2336, 2377, 3111, 3209, 3297, 3322, 3336, 3546, 3681, 4083, 4254, 4275, 4435, 4609, 4619, 5212, 5523, 5707, 5787, 5953, 5987, 6068, 656, 6682, 6870, 7356, 7557, 7742, 782, 7833, 7986, 8044, 8282, 8512, 8678, 8882, 9023, 9374, 9658, 9683

Francia szövegrészleteket tartalmazó cikkek azonosítója: 8069

Német szövegrészleteket tartalmazó cikkek azonosítója: 10084, 10335, 4197, 5875, 612, 6287, 6306, 6796, 7935, 84, 8729, 9269, 9591, 9982

Olasz szövegrészleteket tartalmazó cikkek azonosítója: 7050, 8355

Román szövegrészleteket tartalmazó cikkek azonosítója: 196

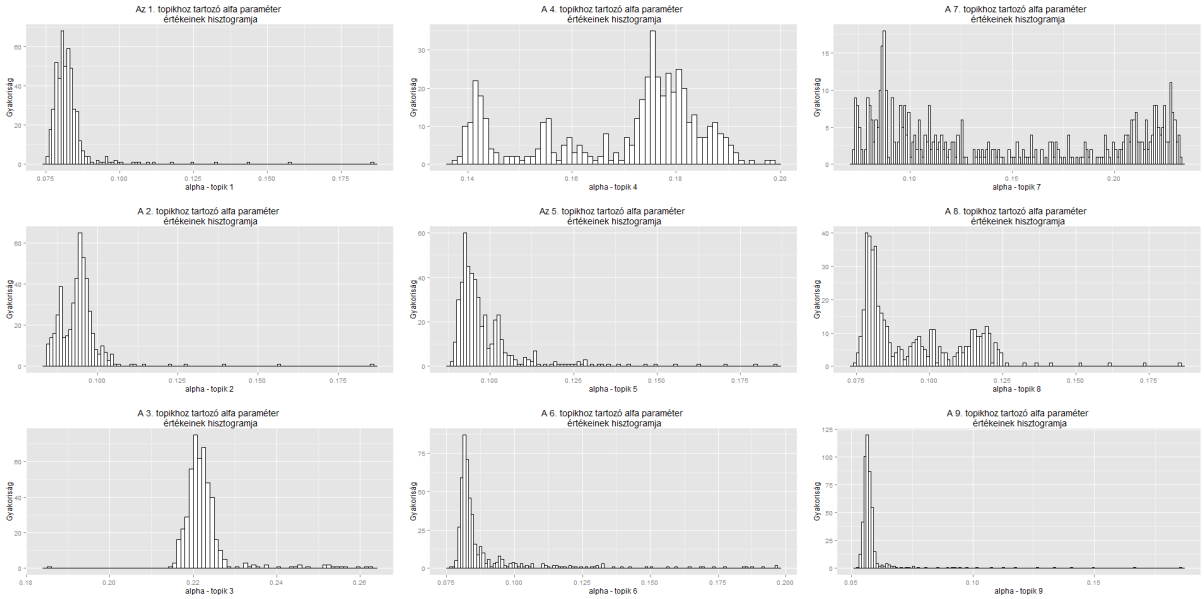
Spanyol szövegrészleteket tartalmazó cikkek azonosítója: 1748, 3738, 5762

Svéd szövegrészleteket tartalmazó cikkek azonosítója: 4536, 7378

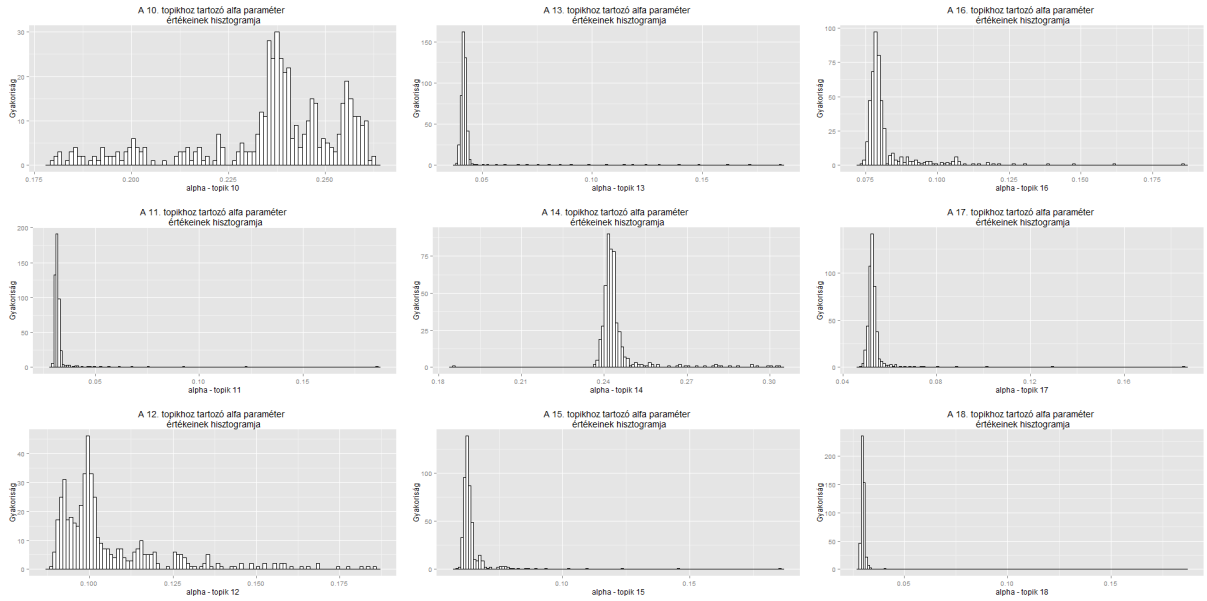
Szerb szövegrészleteket tartalmazó cikkek azonosítója: 8225

Törölt fájlok azonosítója: 6441, 5762, 9079

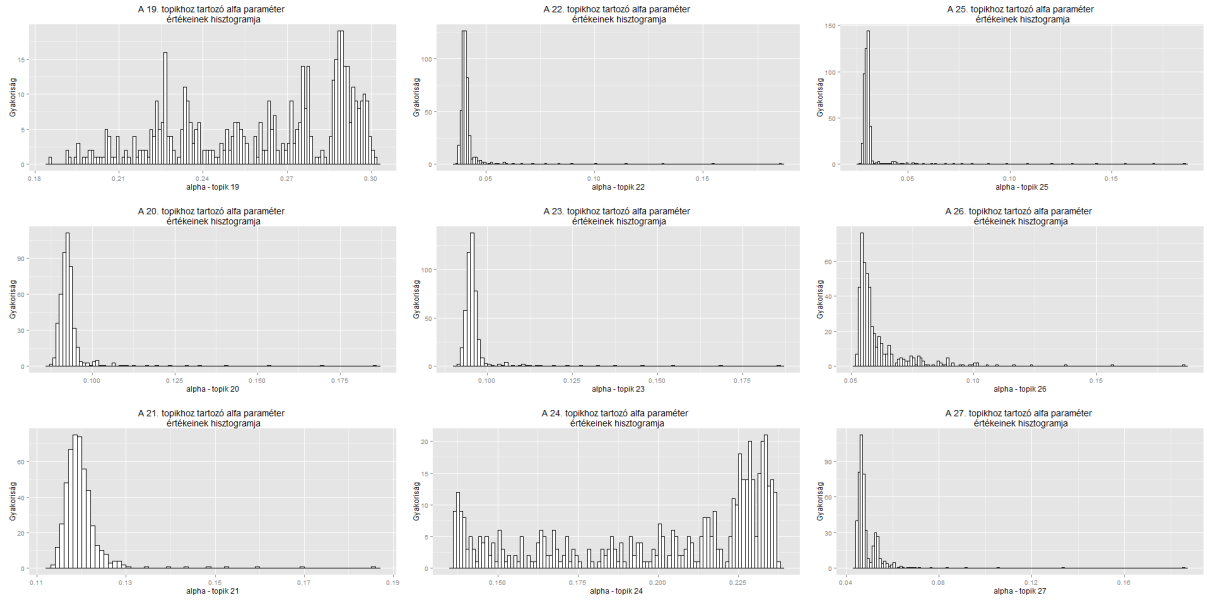
J. GIBBS-MINTAVÉTEL KONVERGENCIÁJÁNAK ELLENŐRZÉSE - ÁBRÁK



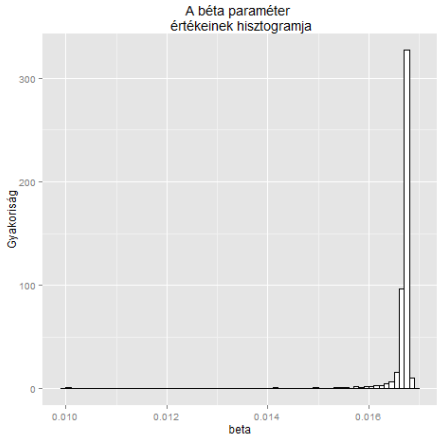
J.1. ábra. Az 1-9. topik α hiperparaméter értékeinek hisztogramjai - 1. lánc, aszimmetrikus α



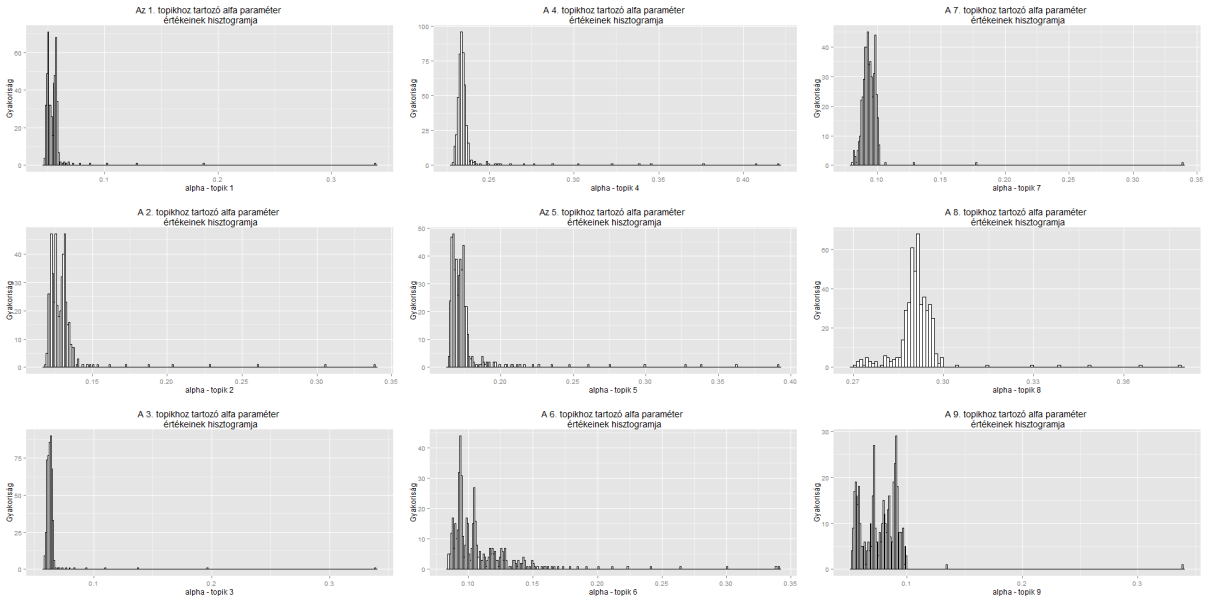
J.2. ábra. A 10-18. topik α hiperparaméter értékeinek hisztogramjai - 1. lánc, aszimmetrikus α



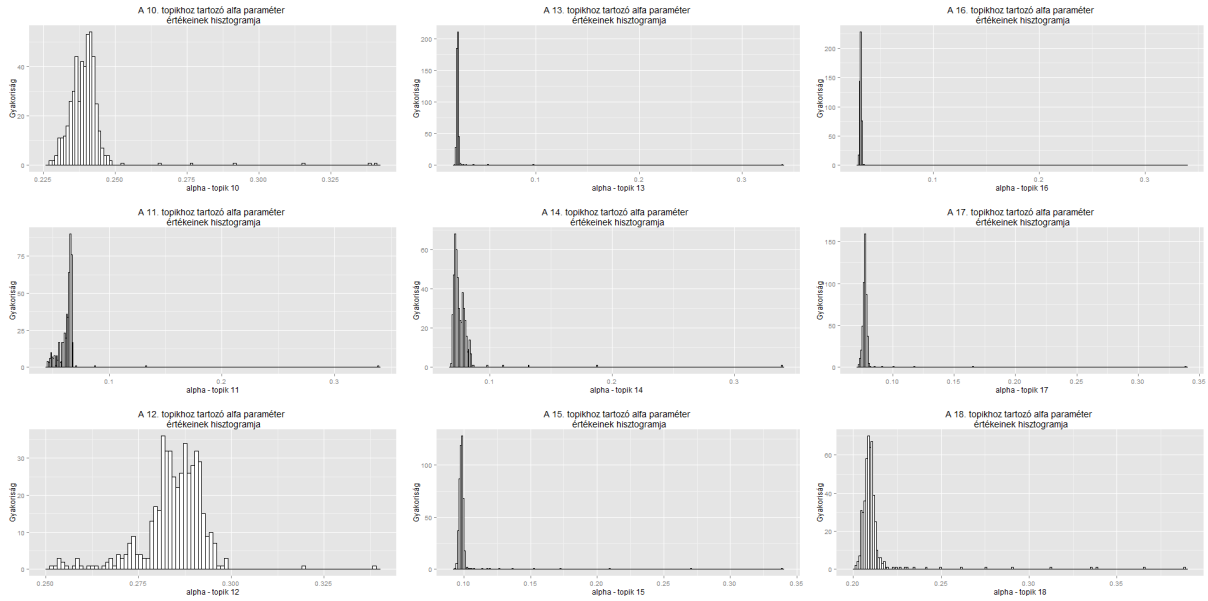
J.3. ábra. A 19-27. topik α hiperparaméter értékeinek hisztogramjai - 1. lánc, aszimmetrikus α



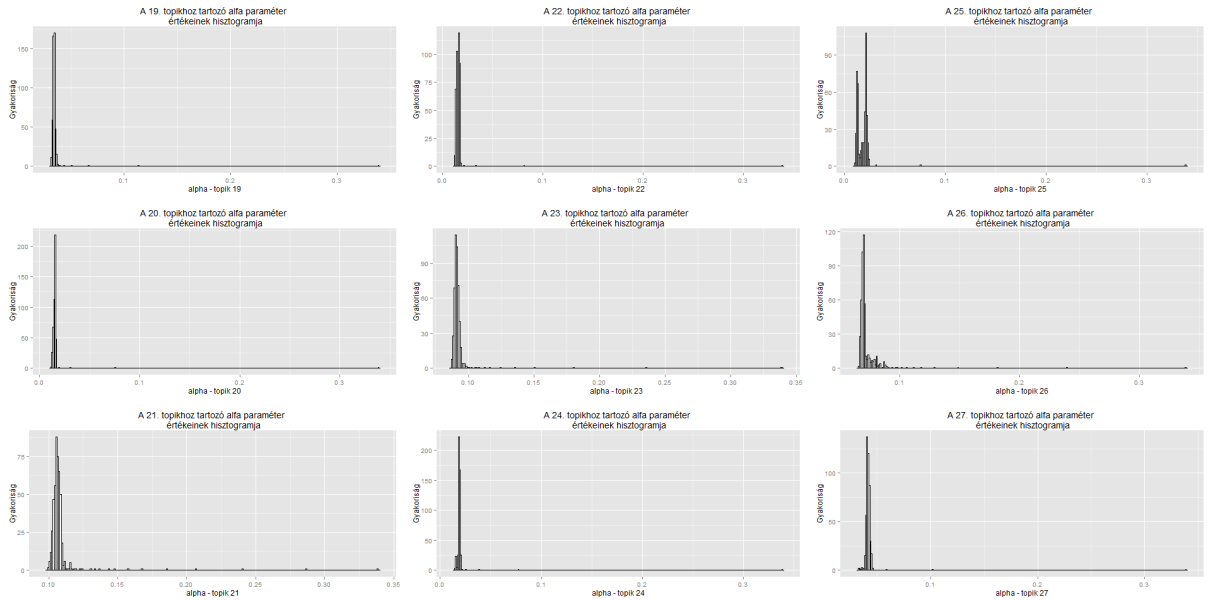
J.4. ábra. A β hiperparaméter értékeinek hisztogramja - 1. lánc, aszimmetrikus α



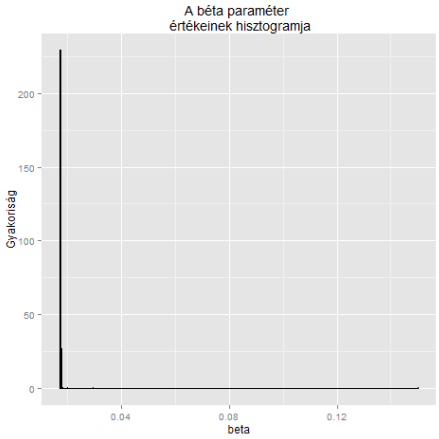
J.5. ábra. Az 1-9. topik α hiperparaméter értékeinek hisztogramjai - 2. lánc, aszimmetrikus α



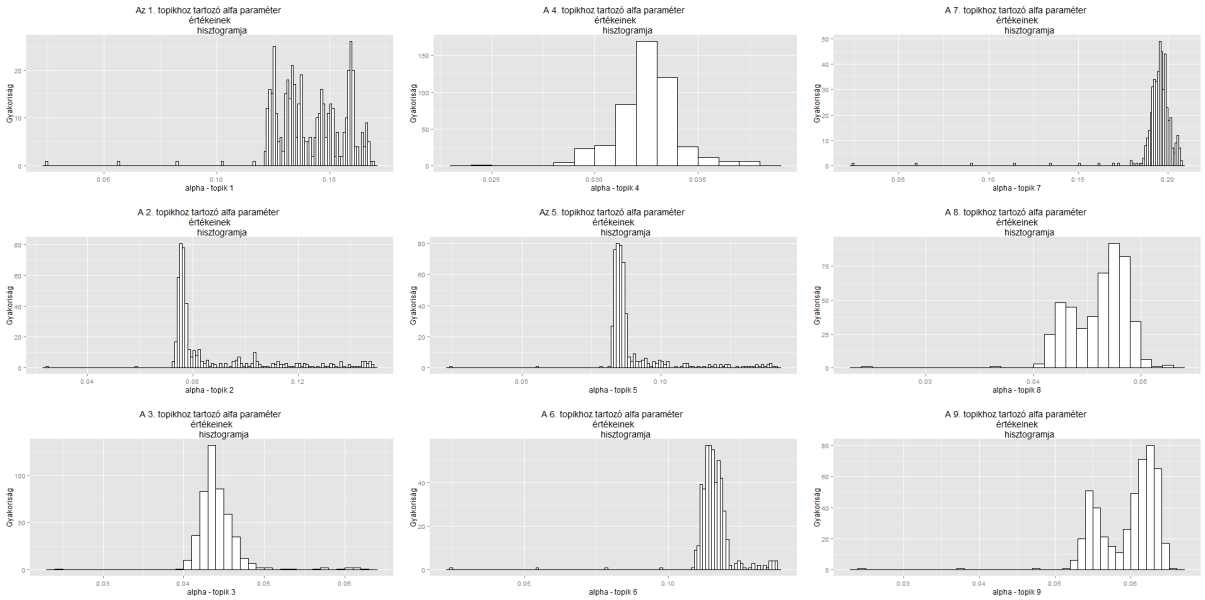
J.6. ábra. A 10-18. topik α hiperparaméter értékeinek hisztogramjai - 2. lánc, aszimmetrikus α



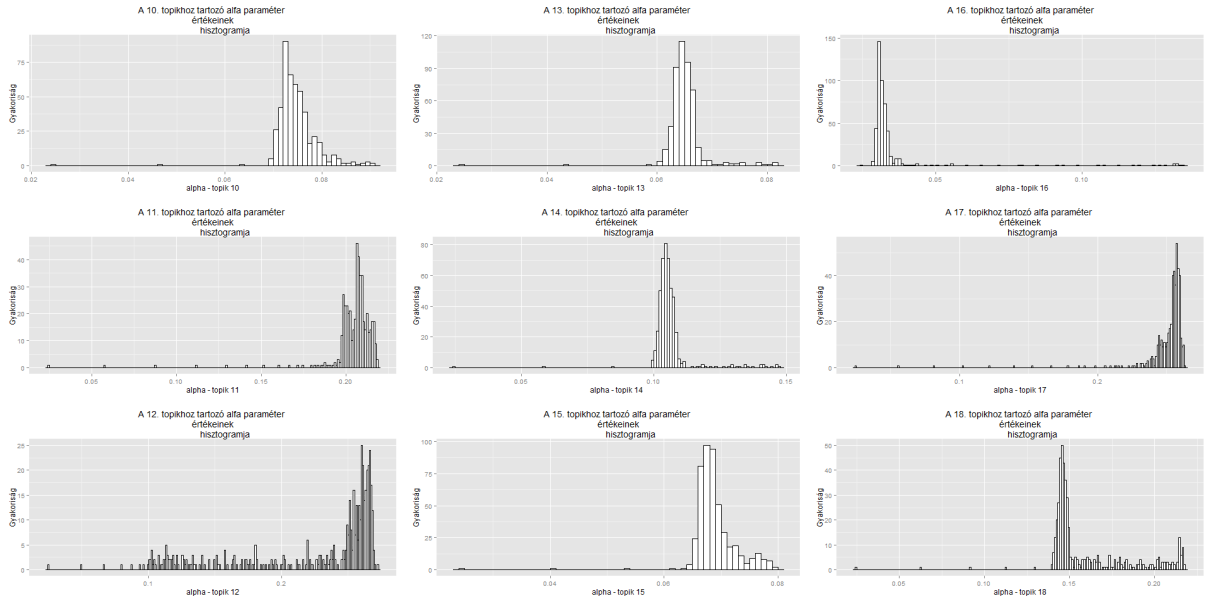
J.7. ábra. A 19-27. topik α hiperparaméter értékeinek hisztogramjai - 2. lánc, aszimmetrikus α



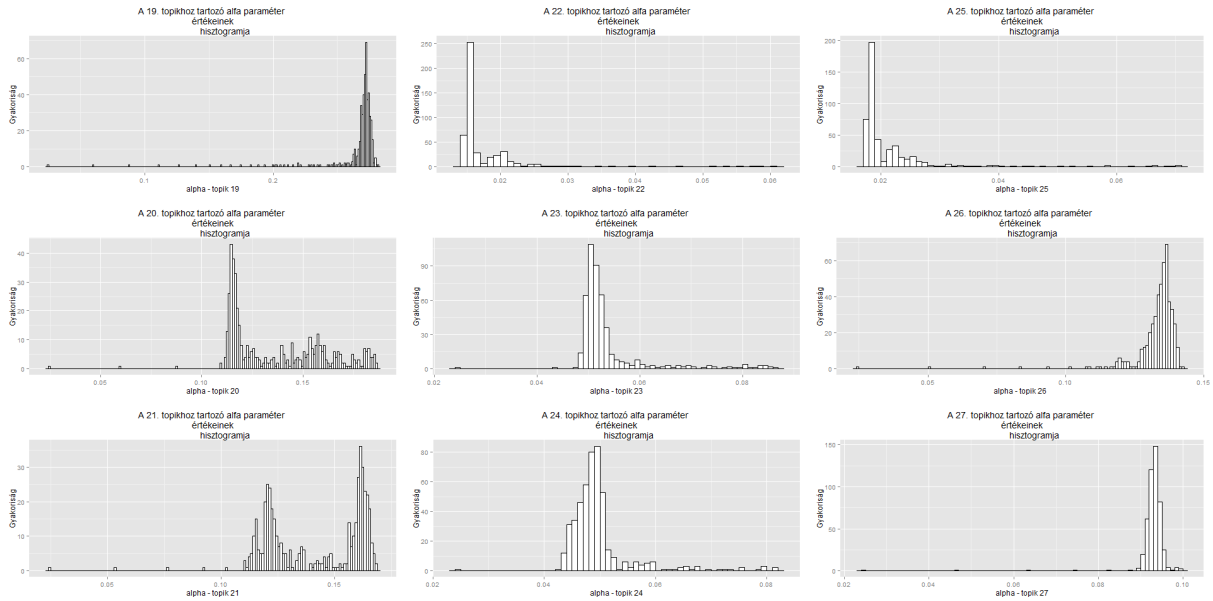
J.8. ábra. A β hiperparaméter értékeinek hisztogramja - 2. lánc, aszimmetrikus α



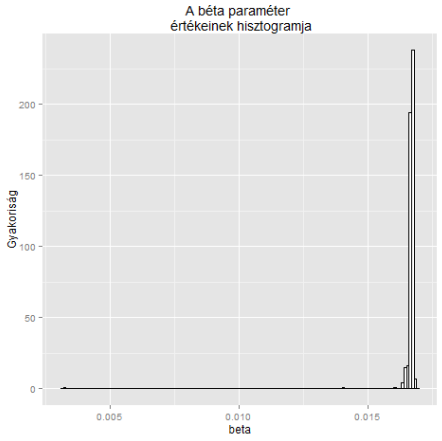
J.9. ábra. Az 1-9. topik α hiperparaméter értékeinek hisztogramjai - 3. lánc, aszimmetrikus α



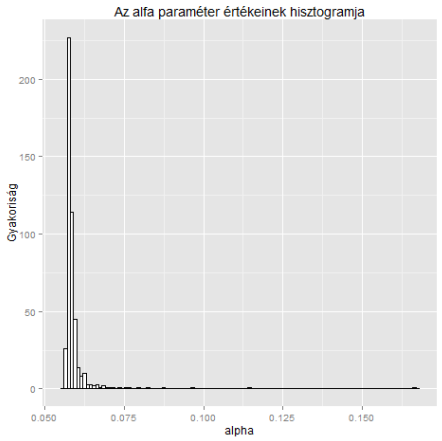
J.10. ábra. A 10-18. topik α hiperparaméter értékeinek hisztogramjai - 3. lánc, aszimmetrikus α



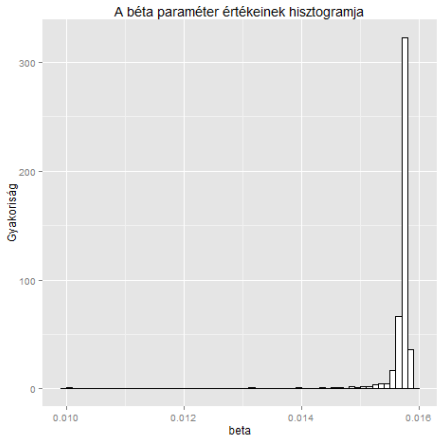
J.11. ábra. A 19-27. topik α hiperparaméter értékeinek hisztogramjai - 3. lánc, aszimmetrikus α



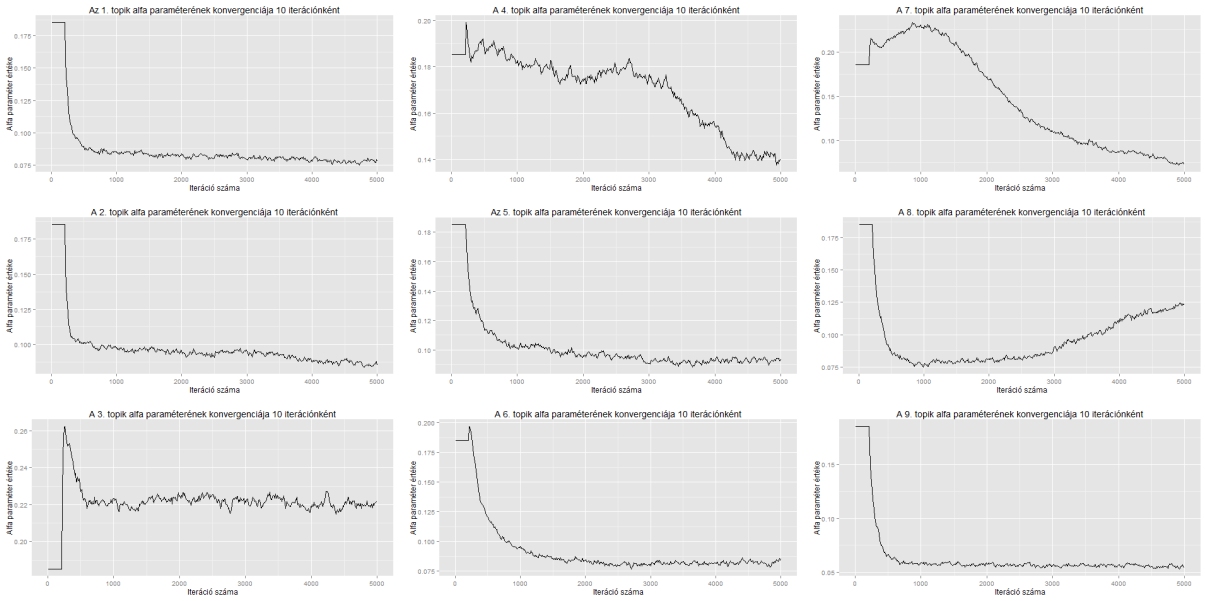
J.12. ábra. A β hiperparaméter értékeinek hisztogramja - 3. lánc, aszimmetrikus α



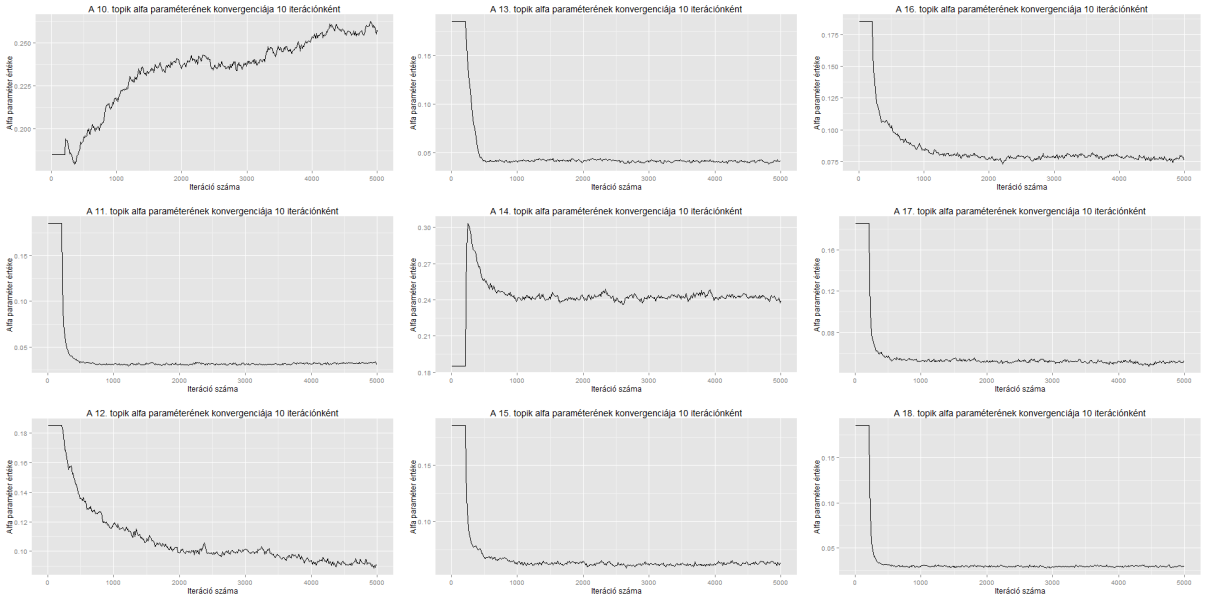
J.13. ábra. Az α hiperparaméter értékeinek hisztogramja - szimmetrikus α



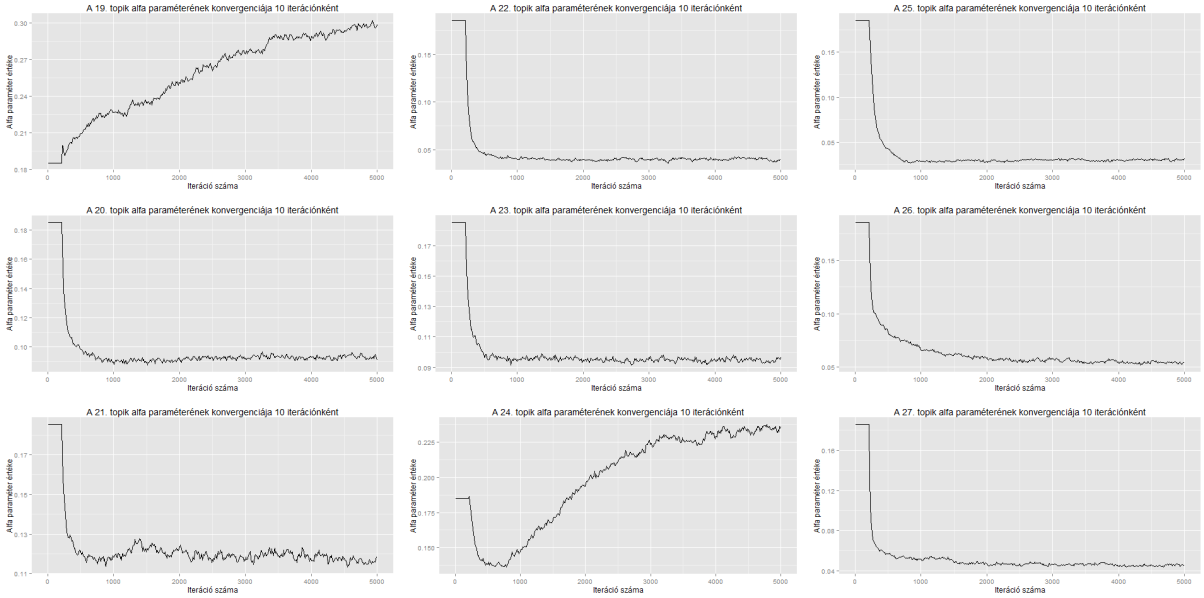
J.14. ábra. A β hiperparaméter értékeinek hisztogramja - szimmetrikus α



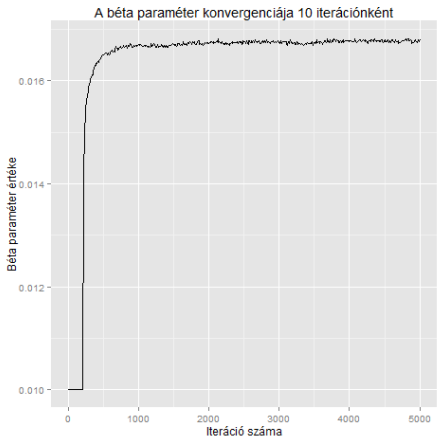
J.15. ábra. Az 1-9. topik α paraméterének konvergenciája 10 iterációnként - 1. lánc, aszimmetrikus α



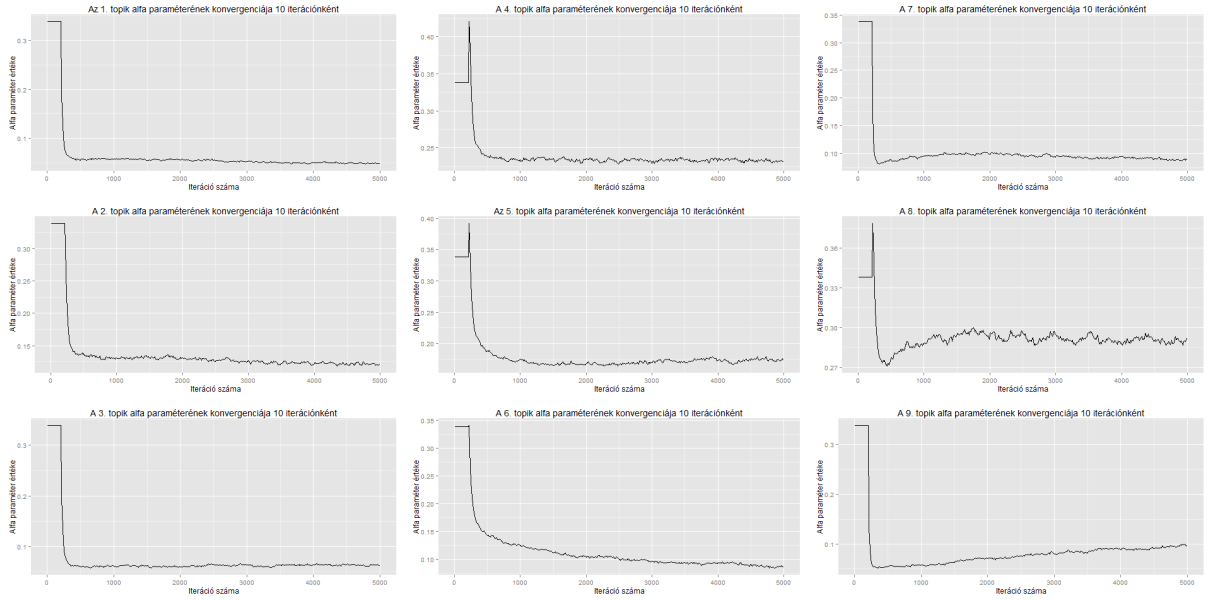
J.16. ábra. A 10-18. topik α paraméterének konvergenciája 10 iterációnként - 1. lánc, aszimmetrikus α



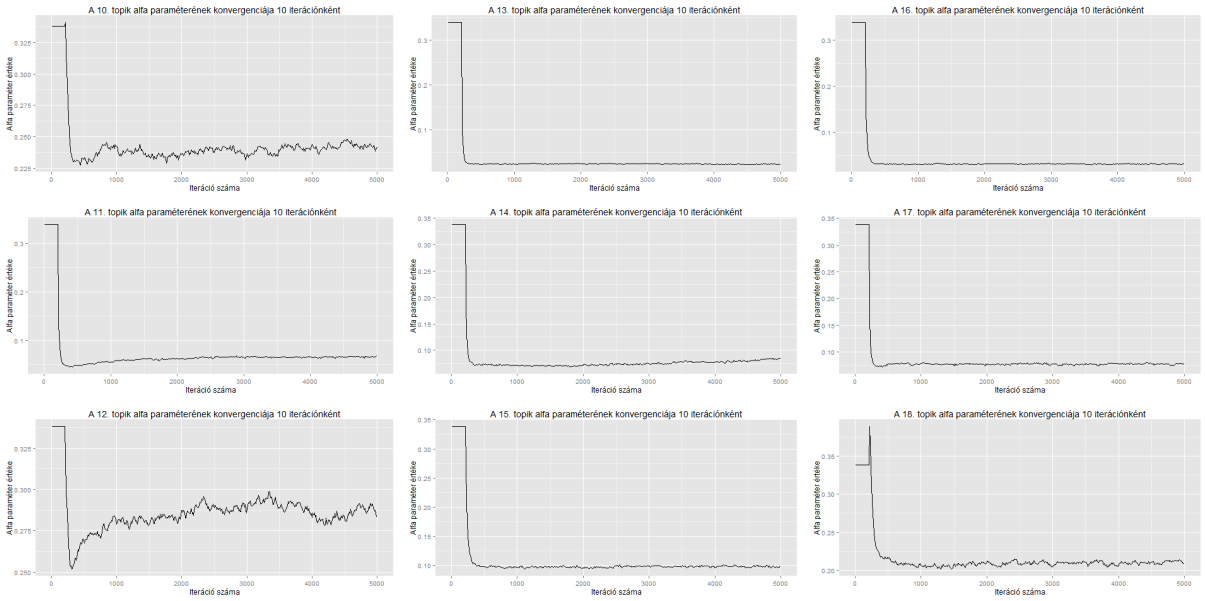
J.17. ábra. A 19-27. topik α paraméterének konvergenciája 10 iterációnként - 1. lánc, aszimmetrikus α



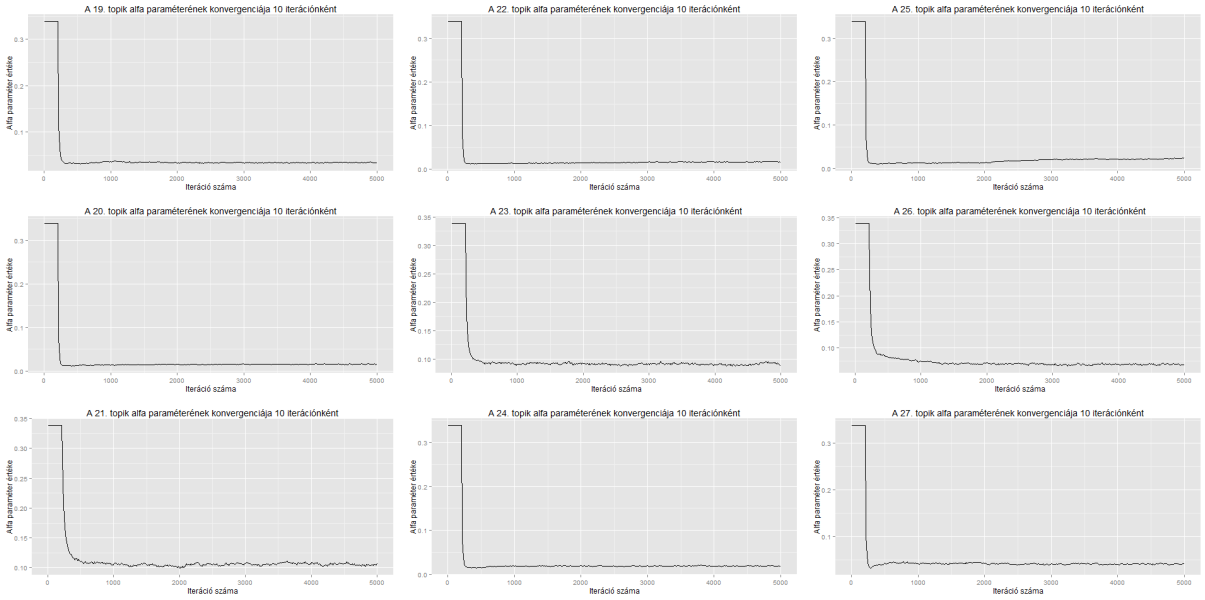
J.18. ábra. A β paraméter konvergenciája 10 iterációnként - 1. lánc, aszimmetrikus α



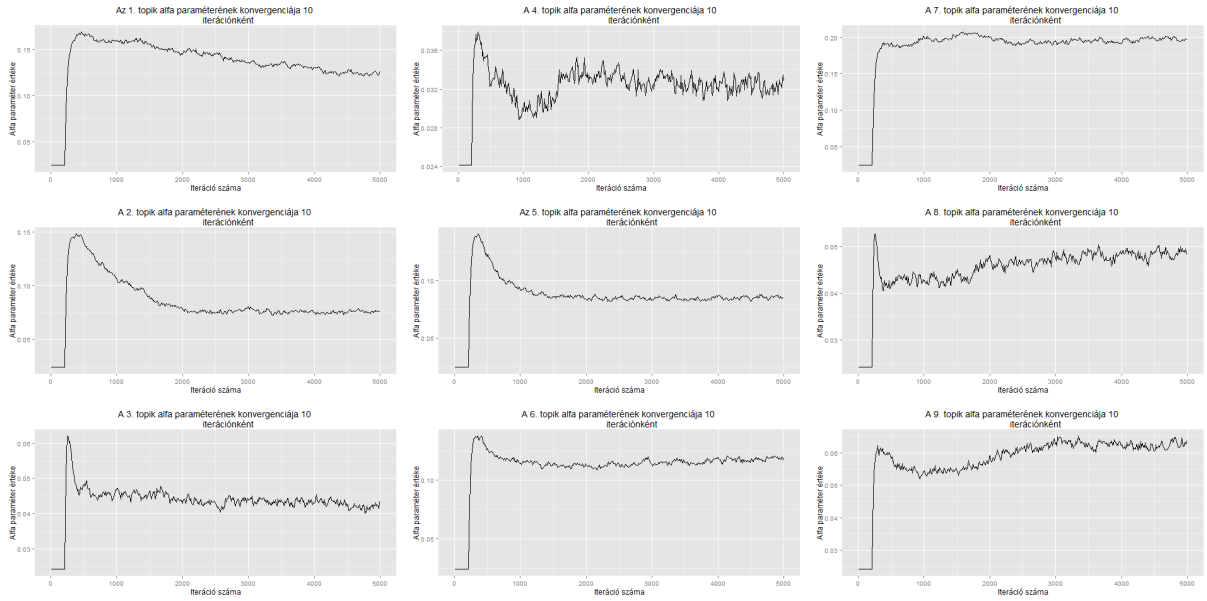
J.19. ábra. Az 1-9. topik α paraméterének konvergenciája 10 iterációnként - 2. lánc, aszimmetrikus α



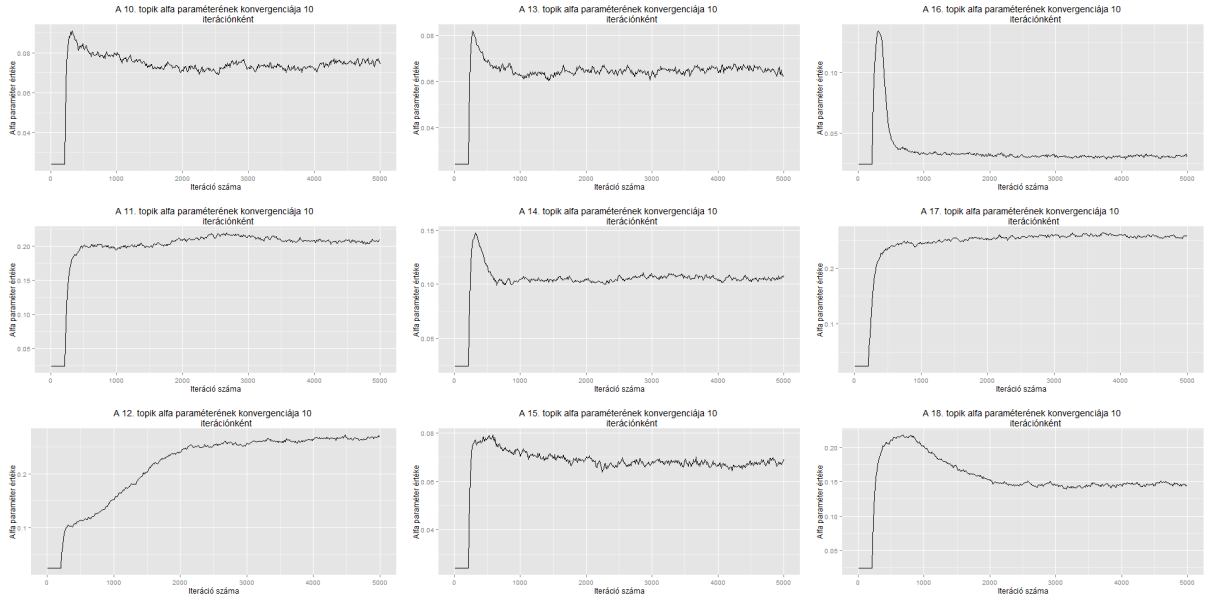
J.20. ábra. A 10-18. topik α paraméterének konvergenciája 10 iterációnként - 2. lánc, aszimmetrikus α



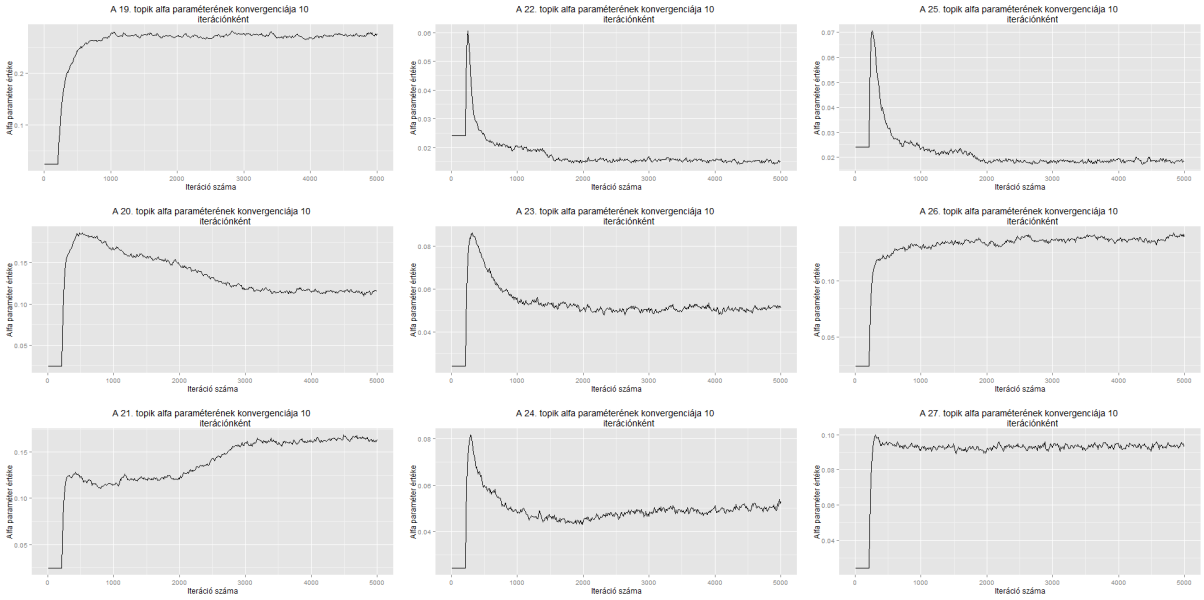
J.21. ábra. A 19-27. topik α paraméterének konvergenciája 10 iterációnként - 2. lánc, aszimmetrikus α



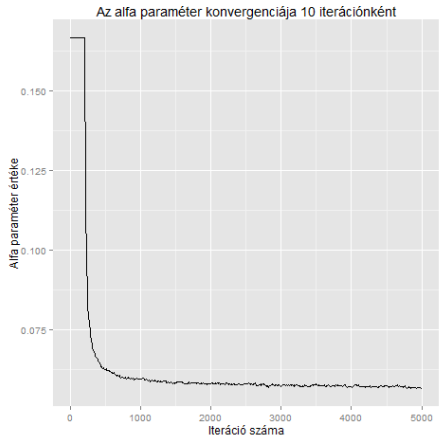
J.22. ábra. Az 1-9. topik α paraméterének konvergenciája 10 iterációnként - 3. lánc, aszimmetrikus α



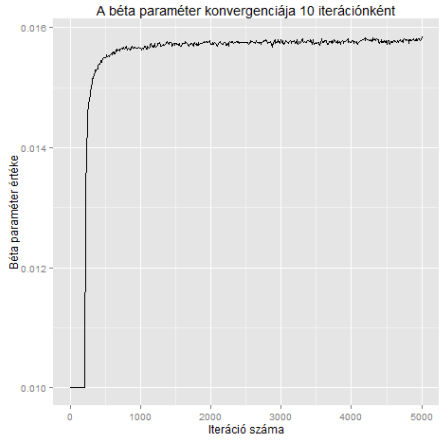
J.23. ábra. A 10-18. topik α paraméterének konvergenciája 10 iterációnként - 3. lánc, aszimmetrikus α



J.24. ábra. A 19-27. topik α paraméterének konvergenciája 10 iterációnként - 3. lánc, aszimmetrikus α

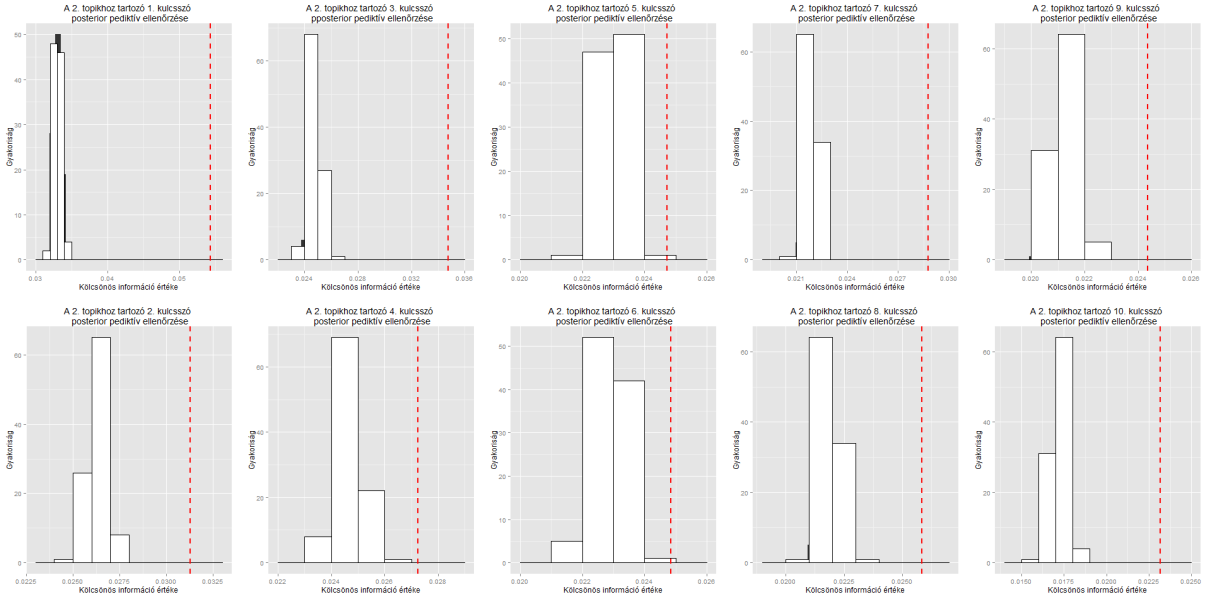


J.25. ábra. Az α paraméter konvergenciája 10 iterációnként - szimmetrikus α

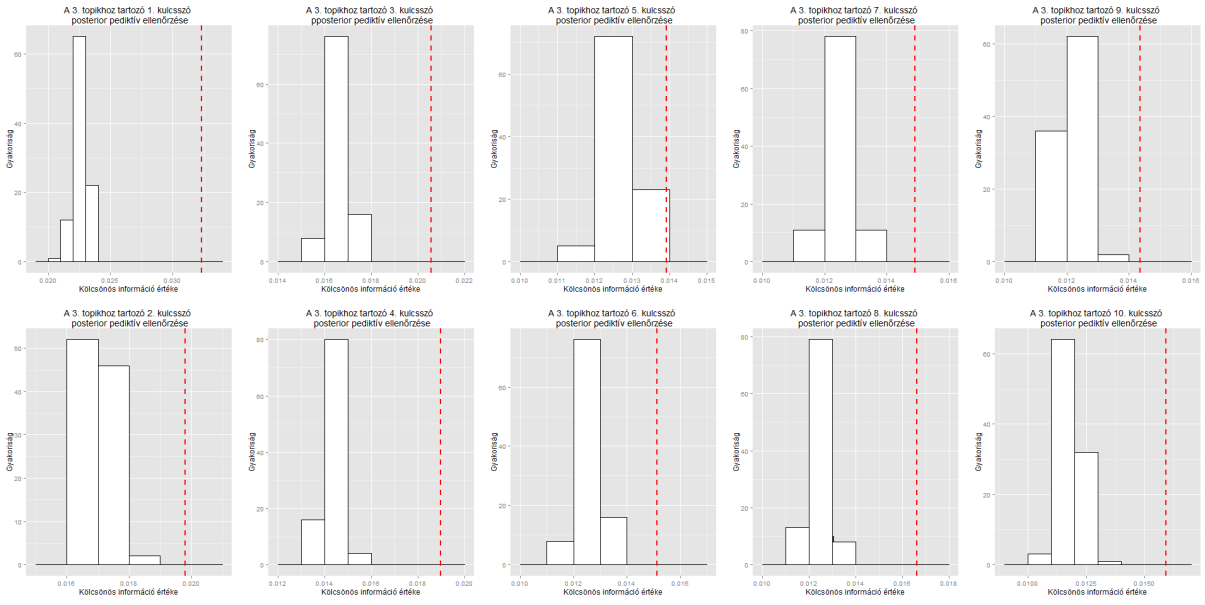


J.26. ábra. A β paraméter konvergenciája 10 iterációnként - szimmetrikus α

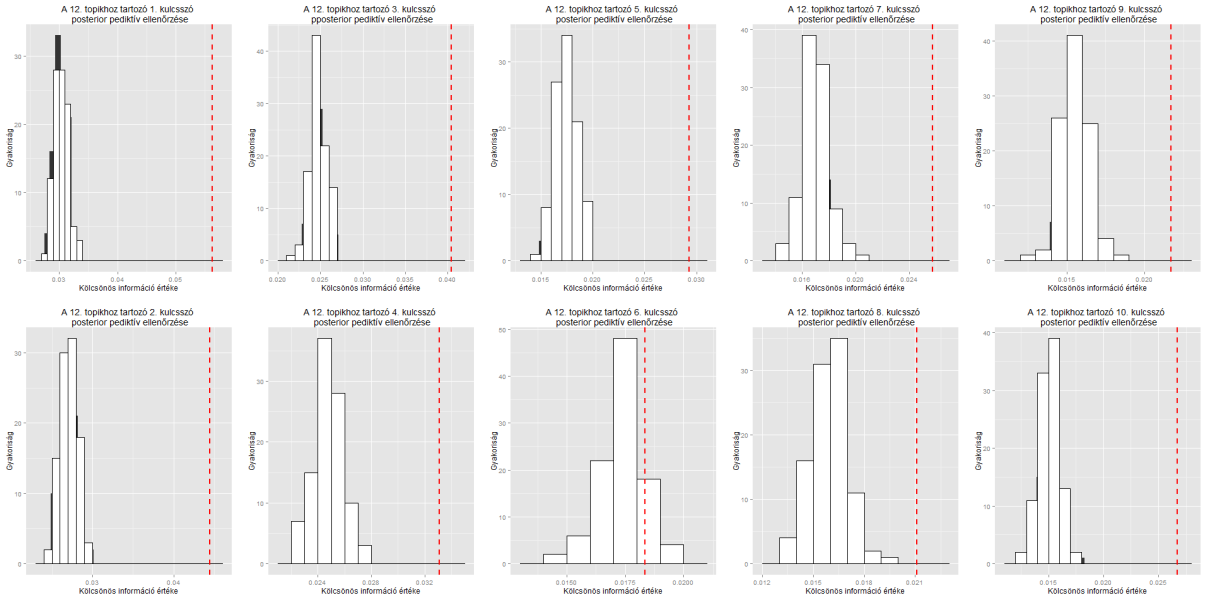
K. POSTERIOR PREDIKTÍV ELLENŐRZÉS - ÁBRÁK



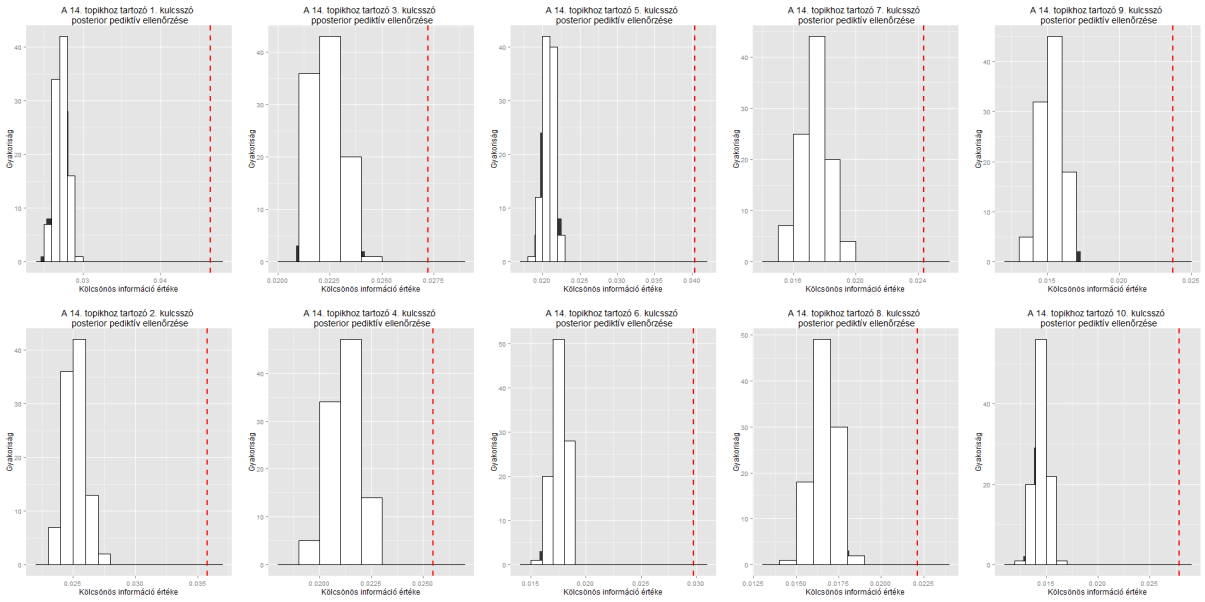
K.1. ábra. Posterior prediktív ellenőrzés [Mimno and Blei, 2011] és [Gelman and Meng, 2005] alapján - 1. lánc, aszimmetrikus α, 3. topik



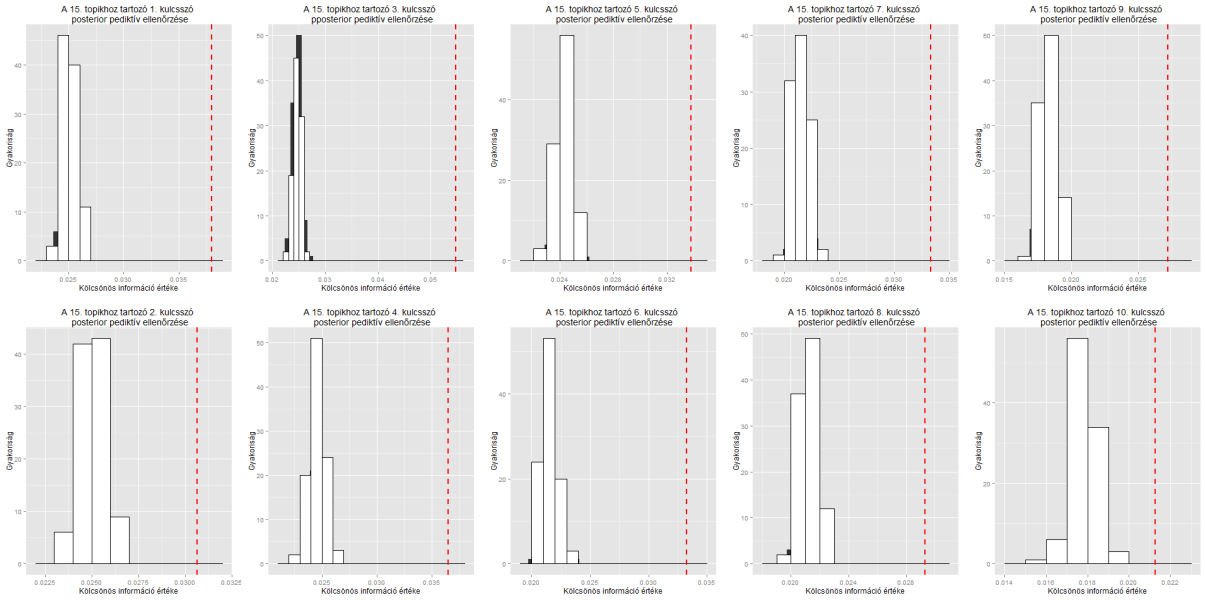
K.2. ábra. Posterior prediktív ellenőrzés [Mimno and Blei, 2011] és [Gelman and Meng, 2005] alapján - 1. lánc, aszimmetrikus α, 4. topik



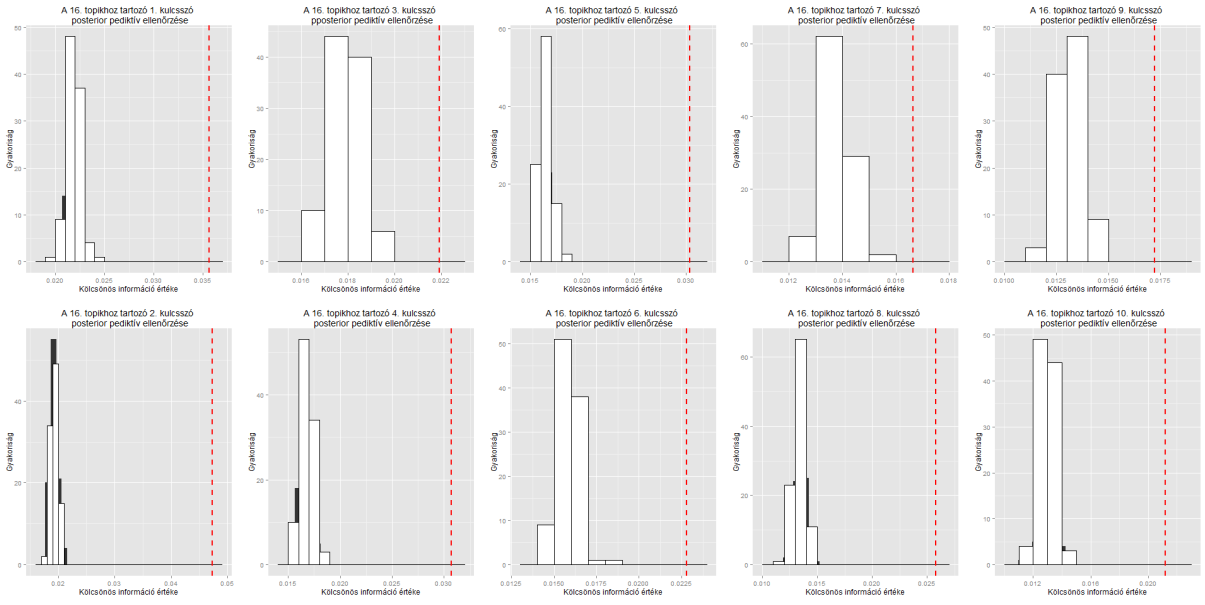
K.3. ábra. Posterior prediktív ellenőrzés [Mimno and Blei, 2011] és [Gelman and Meng, 2005] alapján - 1. lánc, aszimmetrikus α, 13. topik



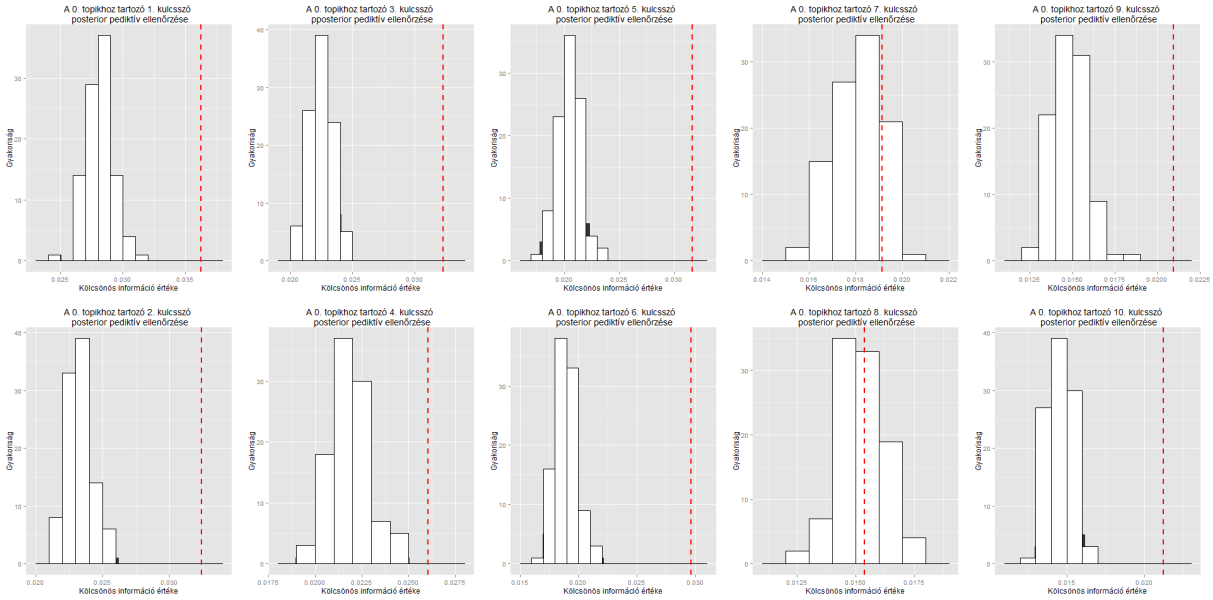
K.4. ábra. Posterior prediktív ellenőrzés [Mimno and Blei, 2011] és [Gelman and Meng, 2005] alapján - 1. lánc, aszimmetrikus α, 15. topik



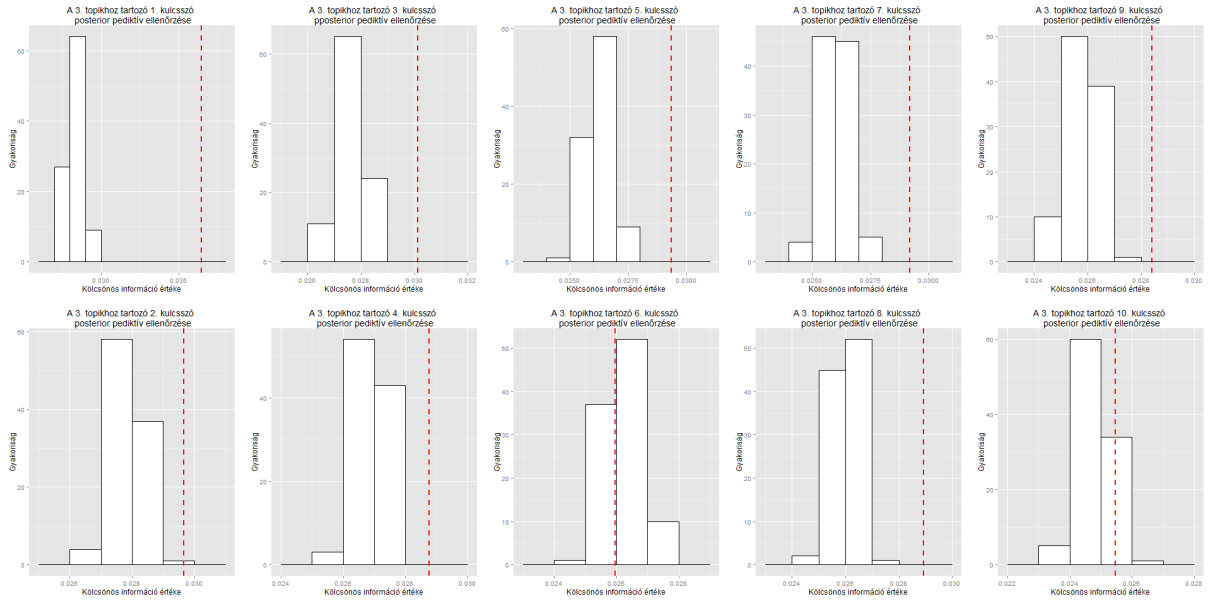
K.5. ábra. Posterior prediktív ellenőrzés [Mimno and Blei, 2011] és [Gelman and Meng, 2005] alapján - 1. lánc, aszimmetrikus α, 16. topik



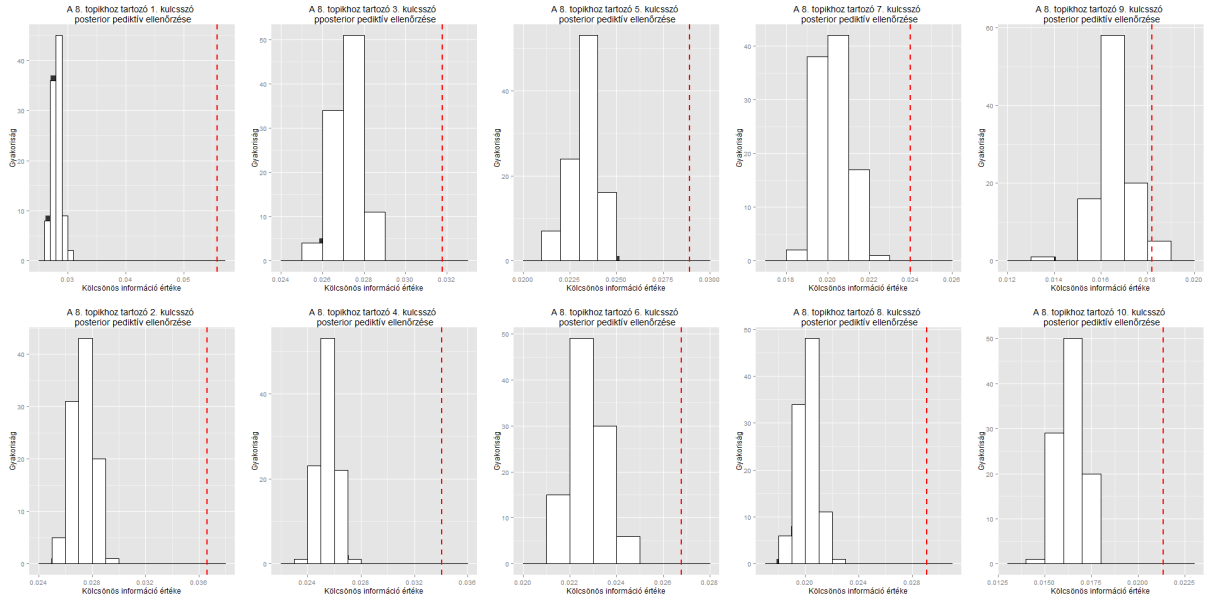
K.6. ábra. Posterior prediktív ellenőrzés [Mimno and Blei, 2011] és [Gelman and Meng, 2005] alapján - 1. lánc, aszimmetrikus α, 17. topik



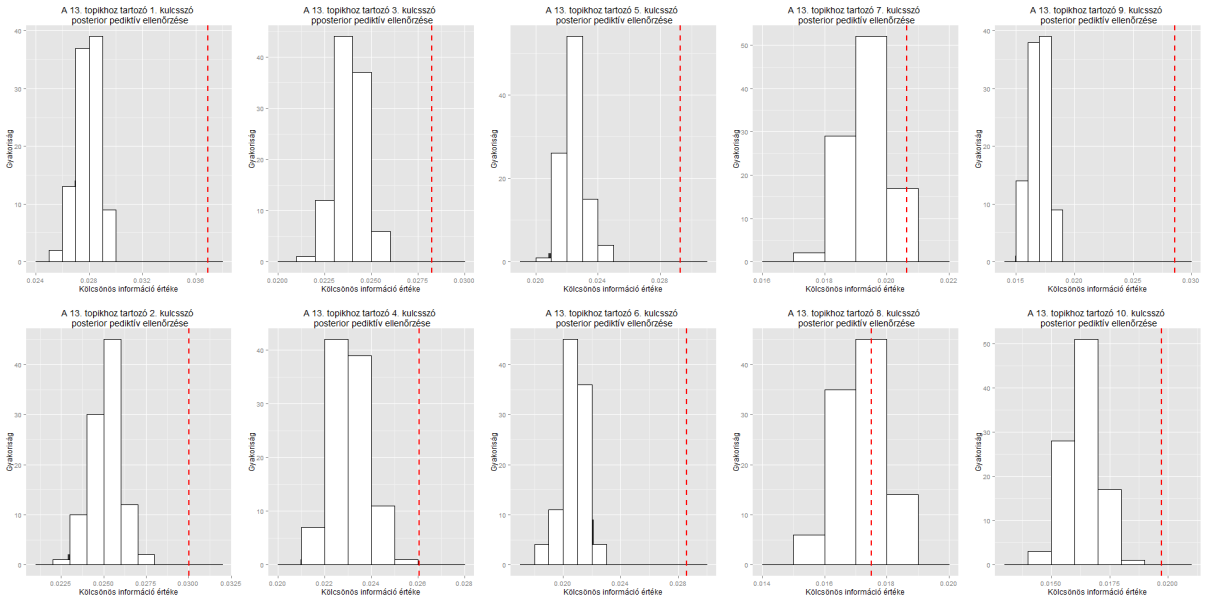
K.7. ábra. Posterior prediktív ellenőrzés [Mimno and Blei, 2011] és [Gelman and Meng, 2005] alapján - szimmetrikus α, 1. topik



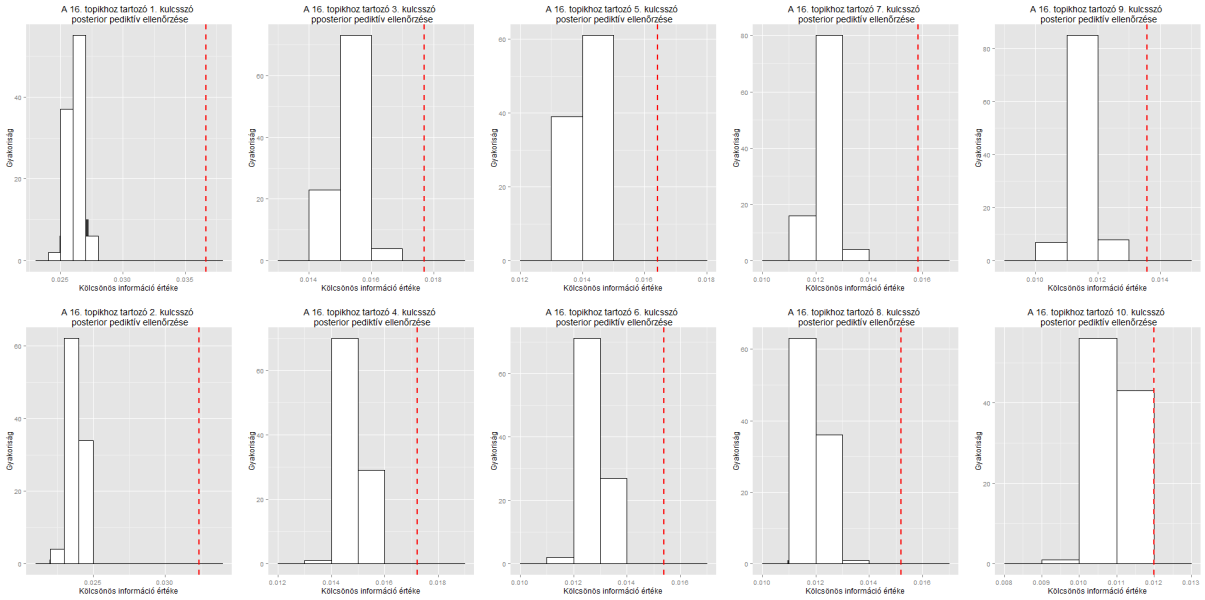
K.8. ábra. Posterior prediktív ellenőrzés [Mimno and Blei, 2011] és [Gelman and Meng, 2005] alapján - szimmetrikus α, 4. topik



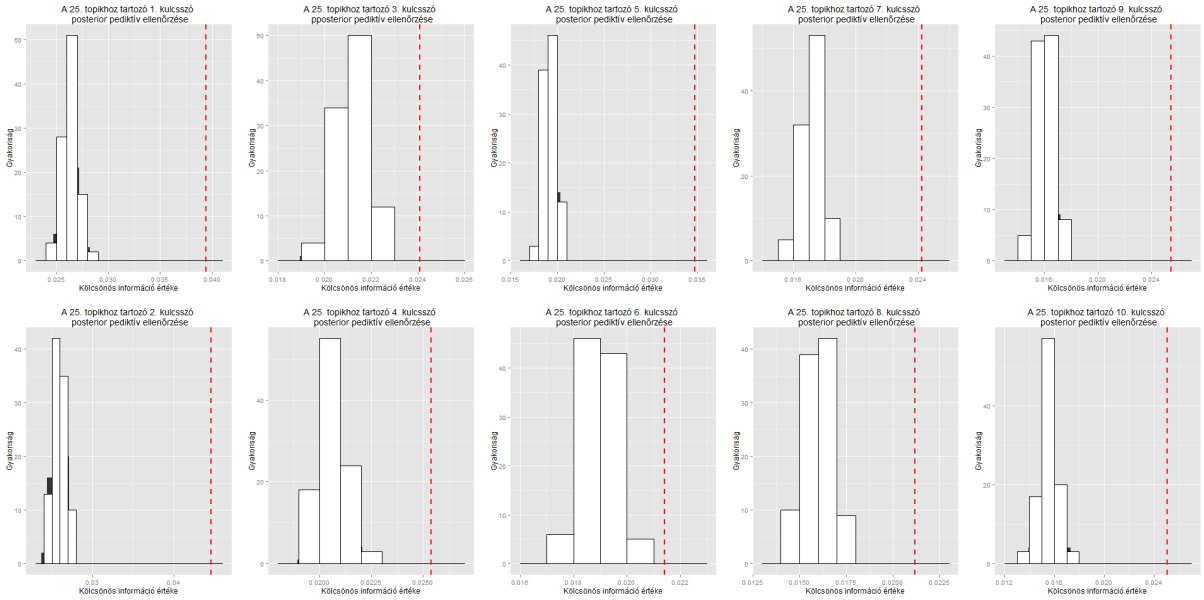
K.9. ábra. Posterior prediktív ellenőrzés [Mimno and Blei, 2011] és [Gelman and Meng, 2005] alapján - szimmetrikus α, 9. topik



K.10. ábra. Posterior prediktív ellenőrzés [Mimno and Blei, 2011] és [Gelman and Meng, 2005] alapján - szimmetrikus α, 14. topik



K.11. ábra. Posterior prediktív ellenőrzés [Mimno and Blei, 2011] és [Gelman and Meng, 2005] alapján - szimmetrikus α, 17. topik



K.12. ábra. Posterior prediktív ellenőrzés [Mimno and Blei, 2011] és [Gelman and Meng, 2005] alapján - szimmetrikus α, 26. topik

L. LÁTENS DIRICHLET ALLOKÁCIÓ MODELLEK

TOPIKJAIHOZ TARTOZÓ ELSŐ 30 KULCSSZÓ

Topik	α értéke	Kulcsszavak
0	0.07821	iskola gyerek szülő tanár diák tanuló igazgató iskolai lány oktatási fiatal osztálytárs család erőszak óvoda kislány probléma hátrányos helyzetű mohácsi iskolás felnőtt kisfiú szakember középiskola szülői önkormányzat suli gimnázium nevelő
1	0.0842	segély munka szociális pénz támogatás önkormányzat polgármester program forint család rendszer lehetőség gyerek állami monok közmunka állam családi összeg pótlék szepessy törvény százalék munkahely rendszeres monoki juttatás dolgozó kormány munkanélküli
2	0.22167	rendőr cigány autó rendőrség helyszín sérülés támadó kórház súlyos baleset eljárás kocsi család mentő támadás életveszélyes garázdaság fiatal verekedés könnyű testi kés segítség sofőr fiatalember járőr erőszak sértés rendőri sérült
3	0.13897	magyar politikai magyarország párt társadalom kormány cigányság zsidó politikus fidesz nemzet hatalom liberális állam jog rasszista politika bűnöző emberi haza jobbik szdsz probléma erőszak bűn többség média törvény mszp orbán
4	0.09567	lány gyilkosság áldozat rendőrség gyilkos cigány halál kislány elkövető brutális holttest nyomozás tettes bűncselekmény gyanúsított emberölés helyszín pest kiskunlacháza meggyilkolt család péntek magyar nyom rendőr nyomozó fiatal kiskunlacházi tragédia bors
5	0.0858	gárda magyar jobbik cigány gárdista roma szervezet rendezvény rendőrség demonstráció képviselő mozgalom elnök jobbikos nemzeti vona helyszín párt tüntetés békés polgármester anyaország magyarország fórum horváth cigányság cigánybűnöző felvonulás szombat sajóbábony
6	0.07199	cigány magyar zsidó rasszista tetves olaszliszka gyerek szegény cionista büdös olaszliszkai horda ávh paraszt réti jog néger kurva kb ártatlan segély lincselés szögi retek cigánybűnöző ft tanár mocsok lényeg becsületes
7	0.12186	cigány roma cigányság társadalom probléma százalék arány kultúra magyar társadalmi többség többségi közösség eredmény csoport kisebbség etnikai munka integráció magyarország iskola fiatal megoldás szegény európai fehér diszkrimináció család esély szociális
8	0.05444	ház roma támadás rendőrség gyilkosság elkövető család fegyver helyszín tűzoltó rasszista tatárszentgyörgy áldozat cigány tatárszentgyörgyi indíték pátkai pátka etnikai lövés támadó tettes nyomozás pécsi magyar polgárőr tarnabod nyomozó nagycsécs jenő
9	0.26019	rendőrség bűncselekmény törvény rendőr bűnöző érdek hatóság figyelem bűnözés munka rendőri lehetőség elkövető intézkedés jog védelem csoport közbiztonság probléma eredmény eszköz eljárás feladat rend adat szükséges rendőrkapitány vizsgálat százalék erőszakos
10	0.03233	vonat vasúti kábel busz utas kábellopás kár vonal máv villamos sín bkv hév közlekedés forint állomás vezeték lopás jegy szakasz kalauz biztonság közlekedési baleset ismeretlen kábeltolvaj ellenőr elkövető szerelvény vasút
11	0.0906	polgármester falu cigány roma polgárőr önkormányzat család rendőr lakos lakosság rendőrség község gyöngyöspata rend polgárőrség ház közbiztonság képviselő probléma magyar hivatal lopás idős bűnöző körzeti kisebbségi polgármesteri jelenlét egyesület rendőri
12	0.04188	orvos beteg kórház gyerek fogyatékos család egészségügyi betegség védőnő nyírmihálydi ellátás vizsgálat orvosi szülő állapot gyógyszer csecsemő súlyos terhes háziorvos vérfertőzés dolgozó családi mentő értelmi újszülött kislány genetikai szakember pótlék
13	0.23891	rendőrség bűncselekmény gyanúsított rendőrkapitányság idős rendőr elkövető előzetes eljárás őrizet gyanú rablás letartóztatás sértett elkövetés nyomozás megalapozott fiatal forint pénz ház lakás nyomozó áldozat bűntett ismeretlen tettes bűnügyi rabló fiatalkorú

L.1. táblázat: Látens Dirichlet allokáció topikjainak első 30 kulcsszava - aszimmetrikus α paraméter, 1. lánc, 1-14. topik, $\beta = 0.016784$

Topik	α értéke	Kulcsszavak
14	0.06174	lakás ház önkormányzat család miskolc miskolci ingatlan épület telep ózd szemét környék képviselő cigány lakótelep állapot probléma városrész ózdi avas telek szomszéd áram polgármester lépcsőház körülmény bérlakás megoldás segítség otthon
15	0.07706	roma cigány kolompár önkormányzat elnök orbán kisebbségi képviselő ocö szervezet forint farkas bizottság fidesz párt választás támogatás pénz politikus cigányság országgyűlési mszp fideszes európai gazdasági ülés program magyar drom lungo
16	0.05203	roma román cigány francia olasz európai kormány cseh hatóság német szlovák bevándorló szervezet tábor szlovákia állampolgár elnök külföldi bolgár uniós illegális sarkozy eu euró magyar németország közösség politikus unió bulgária
17	0.03071	cozma veszprém sztojka veszprémi román enyingi sportoló szórakozóhely rendőrség tanú verekedés gyilkosság kézilabdázó patrióta lokál siófoki vallomás szerb játékos bár szív gyanúsított siófok diszkó enying banda kés kézilabdás sesum vese
18	0.29905	cigány magyar roma fehér fajvédő bűnöző hazugság népszabadság cigányfajvédő származás náci tény gádzsó zsindex cigók médium balliberális média származású nemzet műsor kivétel komoly rasszista sajtó egyetlen gyakorlati gyilkosság figyelem mohácsi
19	0.09199	tolvaj forint kár lopás ismeretlen tettes érték rendőrség gazda temető rendőr elkövető víz fém templom forintos rongálás szolnok autó értékű darab okozott tanya fémtolvaj épület helyszín szolnoki betörő ellopott szobor
20	0.11795	fiatal biztonsági lány őr tér telefon kamera üzlet rendőrség elkövető cigány ismeretlen táska kerület budapest felvétel pénz fekete bolt környék kés fiatalember rablás körüli eladó telefonszám fegyver magas győri lopás
21	0.03976	magyar cigány kanada lány magyarország család hatóság rabszolga osztrák áldozat prostituált banda prostitúció róbert rendőrség magyarországi munka ausztria euró svájci szarvasi pénz strici bűnöző svájc emberkereskedelem holland futtató csoport rabszolgatartó
22	0.09664	vádlott bíróság ítélet tárgyalás börtön elkövetett büntetés vád bíró ügyészség jogerős sértett tanú vallomás emberölés előzetes cigány börtönbüntetés szabadságvesztés ügyvéd bűntett ügyész bűncselekmény eljárás per felfüggesztett elsőrendű vádirat védő testi
23	0.23613	ház cigány család gyerek falu pénz idős munka szomszéd feleség néni baj bácsi fia kert autó férj környék fiatal lakás otthon segítség rend hölgy testvér öreg lassú nehéz ismerős régi
24	0.03164	magyar film kultúra amerikai daróczi zene kulturális fiatal magyarország előadás india havas plakát árpád ismert művész műsor junior nyelv dokumentumfilm buddhista közönség amerika dopeman edző dal színész király szöveg orsós
25	0.05347	finn cigány amerikai bűnöző francia olasz svéd rendőrség börtön derék korcs tábor fegyver korcsok kivétel származás normálisabb állampolgárság usa söpredék tót finnországi kemény rendőr gyakorlat gyakorlati nép gyilkos bűnözés állam
26	0.04641	uzsorás pénz forint család kölcsön összeg tartozás adós kamat vajda uzsora hitel apeh mali tiszabő rendőrség ingatlan bank ház polgármester szerződés győzike csalás havi számla forintos ft zsarolás lakatos tevékenység

L.2. táblázat: Látens Dirichlet allokáció topikjainak első 30 kulcsszava - aszimmetrikus α paraméter, 1. lánc, 15-27. topik, $\beta = 0.016784$

Topik	α értéke	Kulcsszavak
0	0.04977	roma román cigány francia olasz európai cseh kormány szlovák tábor hatóság szervezet állampolgár bevándorló szlovákia bolgár illegális közösség euró sarkozy külföldi eu elnök csehország uniós bulgária unió lakosság párizs bűnöző
1	0.12082	magyar cigány gárda roma jobbik polgármester gárdista rendőrség falu rendezvény család képviselő szervezet demonstráció nemzeti elnök békés önkormányzat mozgalom olaszliszka helyszín szögi tüntetés kisebbségi fórum horváth rendőr rasszista kisebbség cigánybűnöző
2	0.0634	roma önkormányzat cigány kolompár elnök kisebbségi orbán ocö szervezet képviselő forint farkas támogatás pénz cigányság alapítvány lungo vizsgálat gazdasági drom vád horváth szerződés kállai választás összeg csalás ügyészség érdek anyaország
3	0.23144	rendőrség idős rendőrkapitányság bűncselekmény elkövető rendőr rablás gyanúsított pénz gyanú forint őrizet előzetes fiatal eljárás elkövetés sértett letartóztatás nyomozás ismeretlen megalapozott lopás tettes áldozat lakás ház érték bűntett rabló nyomozó
4	0.175	cigány roma magyar cigányság társadalom probléma magyarország társadalmi többség kultúra kisebbség többségi százalék csoport közösség politikai arány etnikai eredmény munka rasszista állam megoldás jog európai fiatal integráció szegény nép emberi
5	0.08818	ház lakás család önkormányzat uzsorás pénz ingatlan épület forint szomszéd cigány kölcsön telep polgármester tartozás környék szemét ózd segítség otthon ózdi adós munka áram hitel képviselő szerződés kamat állapot körülmény
6	0.08761	jobbik párt magyar fidesz kormány politikai képviselő magyarország politikus mszp választás elnök parlamenti orbán szocialista szdsz fideszes parlament mohácsi gyurcsány anyaország európai viktor hatalom liberális miniszterelnök bizottság politika vona viktória
7	0.29021	cigány magyar fehér roma bűnöző fajvédő származás hazugság cigók cigányfajvédő finn népszabadság náci gyilkos kivétel balliberális gádzsó nemzet gyilkosság bűnözés zsindex sajtó műsor médium hivatalos média normális gyakorlati elkövető rasszista
8	0.09475	falu ház polgármester család idős kert község tolvaj bácsi gazda lopás szomszéd rendőr környék munka néni önkormányzat porta rend lakosság betörés áram gyerek baj roma lakos polgárőr autó polgárőrség cigány
9	0.24146	rendőrség bűncselekmény rendőr törvény bűnöző érdek rendőri bűnözés közbiztonság hatóság intézkedés jog rend elkövető figyelem probléma védelem munka szolgálat lehetőség eredmény feladat eljárás hivatal szükséges jelentős komoly büntetés eszköz súlyos
10	0.06617	gyerek orvos család szülő beteg kislány kórház kiskorú fogyatékos kisfiú egészségügyi családi orvosi súlyos betegség vizsgálat lány állapot védőnő ellátás körülmény felnőtt nyírmihálydi gyógyszer csecsemő gyermekvédelmi intézet terhes mentő polgármester
11	0.28491	cigány magyar gyerek pénz fiatal lány baj család munka telefon fia ismerős feleség hölgy szegény idős környék vér arc ház isten fehér fekete srác nyak segítség ruha lassú férj nehéz
12	0.02173	fegyver víz lőfegyver amerikai magyar árvíz önvédelmi gát engedély zsilip pisztoly gázpisztoly lőszer folyó fegyvertartás védekezés fegyveres önvédelem rab usa amerika árvízi felsőzsolca hernád töltés birtoklás fosztogató éles part vízügyi
13	0.08519	kár tolvaj forint lopás ismeretlen kábel vasúti tettes érték elkövető rongálás kábellopás tűzoltó temető biztonsági őr rendőrség vezeték vonal fém vonat templom okozott fémtolvaj vas sín forintos máv szobor telephely

L.3. táblázat: Látens Dirichlet allokáció topikjainak első 30 kulcsszava - aszimmetrikus α paraméter, 2. lánc, 1-14. topik, $\beta = 0.017465$

Topik	α értéke	Kulcsszavak
14	0.09754	vádlott bíróság ítélet tárgyalás börtön elkövetett büntetés bíró előzetes jogerős sértett vád ügyészség vallomás emberölés tanú bűncselekmény ügyvéd börtönbüntetés szabadságvesztés bűntett ügyész eljárás per cigány letartóztatás felfüggesztett elsőrendű vádirat gyilkos
15	0.03071	cozma veszprém veszprémi sztojka román enyingi szórakozóhely sportoló rendőrség gyilkosság tanú verekedés patrióta kézilabdázó lokál vallomás siófoki gyanúsított játékos bár szerb banda enying szív diszkó gyilkos kés siófok kézilabdás sesum
16	0.0785	autó rendőr baleset kocsi utas sofőr cigány helyszín busz jármű rendőrség vonat gépkocsi sérülés parkoló személygépkocsi jogosítvány mentő villamos ittas közlekedő motoros közlekedési járőr rendőrautó ülő motor utazó sebesség megálló
17	0.21145	rendőr rendőrség cigány támadó sérülés kórház súlyos helyszín eljárás fiatal család támadás garázdaság testi kés életveszélyes elkövető sértés verekedés mentő fiatalember könnyű áldozat ház segítség erőszak őrizet feljelentés hétfő kedd
18	0.03332	miskolc miskolci pásztor lakatos vajda rendőrkapitány boon avas draskovics cigányvajda avasi lyukóvölgy miniszter kapitány lakótelep káli borsod fészekrakó sajtótájékoztató polgármester képviselő munk városrész bábonyibérc munka nyilatkozat rablás tény számozott városvezetés
19	0.015	film győzike junior dokumentumfilm damu közönség sziget művész h színész cigany adócsalás koncert énekes dal lengyel apeh mar rasszizmus belga kamera zenekar traveller milliós lista műsor színpad blog énekesnő madonna
20	0.10617	rendőrség gyilkosság ház áldozat támadás elkövető roma gyilkos helyszín lány család bűncselekmény nyomozás cigány gyanúsított tettes halál rendőr holttest nyomozó pécsi pest vizsgálat tatárszentgyörgy elkövetett emberölés rasszista nyom brutális kiskunlacháza
21	0.0178	cigány finnországi tücsök india csendőr havas kultúra vérbosszú hangya magyarország forrás bünöző elkövetők wc törzs cigányfaj jászapáti cigánybünözés hindu áron szülők csendőrség lumberg szar henrik bünözés cigányfajvédők században cionista nép
22	0.08931	segély munka szociális pénz támogatás forint önkormányzat polgármester program család gyerek rendszer lehetőség százalék összeg monok állami közmunka szepessy dolgozó családi pótlék állam törvény juttatás kormány munkahely havi munkanélküli monoki
23	0.01907	polgárőr gyöngyöspata roma gyöngyösi gyöngyöspatai egyesület gyöngyös tasz hajdúhadház field szervezet polgárőrség véderő patai juhász heol farkas falu richard lmp szjpe vöröskereszt közösség egyenruhás eszes hajdúhadházi pintér csoport etnikai konfliktus
24	0.02397	finn német svéd bevándorló osztrák németország svédország nyelv helsinki holland muzulmán fordítás török fehér ausztria mozlim külföldi brit származás zsidó fekete iszlám menekült nyugati dán nigger börtön finnországi adat kultúra
25	0.06811	iskola gyerek szülő tanár diák tanuló iskolai igazgató oktatási roma lány család osztálytárs fiatal óvoda hátrányos probléma mohácsi helyzetű erőszak iskolás gimnázium suli középiskola tanulás felnőtt önkormányzat kislány szakember tanítás
26	0.04045	magyar lány cigány kanada rendőrség család magyarország hatóság rabszolga áldozat prostituált prostitúció banda francia munka szarvasi strici magyarországi nyomozás szarvas futtató emberkereskedelem svájci hajléktalan rabszolgatartó csicska róbert svájc ophélie rendőr

L.4. táblázat: Látens Dirichlet allokáció topikjainak első 30 kulcsszava - aszimmetrikus α paraméter, 2. lánc, 15-27. topik, $\beta = 0.017465$

Topik	α értéke	Kulcsszavak
0	0.1269	ház család falu lakás önkormányzat polgármester szomszéd cigány környék épület roma gyerek ingatlan idős pénz kert otthon rend telep probléma szemét hivatal lakos porta segítség állapot munka telek község régi
1	0.07505	gyerek orvos szülő kislány család kórház beteg lány kiskorú kisfiú fogyatékos állapot családi orvosi egészségügyi súlyos védőnő gyógyszer gyermekvédelmi csecsemő nagymama ellátás terhes szolgálat intézet fiatal felnőtt korhatár körülmény mária
2	0.04331	magyar lány kanada cigány magyarország család rendőrség hatóság rabszolga osztrák áldozat prostitúció prostituált banda amerikai francia ausztria magyarországi budapest szarvasi strici emberkereskedelem svájci munka euró szarvas külföldi svájc rostás csicska
3	0.03312	miskolc miskolci pásztor lakatos vajda rendőrkapitány ózd ózdi avas boon draskovics avasi lyukóvölgy cigányvajda munka kapitány borsod káli fészekrakó miniszter képviselő városrész lakótelep munk sajtótájékoztató rablás bábonyibérc gyár telep városi
4	0.0864	segély munka szociális pénz támogatás polgármester önkormányzat program forint család gyerek rendszer lehetőség közmunka összeg állami állam családi pótlék százalék monok törvény munkanélküli juttatás szepessy munkahely monoki rendszeres havi szegény
5	0.11672	cigány roma magyar polgármester gárda jobbik falu rendőrség gárdista polgárőr képviselő család szervezet önkormányzat rendezvény gyöngyöspata elnök demonstráció rendőr lakos lakosság békés fórum ház kisebbségi polgárőrség mozgalom rendőri jobbikos egyesület
6	0.1963	cigány sérülés támadó rendőrség súlyos kórház rendőr fiatal áldozat kés család támadás elkövető testi életveszélyes helyszín garázdaság sértés verekedés fiatalember eljárás könnyű lány mentő állapot sértett arc segítség sérült szombat
7	0.04714	roma román cigány francia olasz bevándorló európai német tábor hatóság állampolgár szervezet kormány svéd illegális németország sarkozy külföldi bolgár euró közösség bűnöző brit bulgária elnök párizs eu róma római finn
8	0.06295	kár forint kábel tolvaj vasúti vonat lopás ismeretlen tettes temető kábellopás rongálás vonal templom fém vezeték elkövető máv sín rendőrség fémtolvaj réz szobor közlekedés bkv színesfém hév érték okozott hétfő
9	0.07351	párt jobbik fidesz kormány képviselő roma magyar politikus magyarország politikai mszp európai választás elnök parlamenti szlovák fideszes szocialista bizottság parlament orbán cseh szdsz nemzeti viktor miniszterelnök szervezet balog mohácsi szlovákia
10	0.20761	rendőrség gyanúsított rendőrkapitányság rendőr bűncselekmény eljárás előzetes gyanú őrizet letartóztatás elkövetés nyomozás megalapozott elkövető nyomozó bűntett sértett fiatal elkövetett bűnügyi bejelentés rablás követő adat fiatalkorú lopás bíróság anyaország helyszín büntetőeljárás
11	0.26912	cigány magyar gyerek munka pénz baj szegény fehér fiatal ismerős nehéz fia lány isten család feleség lassú erős komoly fekete telefon száj láb srác arc rend pillanat vér rendes igazi
12	0.06204	kolompár cigány önkormányzat roma elnök kisebbségi orbán ocö szervezet forint képviselő farkas támogatás pénz cigányság alapítvány lungo drom választás vád ügyészség gazdasági vizsgálat szerződés bizottság pályázat daróczi hivatal horváth jogosulatlan
13	0.10853	gyilkosság rendőrség ház áldozat roma támadás elkövető gyilkos lány cigány helyszín család tettes nyomozás bűncselekmény halál gyanúsított holttest rendőr pest nyomozó magyar vizsgálat nyom pécsi tatárszentgyörgy brutális meggyilkolt kiskunlacháza emberölés

L.5. táblázat: Látens Dirichlet allokáció topikjainak első 30 kulcsszava - aszimmetrikus α paraméter, 3. lánc, 1-14. topik, $\beta = 0.016645$

Topik	α értéke	Kulcsszavak
14	0.06909	iskola gyerek tanár szülő diák tanuló iskolai igazgató oktatási lány osztálytárs roma probléma hátrányos erőszak mohácsi óvoda család fiatal helyzetű önkormányzat iskolás tanulás középiskola gimnázium alapítvány suli szakember szegedi erzsébet
15	0.03055	cozma veszprém veszprémi sztojka román enyingi szórakozóhely sportoló gyilkosság kézilabdázó patrióta verekedés tanú rendőrség lokál játékos bár vallomás banda enying szerb siófoki szív gyanúsított film gyilkos diszkó kés kézilabdás ügyvéd
16	0.25746	cigány magyar fehér roma bűnöző finn fajvédő származás cigók népszabadság kivétel hazugság cigányfajvédő balliberális gyilkos gádzsó zsindex nemzet hivatalos gyilkosság médium bűnözés gyakorlati normális jellemző plusz finom náci itteni korcs
17	0.14422	magyar cigány rasszista zsidó magyarország politikai gárda bűnöző társadalom jog média rasszizmus cigányság nemzet náci hatalom haza liberális nemzeti nép politikus magyarság állam áldozat állampolgár bűn erőszak olaszliszka nyílt kisebbség
18	0.2743	rendőrség bűncselekmény törvény bűnöző érdek bűnözés hatóság rendőr probléma rend rendőri munka jog közbiztonság figyelem lehetőség elkövető intézkedés védelem csoport eszköz eredmény súlyos feladat adat szolgálat megoldás szükséges lakosság erő
19	0.11492	rendőr autó helyszín kocsi baleset cigány rendőrség sofőr járőr jármű utas busz rendőri tűzoltó mentő gépkocsi segítség intézkedés ercsi kolléga hivatalos erőszak rendőrautó kapitányság fegyver sérülés szolgálat intézkedő ittas lövés
20	0.16296	idős pénz forint rablás elkövető ismeretlen érték ház rabló lopás rendőrség lakás tolvaj fiatal áldozat néni tettes táska készpénz rendőr sértett mobiltelefon üzlet biztonsági betörő bűncselekmény betörés ékszer kés őr
21	0.0147	finn cigány finnországi nyírmihálydi roma betegség beteg havas vizsgálat vérbosszú film tbc helsinki telefon genetikai bünöző nyisztor vírus elkövetők cigánybünözés wc fertőzés mutáció hivatalos lumberg tünet sanomat motoros szülők hepatitisz
22	0.0528	uzsorás pénz forint család kölcsön tartozás összeg adós kamat uzsora hitel ház mali tiszabő polgármester ingatlan szerződés apeh csalás bank ft rendőrség zsarolás forintos havi lakás számla milliós áram ügyvéd
23	0.05265	tolvaj gazda víz lopás forint kár kert falu fatolvaj tanya határ falopás árvíz környék illegális gyümölcs termés mezőőr polgárőr érték ló munka termény helyszín zsilip fűrész folyó mezőgazdasági község gát
24	0.0183	fegyver áram bácsi lőfegyver kesznyéteni barna törvény állam önvédelem amerikai szepessy monok önvédelmi kert uborka jogos kesznyéten szoboszlai fegyvertartás engedély pisztoly wahorn börtön zsolt fekete gázpisztoly védelem áldozat támadó bódi
25	0.0183	fegyver áram bácsi lőfegyver kesznyéteni barna törvény állam önvédelem amerikai szepessy monok önvédelmi kert uborka jogos kesznyéten szoboszlai fegyvertartás engedély pisztoly wahorn börtön zsolt fekete gázpisztoly védelem áldozat támadó bódi
26	0.0929	vádlott bíróság ítélet tárgyalás börtön büntetés bíró elkövetett vallomás ügyészség vád jogerős tanú emberölés sértett cigány bűncselekmény börtönbüntetés szabadságvesztés ügyész ügyvéd előzetes eljárás bűntett per felfüggesztett elsőrendű védő vádirat fiatalkorú

L.6. táblázat: Látens Dirichlet allokáció topikjainak első 30 kulcsszava - aszimmetrikus α paraméter, 3. lánc, 15-27. topik, $\beta = 0.016645$

Topik	α értéke	Kulcsszavak
0	0.05682	tolvaj gazda kert lopás áram bácsi forint kár fatolvaj tanya gyümölcs határ barna kesznyéteni falopás tulajdon ló porta termés termény szomszéd falu környék zöldség ház tyúk mezőőr érték nyom illegális
1	0.05682	roma román cigány francia olasz európai cseh bevándorló kormány német hatóság szlovák szervezet tábor állampolgár szlovákia bolgár illegális elnök sarkozy eu közösség euró külföldi németország svéd csehország bulgária romák lakosság
2	0.05682	munka segély szociális pénz támogatás önkormányzat polgármester program család forint rendszer százalék lehetőség állami monok összeg közmunka törvény dolgozó szepessy családi munkahely gyerek állam szegény munkanélküli rendszeres pótlék juttatás kormány
3	0.05682	gyanúsított rendőrkapitányság rendőrség bűncselekmény rendőr gyanú eljárás őrizet előzetes letartóztatás elkövetés megalapozott elkövető nyomozás bűntett fiatal rablás nyomozó sértett elkövetett bűnügyi bejelentés követő bíróság fiatalkorú idős lopás anyaország adat ház
4	0.05682	rendőr rendőrség család helyszín cigány rendőri autó eljárás járőr fegyver hivatalos intézkedés támadó erőszak feljelentés szolgálat ház segítség kapitányság ercsi lövés rendőrkapitányság garázdaság kiérkező ügyészség csoport egyenruhás intézkedő támadás bejelentés
5	0.05682	cigány roma százalék arány fehér finn adat statisztika hivatalos kultúra többség felmérés kutatás finnországi magas eredmény cigányfajvédő cigányság alacsony fajvédő származású jellemző fekete forrás bűnözés fiatal börtön magyarország magyarországi megkérdezett
6	0.05682	orvos beteg kórház egészségügyi betegség mentő nyírmihálydi orvosi ellátás dolgozó vizsgálat háziorvos mentőautó kezelés mentőszolgálat genetikai ügyeletes rendelő állapot ellátó gyógyszer kórházi pál ügyelet hozzátartozó munkatárs tbc wc fertőzés pszichiátriai
7	0.05682	roma ház támadás rendőrség gyilkosság cigány elkövető rasszista áldozat helyszín család tatárszentgyörgy tűzoltó fegyver magyar etnikai tatárszentgyörgyi indíték bűncselekmény tettes fajvédő pátkai lövés polgárőr pátka nyomozás támadó pécsi bizonyíték nyomozó
8	0.05682	lány gyilkosság rendőrség áldozat gyilkos cigány elkövető halál holttest brutális nyomozás tettes család rendőr bűncselekmény gyanúsított pest emberölés magyar kiskunlacháza helyszín kislány nyomozó péntek meggyilkolt nyom idős kiskunlacházi tragédia ház
9	0.05682	cigány magyar finn bűnöző fehér származás népszabadság cigók nemzet rendőr kivétel korcs gyilkos börtön amerikai balliberális alábbi gyakorlati francia korcsok rajvédő rendőrség gyilkosság olasz bűnözés normálisabb kemény normális sajtó derék

L.7. táblázat: Látens Dirichlet allokáció topikjainak első 30 kulcsszava - szimmetrikus α paraméter, 1-10. topik, $\beta = 0.015855$

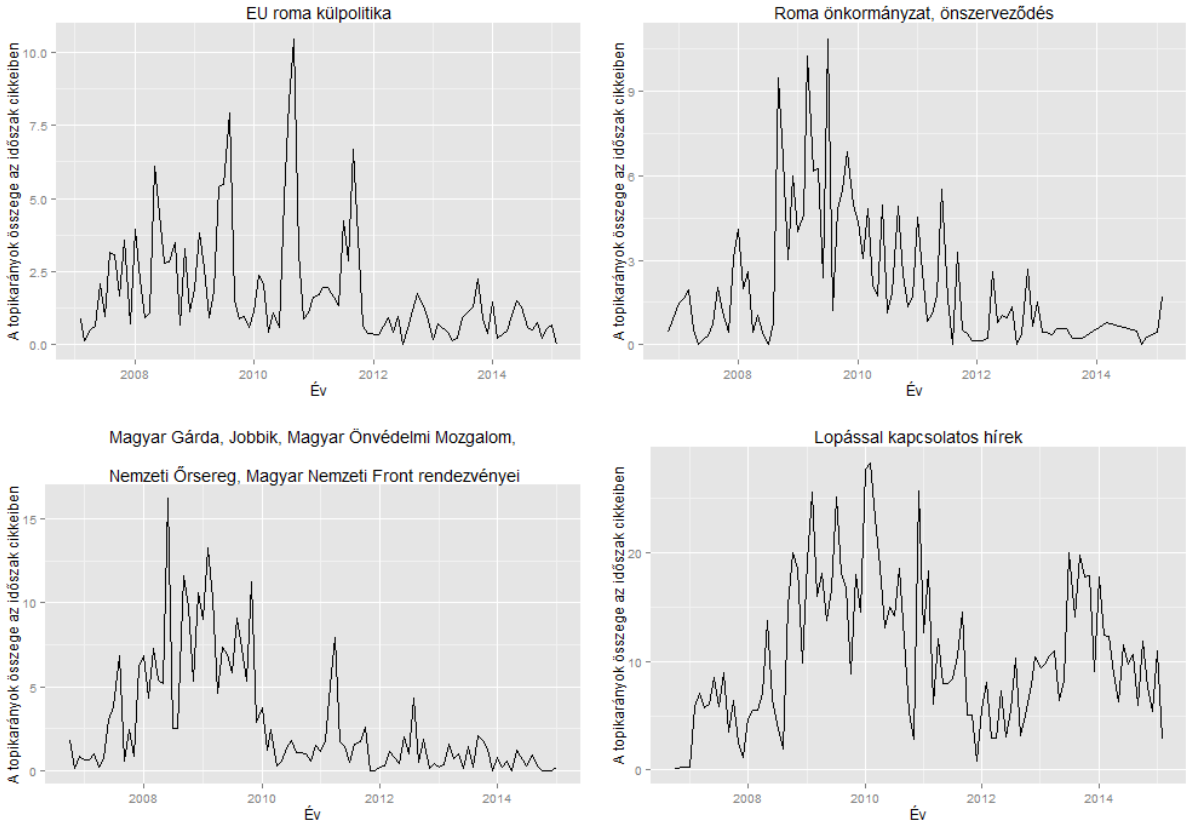
Topik	α értéke	Kulcsszavak
10	0.05682	cigány sérülés kórház súlyos támadó rendőrség fiatal áldozat testi elkövető kés sértés életveszélyes támadás helyszín fiatalember sértett mentő eljárás verekedés könnyű garázdaság rendőr állapot sérült szóváltás szombat lány kísérlet arc
11	0.05682	autó baleset cigány kocsi utas sofőr busz rendőr jármű helyszín vonat gépkocsi sérülés villamos rendőrség mentő jogosítvány ittas megálló könnyű ülő közlekedő közlekedési fiatalember súlyos személygépkocsi segítség parkoló utazó motoros
12	0.05682	iskola gyerek tanár szülő diák tanuló iskolai igazgató oktatási roma hátrányos lány osztálytárs probléma fiatal erőszak helyzetű mohácsi cigány család gimnázium iskolás tanulás középiskola fegyelmi erzsébet suli alapítvány tanítás nehéz
13	0.05682	kár forint tolvaj lopás kábel vasúti ismeretlen tettes temető kábellopás rendőrség rongálás templom vonal fém víz vonat vezeték elkövető fémtolvaj sín máv szobor érték réz hév színesfém okozott vas közlekedés
14	0.05682	műsor film tv daróczi kultúra televízió amerikai zene adás kulturális szöveg rádió riporter győzike havas kamera telefon újságíró tévé szereplő interjú műsorvezető junior előadás budapest közönség kereskedelmi művész ismert v
15	0.05682	cozma veszprém sztojka veszprémi román szórakozóhely enyingi sportoló rendőrség gyilkosság tanú verekedés patrióta kézilabdázó lokál vallomás bár játékos cigány szerb banda diszkó gyanúsított siófoki szív enying kés gyilkos siófok ügyvéd
16	0.05682	cigány magyar roma fehér fajvédő rasszista zsidó náci bűnöző hazugság média cigányfajvédő pénz zsindex tény mohácsi hiba gádzsó magyarország gyerek paraszt munka baj normális nép büdös jog szegény komoly hülye
17	0.05682	magyar cigány magyarország kanada lány hatóság osztrák prostituált ausztria banda áldozat magyarországi brit prostitúció külföldi rendőrség család euró svájci strici bűnöző állampolgár holland haza svájc roma anglia nemzetközi nemzet hollandia
18	0.05682	vádlott bíróság ítélet tárgyalás börtön elkövetett büntetés bíró ügyészség jogerős vád vallomás sértett tanú bűncselekmény emberölés börtönbüntetés bűntett szabadságvesztés előzetes ügyész eljárás ügyvéd per cigány felfüggesztett elsőrendű vádirat védő ítélőtábla
19	0.05682	rendőrség bűncselekmény törvény bűnöző elkövető érdek rendőr bűnözés rendőri jog munka közbiztonság rendőrkapitány hatóság védelem miskolci lehetőség adat figyelem pásztor intézkedés probléma miskolc súlyos vizsgálat büntetés feladat miniszter csoport eszköz

L.8. táblázat: Látens Dirichlet allokáció topikjainak első 30 kulcsszava - szimmetrikus α paraméter, 11-20. topik, $\beta = 0.015855$

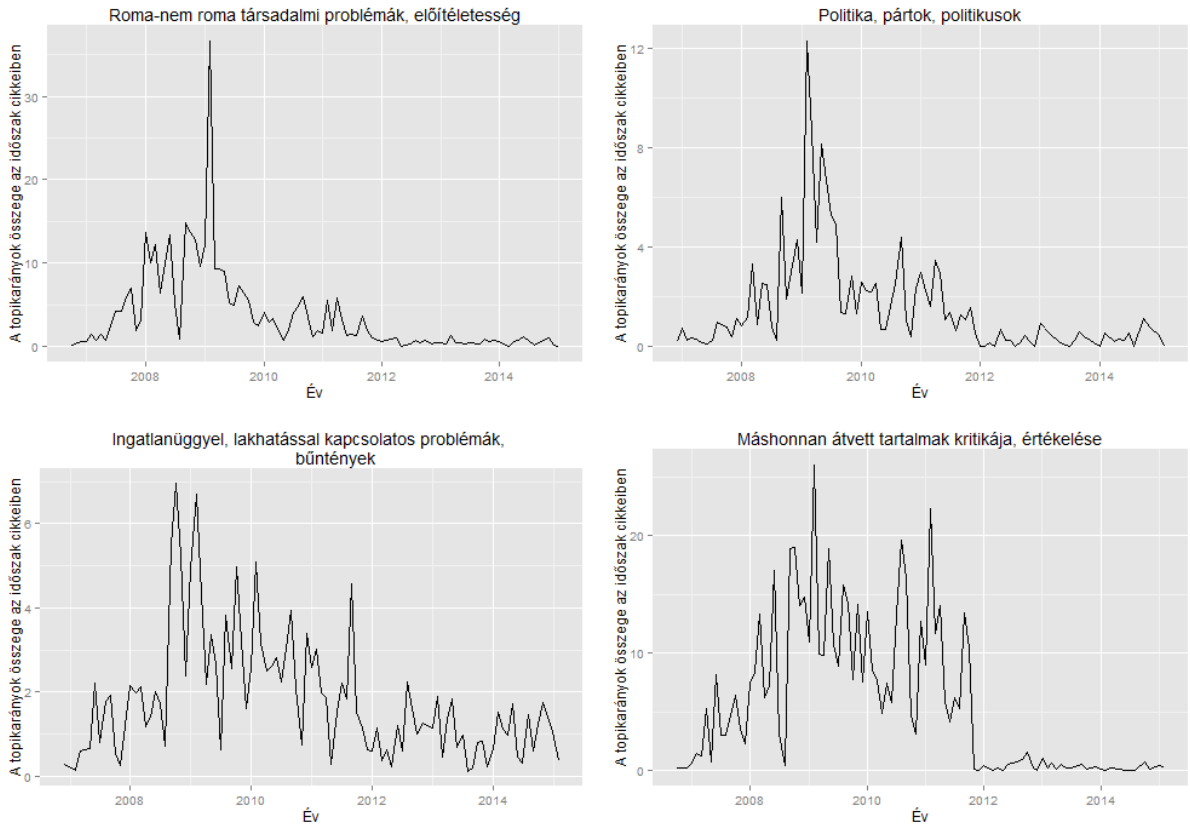
Topik	α értéke	Kulcsszavak
20	0.05682	gyerek szülő család kislány lány kiskorú kisfiú családi cigány felnőtt fogyatékos óvoda fiatal fiatalkorú gyermekvédelmi korhatár védőnő csecsemő intézet büntethető pótlék iskola gyermekkorú terhes szolgálat nevelő rendszeres körülmény felelősség erőszak
21	0.05682	uzsorás pénz forint család kölcsön tartozás rendőrség összeg ház adós hitel kamat uzsora ingatlan vállalkozó mali apeh munka áldozat rabszolga szerződés tiszabő ft csalás szarvasi bank számla feleség polgármester segítség
22	0.05682	polgármester falu roma cigány önkormányzat család polgárőr lakosság lakos község magyar ház gyöngyöspata probléma polgárőrség rendőrség közbiztonság idős képviselő rend rendőr hivatal kisebbségi polgármesteri bűnözés lopás gond munka bűnöző környék
23	0.05682	cigány ház család gyerek pénz munka falu feleség fia baj szomszéd idős fiatal férj lány autó magyar rendőr ismerős környék telefon néni hölgy bácsi nehéz kocsma testvér szegény rend öreg
24	0.05682	cigány roma magyar társadalom cigányság probléma társadalmi magyarország politikai többség közösség kultúra többségi kisebbség csoport megoldás etnikai állam emberi nép nemzet integráció európai rasszista bűnöző politika jog eredmény munka érdek
25	0.05682	ház lakás önkormányzat család miskolc cigány épület ingatlan miskolci telep környék képviselő szemét lakótelep ózd szomszéd állapot városrész avas áram ózdi telek otthon víz rend probléma segítség lépcsőház pénz körülmény
26	0.05682	idős forint pénz rablás elkövető ismeretlen ház rendőrség lopás érték tolvaj rabló lakás tettes rendőr táska áldozat fiatal sértett készpénz néni biztonsági kár bűncselekmény üzlet kés őr rendőrkapitányság betörő ékszer
27	0.05682	cigány roma önkormányzat kolompár elnök kisebbségi orbán ocö szervezet forint képviselő farkas támogatás pénz cigányság magyar lungo drom alapítvány gazdasági hivatal szerződés kállai vád feljelentés pályázat bogdán vizsgálat érdek összeg
28	0.05682	jobbik párt fidesz képviselő kormány magyar politikus politikai magyarország mszp elnök választás parlamenti szocialista bizottság szdsz fideszes parlament orbán európai országgyűlési viktor anyaország mohácsi választási miniszterelnök nemzeti gyurcsány szavazat államtitkár
29	0.05682	magyar gárda cigány jobbik gárdista roma rendőrség rendezvény szervezet demonstráció nemzeti mozgalom elnök békés tüntetés olaszliszka képviselő szögi helyszín magyarország vajda cigányság polgármester horváth tamás politikai vona jobbikos felvonulás félelem

L.9. táblázat: Látens Dirichlet allokáció topikjainak első 30 kulcsszava - szimmetrikus α paraméter, 21-30. topik, $\beta = 0.015855$

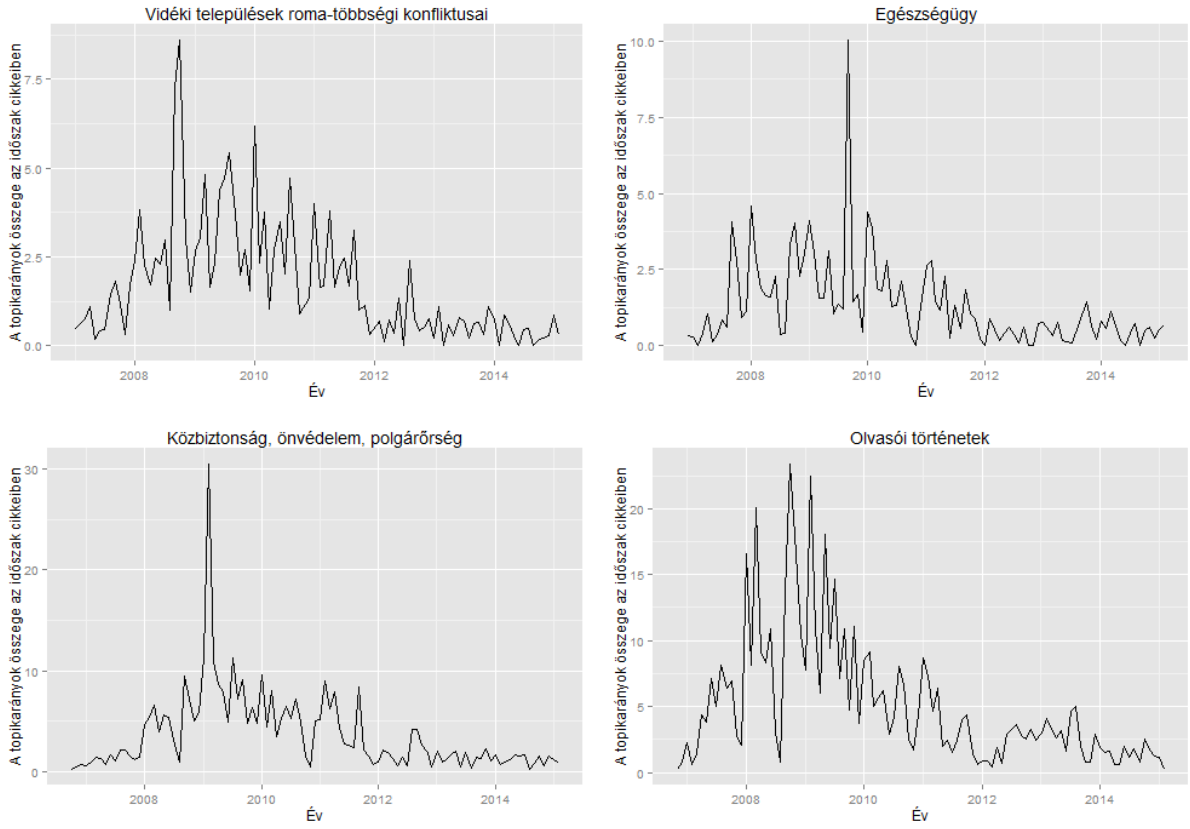
M. A TOPIKOK IDŐBELI VÁLTOZÁSA



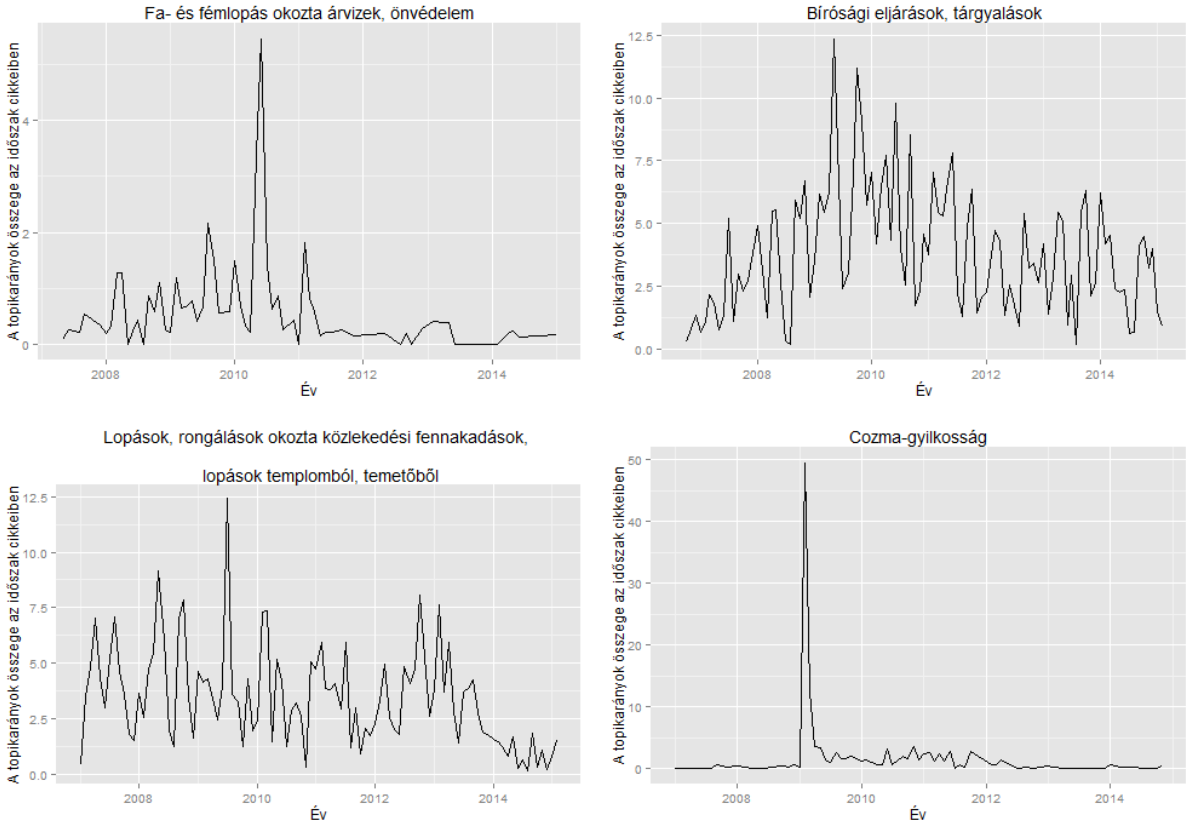
M.1. ábra. Topikok időbeli változása - 1-4. topik



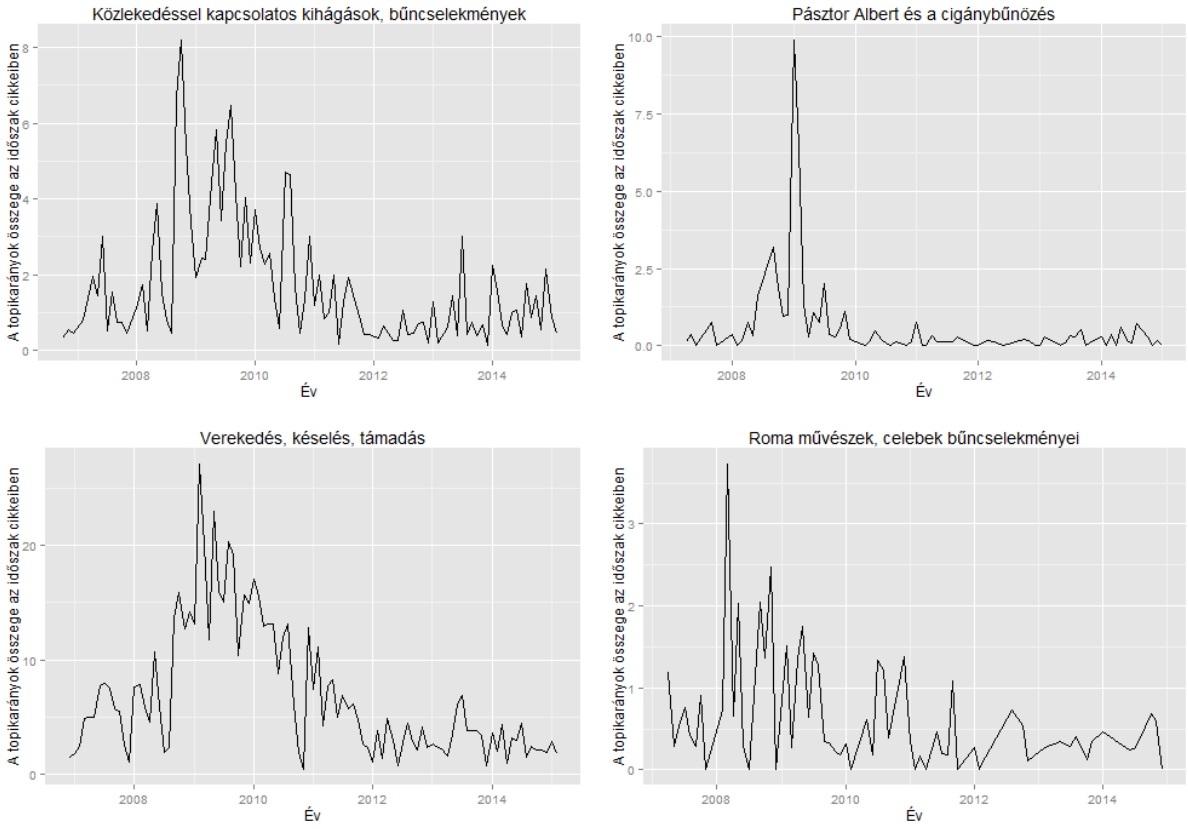
M.2. ábra. Topikok időbeli változása - 5-8. topik



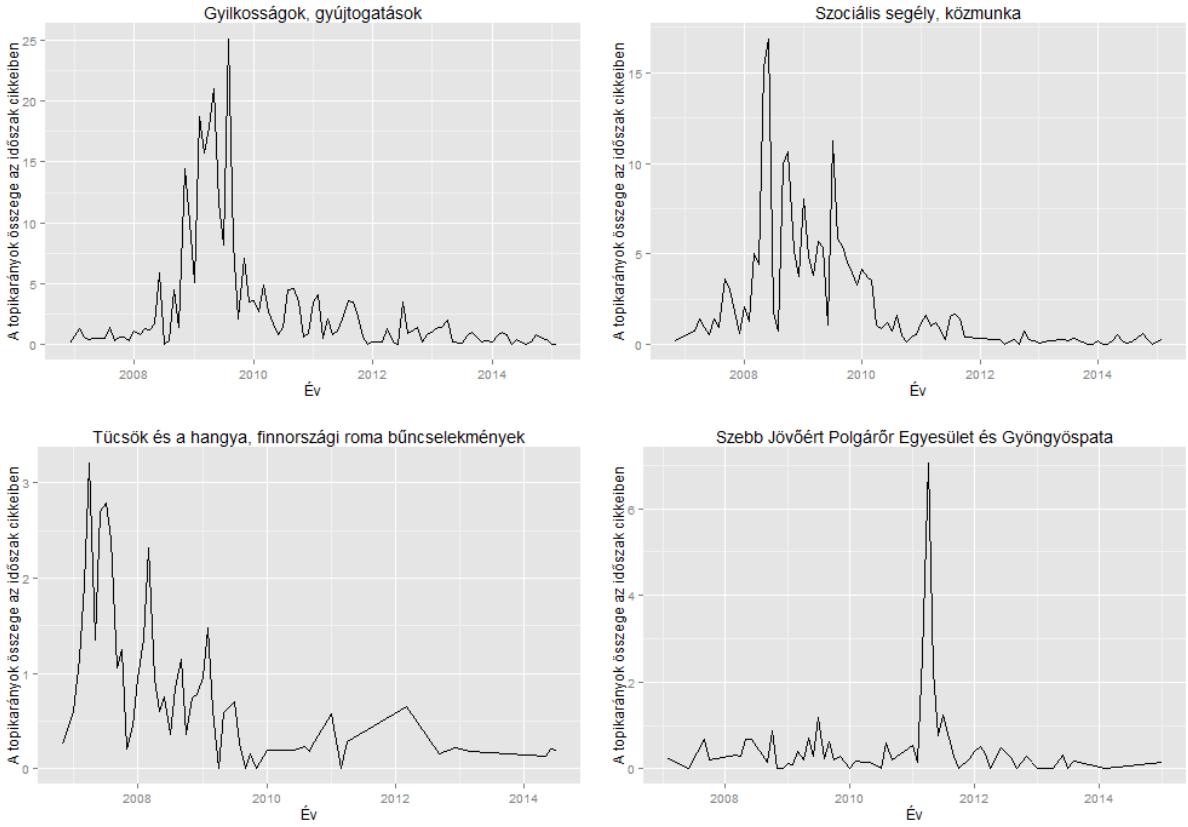
M.3. ábra. Topikok időbeli változása - 9-12. topik



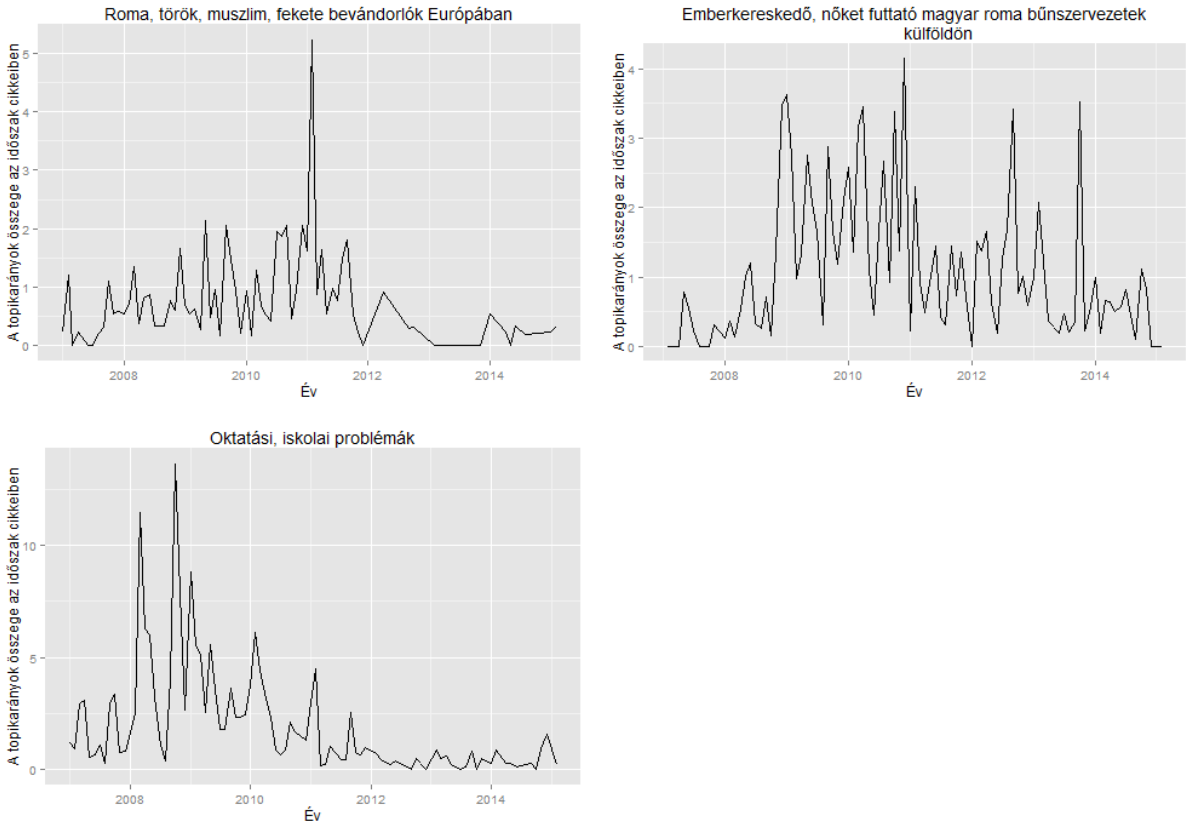
M.4. ábra. Topikok időbeli változása - 13-16. topik



M.5. ábra. Topikok időbeli változása - 17-20. topik



M.6. ábra. Topikok időbeli változása - 21-24. topik



M.7. ábra. Topikok időbeli változása - 25-27. topik