Statisztikai Szemle

Közzététel: 2022. augusztus 24.

A tanulmány címe:
A topikmodellezés lehetőségei és korlátai egy törvénykorpusz példáján

Szerzők:

Gelányi Péter,

a Társadalomtudományi Kutatóközpont Politikatudományi Intézetének kutatási asszisztense
E-mail: gelanyi.peter@tk.hu

Sebők Miklós,

a Társadalomtudományi Kutatóközpont Politikatudományi Intézetének intézetigazgatója
E-mail: sebok.miklos@tk.hu

Ring Orsolya,

a Társadalomtudományi Kutatóközpont Politikatudományi Intézetének kutatója
E-mail: ring.orsolya@tk.hu

DOI: https://doi.org/10.20311/stat2022.8.hu0783

Gelányi Péter – Sebők Miklós – Ring Orsolya

A topikmodellezés lehetőségei és korlátai egy törvénykorpusz példáján

The opportunities and constraints of topic modelling – the case of a corpus of laws

Gelányi Péter, a Társadalomtudományi Kutatóközpont Politikatudományi Intézetének kutatási asszisztense
E-mail: gelanyi.peter@tk.hu
Sebők Miklós, a Társadalomtudományi Kutatóközpont Politikatudományi Intézetének intézetigazgatója
E-mail: sebok.miklos@tk.hu
Ring Orsolya, a Társadalomtudományi Kutatóközpont Politikatudományi Intézetének kutatója
E-mail: ring.orsolya@tk.hu

A topikmodellezés a felügyelet nélküli tanulás egy fajtája, amelynek segítségével egy korpusz dokumentumait kategorizálhatjuk szemantikailag értelmezhető témakörök alapján kialakított csoportokba. A módszernek számos potenciális felhasználási lehetősége van a társadalomtudományok területén. Jelen tanulmány a topikmodellezés előnyeit és buktatóit tekinti át, illetve mutatja be egy kutatás példáján keresztül, amelynek során az 1990 és 2018 között elfogadott magyar törvényekből álló korpuszokból alakítottunk ki topikmodelleket. Célunk az LDA felhasználási lehetőségeinek felmérése volt. A kutatás során ciklusonként kialakított alkorpuszokon futtattunk topikmodelleket, majd az ugyanezen a korpuszon végzett kézi kódolás eredményeivel összehasonlítva értékeltük ki azokat. Eredményeink alapján az LDA más módszerekhez képest jelentősen kisebb mértékű erőforrás-befektetés mellett is alkalmas szemantikailag értelmezhető és koherens kategóriák kialakítására, amelyek a további vizsgálatok szempontjából relevánsak lehetnek. Ugyanakkor nem tanácsoljuk az algoritmus validáció nélküli használatát. A topikmodellezés elsődleges alkalmazási lehetőségét a vizsgált dokumentumok előzetes feldolgozásában, strukturálásában látjuk.

Tárgyszó: topikmodellezés, LDA, felügyelet nélküli tanulás

Topic modelling is a form of unsupervised learning, it is used to categorize the documents of a corpus into groups based on semantically interpretable topics. This method has a number of potential applications in the context of social science research. This study provides an overview of the opportunities and constraints of topic modelling within a social science context, through a concrete research example, that applies topic modelling to a corpus consisting of hungarian laws from 1990 to 2018. Our aim is to provide an evalueation of the potential research usage of LDA. We evaluated our models based on comparisons with hand coding research done on the same corpus. Our results show, the categories generated by our models were semanticly interpretable, and were relevant to potential further study of the corpus, we stress the importance of validation. We see the primary use of topic modelling in the preliminary processing and structuring of data.

Keywords: topic modelling, LDA, unsupervised learning

A társadalomtudományokban az 1980-as és az 1990-es évek fordulóján jelentek meg az első olyan statisztikai eljárások, amelyek képesek voltak nagyobb szövegkorpuszokat komplex módon elemezni, de a módszer használatának felfutása csak a 2000-es években indult meg *(Miner et al., 2012; Grimmer-Stewart, 2013).* A világszintű áttörés mellett a magyar társadalomtudományban a kvantitatív szövegelemzés jobbára még gyerekcipőben jár. A témával foglalkozó kutatóknak így egyaránt feladatuk a felhasznált módszerek alkalmazási lehetőségeinek és az új kutatási eredményeknek a bemutatása. Ennek ellenére vannak példák ezeknek a módszereknek a társadalomtudományi kutatásokban való hazai alkalmazására más módszerekkel párhuzamosan, vegyes módszertanú kutatásokban, vagy akár önmagukban is *(Galántai et al., 2018; Hajósi, 2020; Kerezsi et al., 2019; Kollár, 2020; Nagy–Molnár, 2017; Sebők–Kacsuk, 2020).* Így az általunk alkalmazott csoportosítási feladatok kapcsán is találhatunk példát szövegbányászati eljárásokra *(Katona et al., 2021; Holecz–Szűcs, 2016; Balogh et al., 2017).* Ezenkívül több magyar tudományos munka foglalkozik a szövegbányászati módszerek társadalomtudományi alkalmazásának lehetőségeivel *(Boda–Sebők, 2018; Bolonyai–Sebők, 2020; Németh et al., 2020; Nyitrai, 2021; Sebők, 2016; Sebők et al., 2018; Sebők et al., 2021; Tikk, 2017).* Ehhez az egyre gyarapodó irodalomhoz kívánunk mi is hozzájárulni egy specifikus szövegbányászati módszer, a topikmodellezés áttekintésével és gyakorlati alkalmazásának bemutatásával.

Tanulmányunkban az egyik legelterjedtebb szövegbányászati feladatot, a csoportosítást vesszük górcső alá. Ezen belül figyelmünket a felügyelet nélküli tanulás *(unsupervised learning)* egy alkalmazási lehetőségére, a topikmodellezésre fordítjuk. A felügyelet nélküli tanulás során az alkalmazott algoritmus nem kap részletes iránymutatást a szövegállomány strukturálásának szabályaira vonatkozóan, az elemzési rendszer előzetes kutatói kialakítása így jobbára az eredményként várt csoportok számának meghatározására szorítkozik. Ez alacsony előzetes erőforrás-befektetést tesz lehetővé *(Quinn et al., 2010),* ugyanakkor az elemzés eredményeit is nehezebb hozzáilleszteni a szóba jövő kutatási kérdések egy részéhez.

Elemzésünk során az 1990 és 2018 között elfogadott magyar törvények korpuszát használtuk fel a topikmodellezés módszertani lehetőségeinek és problémáinak feltérképezésére. A kutatás során arra voltunk kíváncsiak, hogy a korpuszunkon kialakított topikmodellek mennyire alkalmasak a törvényhozás trendjeinek vizsgálatára, összehasonlítva a kézi kódolással. A tanulmányban először áttekintjük a topikmodellezés alapjait és témánk szempontjából releváns politikatudo-

mányi alkalmazásait. Ezt követően röviden vázoljuk a kutatási tervet, majd bemutatjuk a teszteléshez felhasznált adatbázist. A következő lépésben részletesen elemezzük a topikmodellezés munkafolyamatát a szövegek importálásától az elemzések értékeléséig. Végül értékeljük az alkalmazott módszerek eredményeit és kitekintünk a további kutatási lehetőségekre.

Eredményeink alapján a kialakított csoportosítások koherensek és szemantikailag értelmezhetőek, ugyanakkor ez önmagában nem jelenti azt, hogy az adott kategóriák a kutatási kérdés szempontjából releváns módon csoportosítják a korpusz dokumentumait. A modellek eredményeinek a kézi kódolással való összevetése alapján látható, hogy az egyes topikok kialakításának módja nem kellően konzisztens ahhoz, hogy a hozzájuk tartozó dokumentumok számának alakulása alapján megbízhatóan következtessünk a topik témájával foglalkozó dokumentumok számának időbeli alakulására.

1. A topikmodell

A felügyelet nélküli tanulási módszerek közé tartozó topikmodellezés során az alkalmazott algoritmus a dokumentumok hasonlóságait felhasználva végez csoportosítást, azaz hoz létre különböző kategóriákat. Segítségével egymástól szemantikailag megkülönböztethető témákat azonosíthatunk a vizsgált korpuszban. Alkalmazásakor nem előre definiált csoportokba rendezünk dokumentumokat, hanem a korpusz belső hasonlóságai alapján hozunk létre kategóriákat, amelyeket később azonosítunk (címkézünk). A topikmodell a szövegbányászati eljárásokon belül az ismeretlen kategóriákba történő, teljes mértékben automatizált vegyes tagságú klasszifikáció elvégzésére használható *(Grimmer–Stewart, 2013, 268. old.)*.

A topikmodellezés legjelentősebb előnyei, hogy viszonylag kevés erőforrást igényel, és a dokumentumokra vonatkozó előzetes ismeretek hiányában is alkalmazható. Az egyes topikmodellek kutatási kérdések széles körének vizsgálatát teszik lehetővé. Lehetőséget teremtenek például szövegek ideológiai tartalmuk alapján való klasszifikációjára *(Lin et al., 2008),* egyes témakörök politikai polarizációs hatásának a felmérésére *(Balasubramanyan et al., 2012),* illetve az egymásnak ellentmondó vélemények modellezésére *(Fang et al., 2012)*. A topikmodellezés alkalmas a politikai figyelem tárgyának kutatására *(Quinn et al., 2010)*, vagy egy konkrét közpolitikai vita keretezéseinek vizsgálatára *(Levy–Freanklin, 2014)*. Ugyancsak lehetőséget nyújt egy konkrét téma sajtóban való különféle megjelenési módjainak csoportosítására, illetve a további kutatás szempontjából

felhasználható anyagok kiválogatására egy nagyobb terjedelmű korpuszból *(Jacobi et al., 2016),* vagy akár a véleményvezérek adott témakörökben elfoglalt álláspontjának megállapítására *(Chen et al., 2010).*

A topikmodellezés során az egyik leggyakrabban alkalmazott modell az LDA *(Latent Dirichlet Allocation),* amely egy generatív probabilisztikus modell *(Blei et al., 2003).* Ennek az algoritmusnak többféle potenciális felhasználási módja létezik, ezek egyike a szövegek csoportosítása. Az LDA eredeti verziójára gyakran utalnak mint alap-LDA-re *(Nikolenko et al., 2017),* megkülönböztetve számos kiegészítőjétől és módosított verziójától, amelyekről a 3. fejezetben ejtünk majd szót.

Az LDA egyik meghatározó eleme, amely a kulcs a modell működésének megértéséhez, a működésében lévő feltételezés arra vonatkozóan, hogy miképpen jöttek létre az egyes dokumentumok. A modell feltételezi, hogy egy dokumentum a következőképpen keletkezik: először kiválasztjuk a rendelkezésünkre álló topikok egy adott megoszlását, majd a topikok véletlenszerűen kiválogatott szavaival megtöltjük a dokumentumot ennek a megoszlásnak megfelelően. Az LDA pedig ezen logika mentén visszafelé dolgozva megállapítja, hogy az ilyen módon létrehozott dokumentumok milyen topikokból állnak, és a topikokban milyen kifejezések vannak jelen. Természetesen a dokumentumok, amelyekkel dolgozunk, nem a fentiekben leírt folyamattal jöttek létre, viszont ez a logika tükrözi a dokumentumok szerkezetét. Általában egy korpusz dokumentumai többféle témát fognak eltérő mértékben tartalmazni, amelyek összefüggésben állnak egyes kifejezésekkel, s ezek az összefüggések lehetnek erősebbek vagy gyengébbek. Tehát, bár a szerzőik emberek, az egyes dokumentumokban rejlő mintázatok megfelelnek az LDA algoritmus feltételezésének *(Blei–Lafferty, 2009, 73. old.).*

Az egyik módja, hogy konceptualizáljuk a topikmodellek működését, hogy elképzeljük, ahogy egyesével végigmegyünk különböző színű kiemelőkkel egy korpusz dokumentumain, és azonos színnel jelöljük ki azokat a szavakat, amelyekről úgy gondoljuk, hogy egy közös témához tartoznak. Ennek a munkának az eredményei egyrészt az azonos színnel megjelölt szavak halmazai, vagyis a szóklaszterek, topikok, amelyek szemantikailag értelmezhető témakörök lesznek. Továbbá nemcsak azt kapjuk meg, hogy milyen topikokat tartalmaz a korpuszunk, hanem az egyes dokumentumokat is annotálhatjuk és rendszerezhetjük aszerint, hogy milyen topikok milyen arányban jelennek meg bennük.

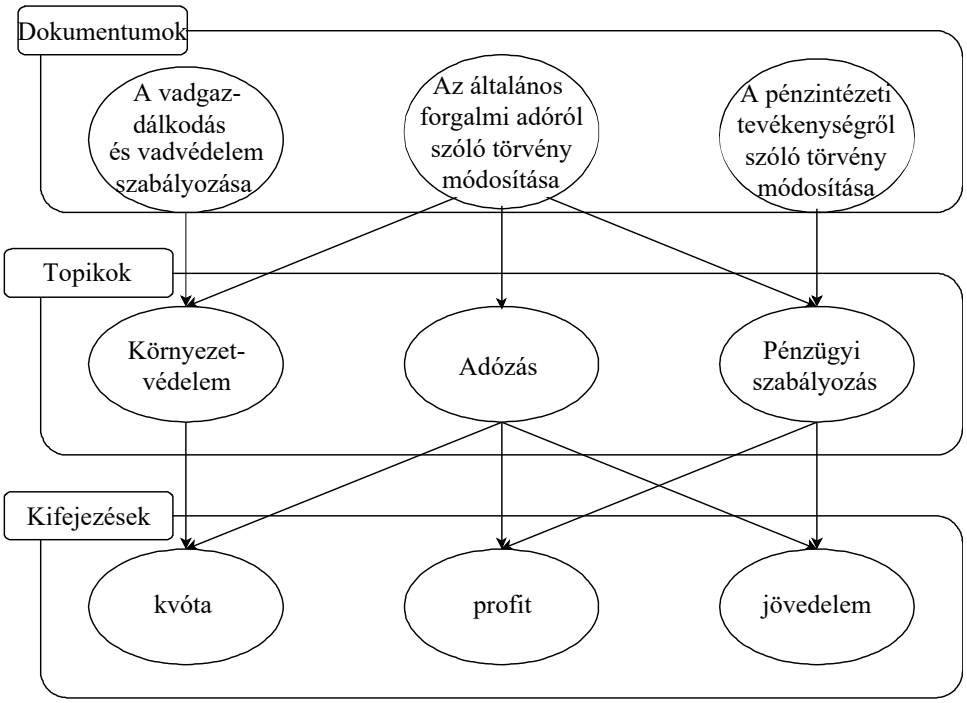
Természetesen mi azért lennénk képesek egy ilyen feladat elvégzésére, mert tisztában vagyunk az egyes szavak jelentésével. Ezzel szemben egy algoritmus meg tudja különböztetni egymástól a szavakat, viszont nem rendelkezik a szükséges információval, hogy értelmezze is azokat, így nem is képes ennek segítségével osztályozni őket, hanem a szavak dokumentumokon belüli megjelenéseinek

számát használja fel. Ez pedig azzal jár, hogy a topikmodellezés érzékeny a szavak között megjelenő bármilyen kapcsolatra, összefüggésre. Mivel két kifejezés számos okból jelenhet meg rendszeresen együtt egy korpuszon belül, a modellünk által létrehozott topikok reflektálhatnak témákat, amelyekkel az egyes dokumentumok foglalkoznak, egyes témák eltérő keretezéseit, beszéd-, illetve írásstílusokat, valamint nyelvtani szerkezeteket is. Ebből következik, hogy egy adott kutatás során nincs lehetőségünk közvetlenül befolyásolni, hogy a korpuszunkra vonatkozó milyen információkat tükröznek majd az egyes topikok *(Jacobi et al., 2016).* Éppen ezért fontosak az olyan előkészítő lépések, mint a korpusz tisztítása és a topikok számának kiválasztása, amelyekről az 5. fejezetben ejtünk szót. Ezeknek a módszereknek az alkalmazása nélkül ugyanis a dokumentumok csoportosítása nem szemantikailag értelmezhető topikok alapján történik.

1. ábra

A dokumentumok, topikok és kifejezések kapcsolata az LDA-modelleken belül

The relation of documents, topics and words within LDA models



Az LDA megértéséhez kulcsfontosságú megérteni a három szint közötti kapcsolatot, amelyet az 1. ábra mutat be egy példán keresztül. A szavak szintjén található minden kifejezés, amely a korpusz dokumentumaiban felbukkan. A topikok pedig ezekre a szavakra vonatkozó probabilisztikus eloszlások. A modell legmagasabb szintjén helyezkednek el a dokumentumok, amelyek a topikokra vonatkozó probabilisztikus eloszlások, ahogy az egyes topikok is a dokumentumok szavaira vonatkozó probabilisztikus eloszlások. Mint ahogy korábban említettük, az LDA a vegyes tagságú klasszifikációk csoportjába tartozik, és ahogyan azt az 1. ábrán látható példa is szemlélteti, az egyes LDA-modellekben egy dokumentum több topikból állhat, illetve egy topik több szóból, és ahogyan egy topik több dokumentumban is jelen lehet, úgy egy szó is több topik része lehet.

2. Az alap-LDA alternatívái: CTM és STM

Ahogyan azt korábban is említettük, az alap-LDA csupán egyike a kutatók számára elérhető számos topikmodellezési módszernek. Ezek az alternatívák mind igyekeznek és képesek is valami újat kínálni az LDA-hez képest, ugyanakkor fontos szem előtt tartani, hogy ezek az extra elemek előzetes feltételezésekkel is járnak a korpuszunkra vonatkozóan. Minden esetben lényeges, hogy megértsük az általunk alkalmazott modell működését, és megbizonyosodjunk arról, hogy az összhangban áll a kutatási kérdésünkkel és a korpuszunkra vonatkozó előzetes információinkkal, feltételezéseinkkel. Az alábbiakban az LDA olyan alternatíváit mutatjuk be, melyek a társadalomtudományi kutatások számára hasznos kiegészítésekkel rendelkeznek.

Az LDA kiegészítései általában két kategóriába sorolhatók: újfajta struktúra szerint rendszerezik a topikmodelleket, vagy valamilyen új információt vonnak be a modell működésébe *(Nikolenko et al., 2017).* Ez az új struktúra lehet például egy hierarchikus modell, amely gráfként reprezentálja a dokumentumok hálózatát *(Chang–Blei, 2010).* A kiegészítő információ pedig lehet például a topikok korpuszban való megjelenésének ideje *(Wang–McCallum, 2006),* vagy a dokumentumok szerzőire vonatkozó metaadat *(Rosen-Zvi et al., 2010, 2012).*

A számos elérhető alternatíva közül kettővel szeretnénk részletesebben is foglalkozni. Az első a korrelált topikmodell (CTM) *(Blei–Lafferty, 2006),* amely képes arra, hogy pótolja az LDA azon hiányosságát, hogy az nem képes az egyes topikok közötti korrelációk modellezésére, mivel dirichlet disztribúciót használ a topikok eloszlásának modellezésére. A CTM logisztikai normál disztribúciót alkalmaz ezen feladat elvégzésére, így elérhetővé teszi az egyes topikok korrelá-

cióját *(Aitchison, 1982)*. A szerzők eredeti példája ennek az információnak a hasznosítására a szakirodalmon belüli keresés, ahol a korrelációkra vonatkozó új információk segítségével a kutatók azonosíthatnak olyan kutatási területeket, amelyek szoros összefüggésben állnak a saját témájukkal, így lehetővé válik a szakirodalom alaposabb áttekintése *(Blei–Lafferty, 2007; Günther–Domahidi, 2017; Yau et al., 2014)*. Ez a modell más kutatások esetén is jól használható, az egyes topikok megjelenéseinek korrelációja számos társadalomtudományi kutatás számára releváns lehet a korpusz megismeréséhez, áttekintéséhez. A CTM tehát hasznos kiegészítést jelent az LDA-topikmodellezéshez képest, ugyanakkor fontos szem előtt tartani, hogy a CTM az LDA-hez képest jelentősen nagyobb számítási kapacitást igényel.

A második modell, amelyet részletesebben is bemutatunk, a strukturált topikmodell (STM) *(Roberts et al., 2014)*. Az STM lehetővé teszi a kutatók számára, hogy a topikmodelljüket kiegészítsék dokumentumokra vonatkozó metaadatokkal. Ezek a metaadatok kovariánsokként jelennek meg a generatív modellen belül, így befolyásolhatják a topikok tartalmát, valamint azt, hogy az egyes dokumentumok milyen mértékben állnak kapcsolatban a topikkal. Ez a lehetőség kétféleképpen is hasznosítható egy kutató számára. Egyrészt új kutatási kérdések vizsgálatának a lehetőségét biztosítja, mivel lehetővé teszi, hogy összekapcsoljunk egy elméleti relevanciával rendelkező metaadatot azzal, hogy egy adott topik milyen módon és milyen gyakorisággal jelenik meg a vizsgált korpuszon belül. Másrészt pedig javíthatja a topikok kialakításának pontosságát, hiszen, ha a modellbe emelt metaadatok a kutatási kérdés szempontjából valóban elméleti relevanciával bírnak, akkor az így kialakított modell topikjai, illetve azok eloszlása várhatóan a kutatásnak megfelelő logikát jobban tükröző eredményeket ad.

Az STM alkalmazása során megjelölhetünk topikelterjedtségi kovariánsokat, valamint egy, a topiktartalomra vonatkozó kovariánst. Tehát az egyes kovariánsoknak nem feltétlenül kell hatniuk a modell korábbiakban említett két elemére. Valamint nem szükséges metaadatokat kovariánsként beemelni a modell mindkét részére vonatkozóan. Ha egyáltalán nem adunk meg kovariánsokat, akkor ezek hiányában a modell a korábban már tárgyalt CTM megfelelőjévé válik *(Roberts et al., 2019)*.

A módszer kidolgozója a modellt, illetve annak hasznosítását a kérdőívek nyitott (vagyis előre definiált válaszlehetőségekkel nem rendelkező) kérdéseinek gyors és költséghatékony feldolgozásának példája mentén mutatja be. Ugyanakkor az STM minden olyan esetben jól alkalmazható eszköz, amikor nagy mennyiségű digitális vagy digitalizálható dokumentumot dolgozunk fel, amihez releváns metaadatok állnak a rendelkezésünkre, tehát számos kutatási kérdés esetében hasznosítható *(Kleinberg et al., 2020; Hu et al., 2019; Barberá et al., 2017; Farrell, 2016)*.

3. A kutatási terv és az adatbázis

A tanulmányunk alapjául szolgáló kutatás esetén választásunk az LDA-topikmodellre esett, mivel ez a legelterjedtebb, továbbá jelentősen egyszerűbb az alkalmazása, ezért az elérhető alternatívák közül ez a modell lehet a legvonzóbb a mélyebb szövegbányászati ismereteket nélkülöző kutatók számára. Elemzésünket a Magyar Országgyűlés által 1990 és 2018 között elfogadott törvények szövegén végeztük, a korpuszunk összesen 4163 egyedi dokumentumból állt. A törvények szövegét a magyar Comparative Agendas Project (CAP-) adatbázisból szereztük be *(Boda–Sebők, 2019).* A topikmodellezés során a parlamenti ciklusok szerint 7 darab alkorpuszt alakítottunk ki, majd azonos szövegtisztítási módszerek alkalmazását követően, minden esetben K = 20 klaszterszámmal alakítottuk ki a modelleket. A klaszterek számát több bevett mérés alkalmazásával állapítottuk meg, amelyeket a következő fejezetben részletezünk. Az egyes topikokat a 100 legnagyobb bétaértékű kifejezéseik alapján, illetve a hozzájuk legerősebben kötődő törvény szövege alapján címkéztük minden ciklus esetén. Az így kialakított, csoportosított törvények segítségével azt vizsgáltuk, hogy miképpen alakult a rendszerváltást követő időszakban a törvényhozás prioritása.

A CAP-adatbázisból származó törvények, amelyeket alkalmaztunk, már korábban csoportosítva lettek kézi kódolás segítségével, szakpolitikai felosztás szerint. Ezt a lehetőséget kihasználva több szempont mentén is összehasonlítottuk a gépi kódolás által kialakított csoportosítást az eredetivel azokban az esetekben, ahol a két kategóriarendszer látszólag átfedésben van egymással. Vizsgáltuk, hogy mennyiben tükrözi az LDA által kialakított besorolásunk a CAP szakpolitikai felosztását, valamint azt is, hogy mindkét esetben azonosíthatók-e ugyanazok a trendek.

4. A topikmodellezés lépéseinek leírása

A vizsgált korpusz egészén egységes módon hajtottunk végre szövegtisztítási eljárásokat. A teljes korpuszt kisbetűsítettük, a központozást, a számokat és a szimbólumokat eltávolítottuk, mint ahogy az R-quanteda csomagjába beépített magyar és angol tiltólistás szavakat is. A szavakat szótövesítettük. Ez eltávolítja az egyes szóvégeket, így az azonos kifejezések különböző módokon ragozott és toldalékolt formáit az esetek többségében egy azonos formára vezeti vissza. Végezetül egy specifikusan ehhez a korpuszhoz általunk összeállított tiltólistán található kifejezéseket is eltávolítottuk a korpuszból, amely a szokásos tiltólistás

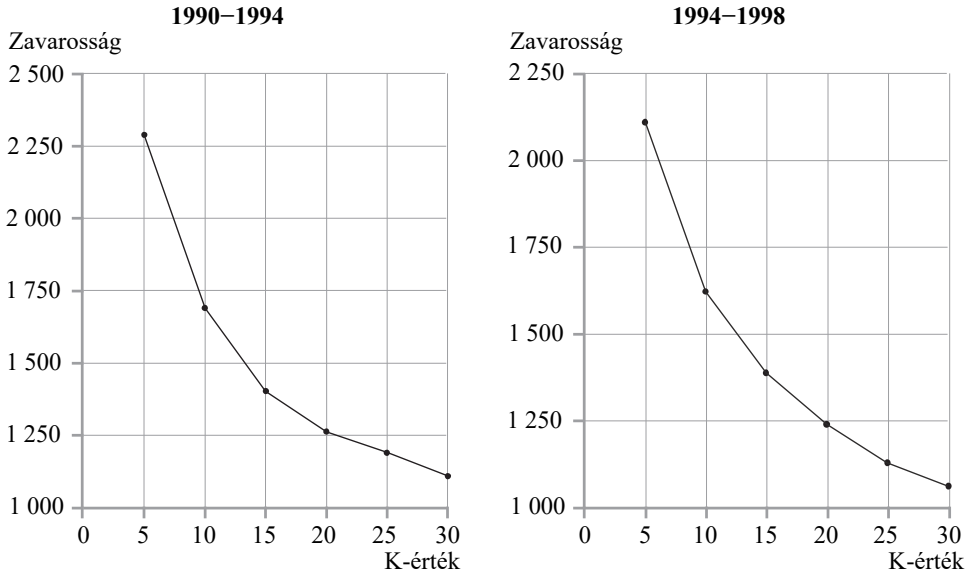
szavakon felül, mint amilyenek például a kötőszavak, olyan kifejezéseket is tartalmazott, amelyek más kontextusban hasznosak lennének, viszont a törvények differencializálásához nem járulnak hozzá (pl. bekezdés, hely, alpont).

A szövegtisztítási eljárásokat még az egységes korpuszon végeztük el, viszont ezt követően 7 alkorpuszra osztottuk a vizsgált időszak 7 ciklusa szerint, és ezeken futtattuk az LDA-modellünket is, mivel így tudtuk vizsgálni a törvényhozás prioritásainak ciklusok közötti változását. A 7 alkorpuszon külön-külön hajtottuk végre a megfelelő K-érték megállapításához szükséges eljárásokat. Ehhez öt különböző módszert alkalmaztunk, és ezek eredményeit hasonlítottuk össze. Az egyes méréseket minden esetben 5, 10, 15, 20, 25 és 30 K-értékekkel végeztük el, ezen mérések eredményei láthatók a 2., 3., 4. és 5. ábrán. Az első és legelterjedtebb eljárás a *perplexity,* vagyis a zavarosság kiszámítása, ami az információs elmélet standard mérőszáma, és megmutatja, hogy egy statisztikai modell milyen pontosan ír le bizonyos adatokat. Segítségével a topikmodellek esetében a topikok által reprezentált szóeloszlásokat tudjuk összehasonlítani a tényleges szóeloszlásokkal. Minél alacsonyabb a zavarosság értéke, annál pontosabban írják le a szóeloszlások a tényleges korpuszt *(Zhao et al., 2015).*

2. ábra

A zavarosságszámítás eredményei ciklusonként

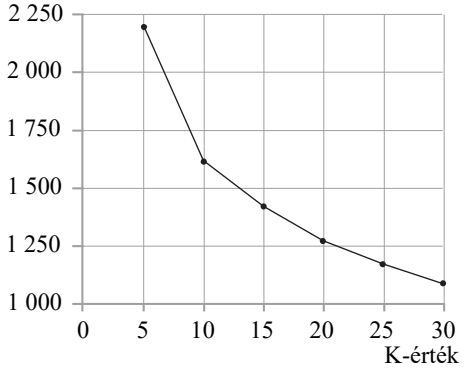
The perplexity results of each subcorpus



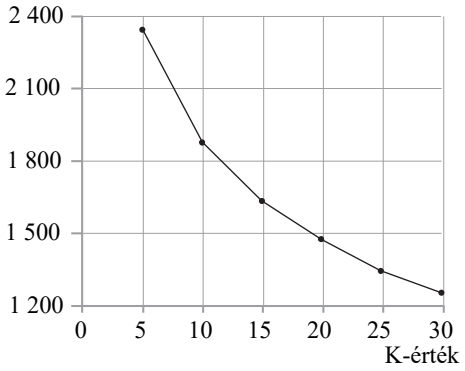
(Az ábra folytatása a következő oldalon)

(folytatás)

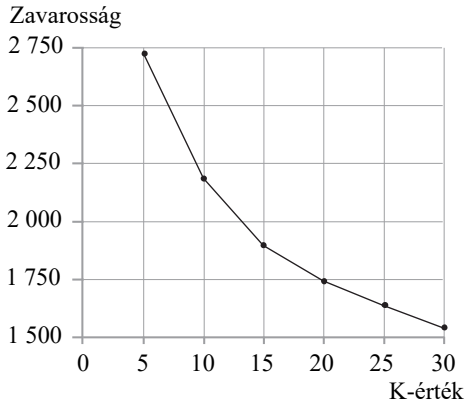
1998–2002



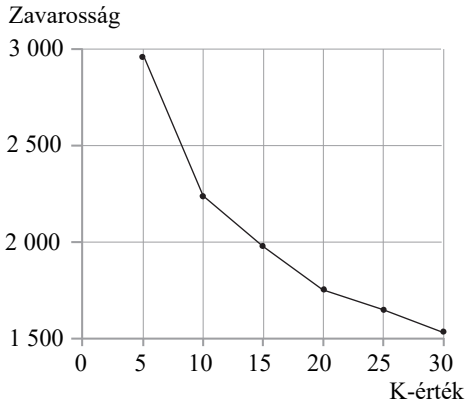
2002–2006



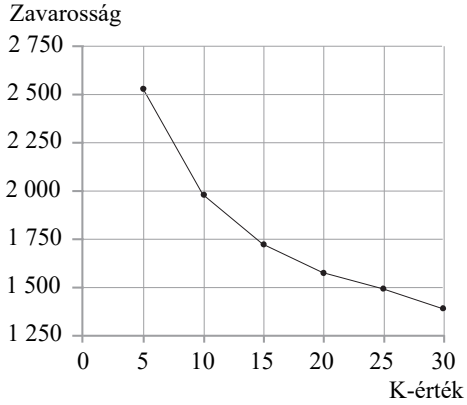
2006–2010



2010–2014



2014–2018



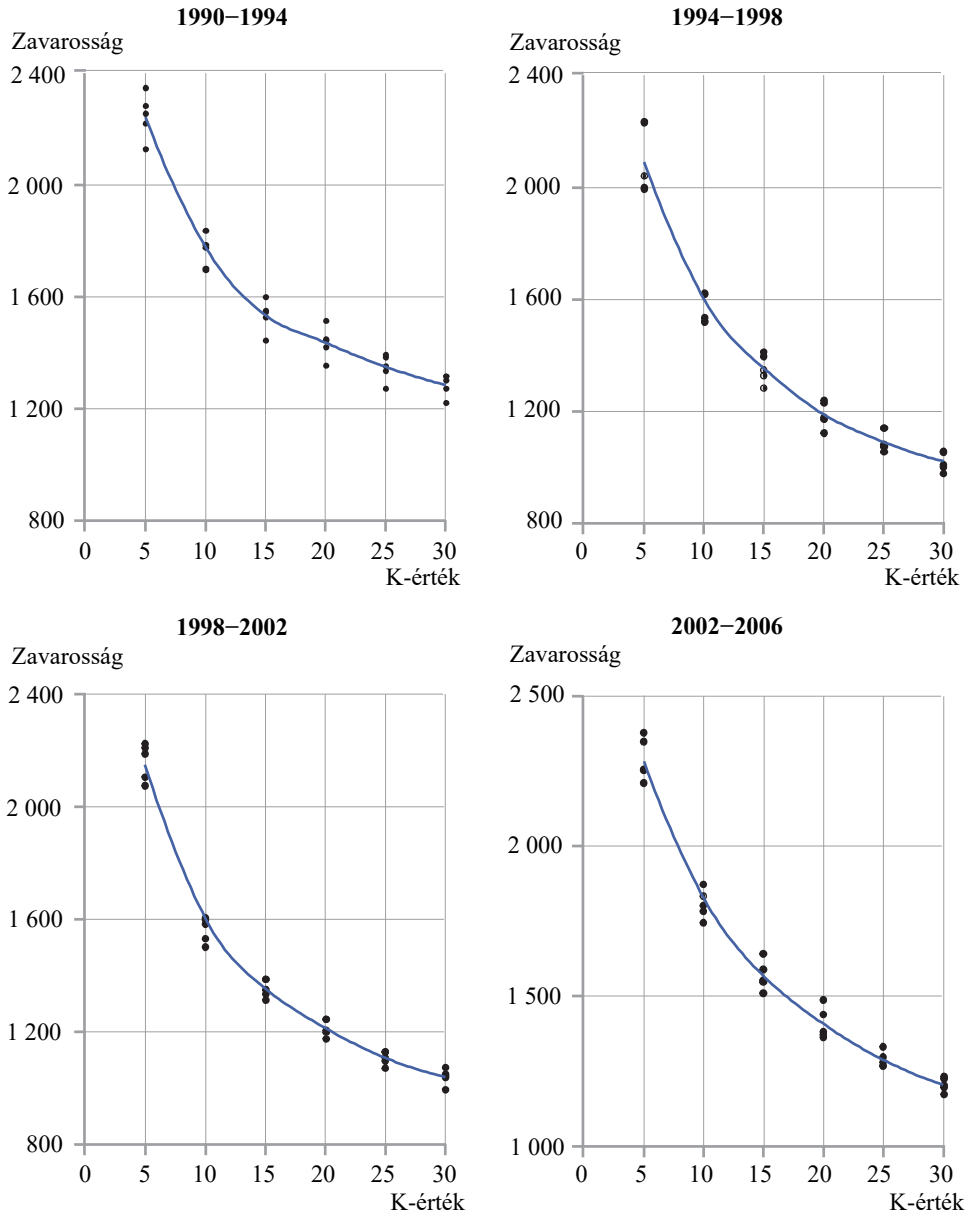
A zavarosságszámítást elvégeztük keresztvalidálással is, minden ciklus korpuszát 5 véletlenszerűen kialakított részre osztottuk, ezekből pedig tanító- és teszthalmazok 5-féle kombinációját hoztuk létre. Minden esetben 4/5 tanítóhalmazon alakítottunk ki egy LDA-t, majd megmértük, hogy az így létrehozott LDA mennyire találja zavarosnak a kimaradt teszthalmazt. A keresztvalidálás nemcsak abban a tekintetben előnyös az egyszerű zavarosságszámításhoz képest, hogy ötször annyi méréssel vizsgálja a zavarosság és a K-érték összefüggését, hanem abban a tekintetben is megbízhatóbb eredményt ad, hogy a zavarosságot mindig egy olyan mintán méri, amely nem szerepelt a tanítóhalmazban, így nem kapunk túlságosan optimista eredményeket *(Browne, 2000, 108–110. old.)*.

A zavarosság vizsgálatán felül még további négy módszert alkalmazunk. Egyrészt a Juan Cao és szerzőtársai által javasolt eljárást, amely az egyes topikok átlagos koszinusztávolsága alapján próbálja az ideális K-értéket megtalálni *(Cao et al., 2009)*. Emellett alkalmazzuk még Romain Deveaud, Eric SanJuan és Patrice Bellot módszerét, amely az összes topikpár közötti információ divergenciamértéke alapján azonosítja a megfelelő K-értéket *(Deveaud et al., 2014)*. Továbbá használjuk Thomas L. Griffiths és Mark Steyvers eljárását, amely különböző K-értékekkel végzett log-likelihood számítások alapján hasonlítja össze a potenciális K-értékeket *(Griffiths–Steyvers, 2004)*. Valamint alkalmazzuk Arun és szerzőtársainak mérését, amely mátrixok segítségével veti össze a topikok szeparációját különböző K-értékek mellett *(Arun et al., 2010)*. Az ezekkel a módszerekkel végzett méréseket szintén megjelenítettük vizuálisan is. Az egyes mérések nem jelölnek ki egyhangúan egy ideális K-értéket, viszont az eredmények átfedése alapján minden időszakon egységesen $K = 20$ értékkel futtattunk LDA-topikmodellt. Az egyes kvantitatív mérések hasznos eszközök, ugyanakkor fontos észben tartani, hogy eredményeik nem abszolútak, s ahogy láthattuk a példánkban is, nem feltétlenül produkálnak egyhangú eredményeket. A megfelelő K-érték megállapításának kérdésében tehát mindig a kutatóé a végső döntés.

3. ábra

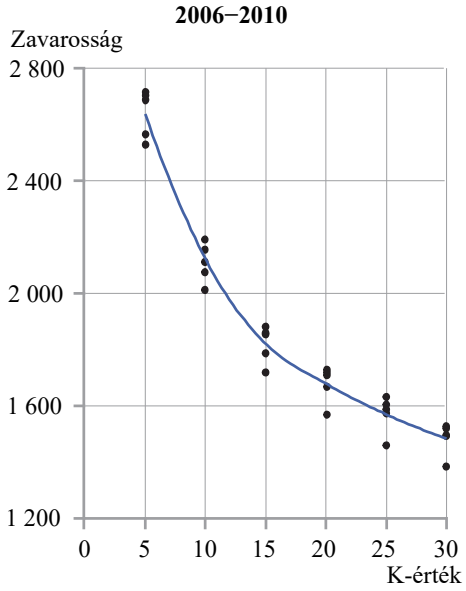
A keresztvalidációval mért zavarosság eredményei ciklusonként

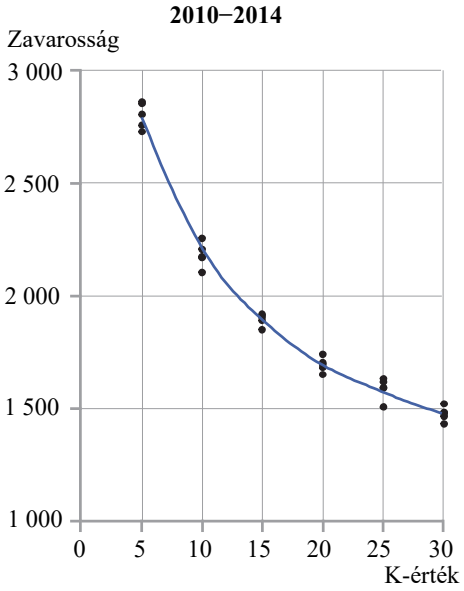
Perplexity results with cross validation on each subcorpus

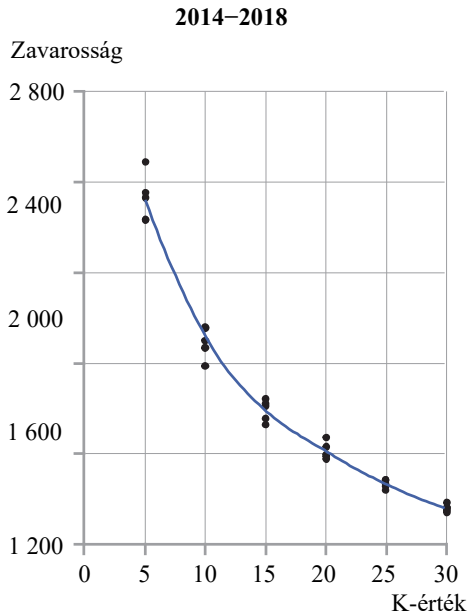


(Az ábra folytatása a következő oldalon)

(folytatás)


2006–2010
Zavarosság
2 800
2 400
2 000
1 600
1 200
0
5
10
15
20
25
30
K-érték

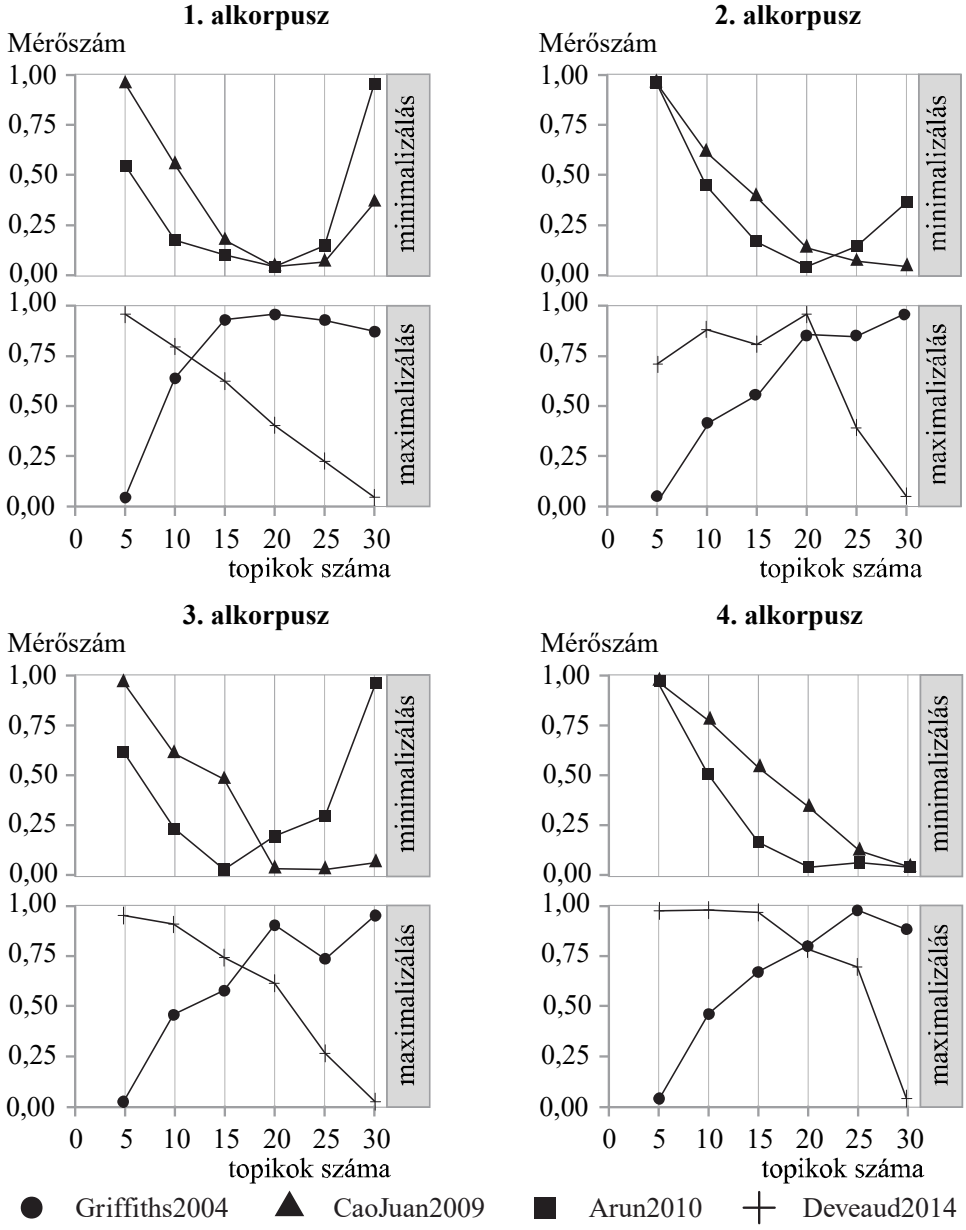

2010–2014
Zavarosság
3 000
2 500
2 000
1 500
1 000
0
5
10
15
20
25
30
K-érték


2014–2018
Zavarosság
2 800
2 400
2 000
1 600
1 200
0
5
10
15
20
25
30
K-érték

4. ábra

A további módszerekkel végzett mérések eredményei az első négy alkorpuszon

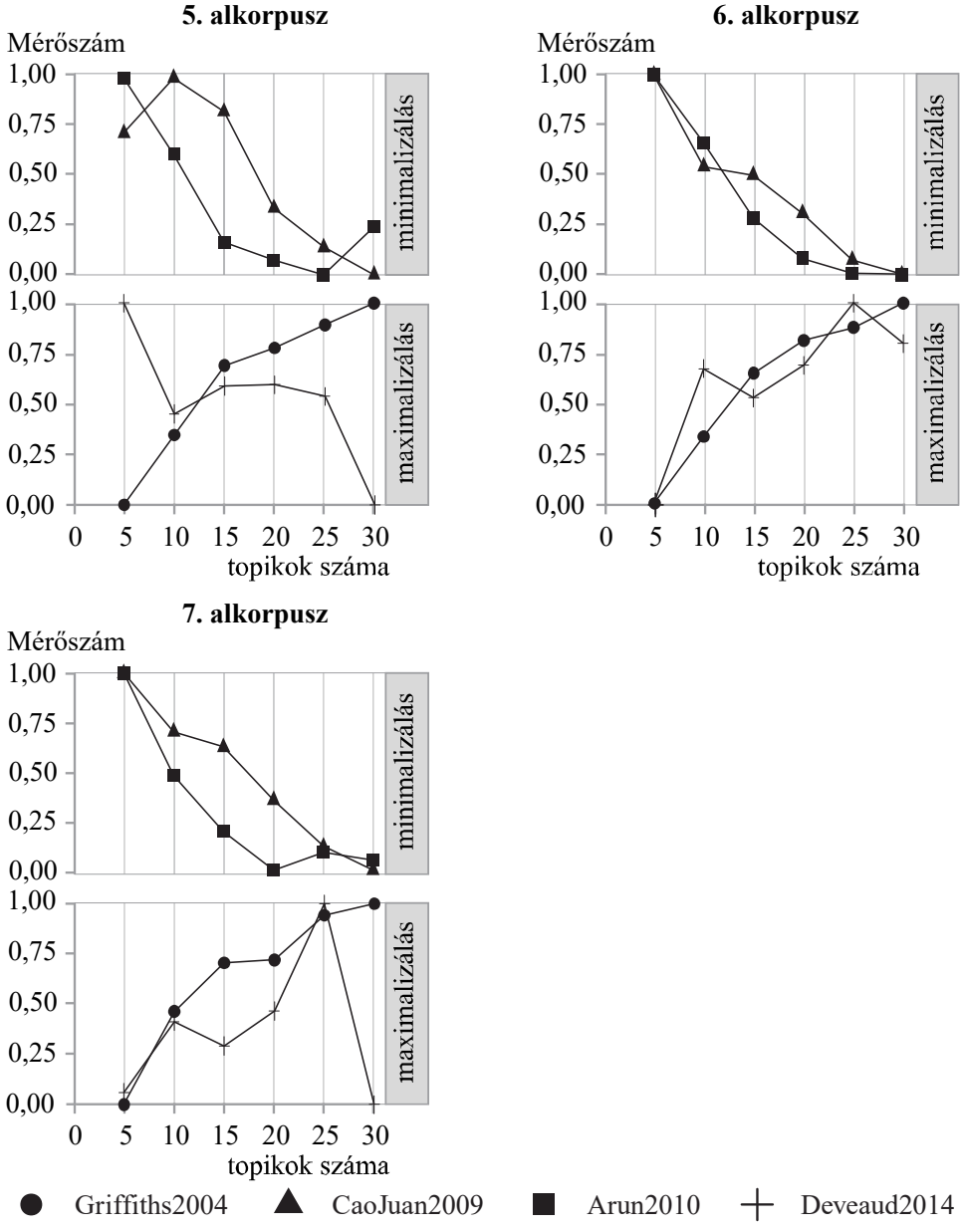
The results of further methods for determining the value of K on the first four subcorpus



5. ábra

A további módszerekkel végzett mérések eredményei az utolsó három alkorpuszon

The results of further methods for determining the value of K on the last three subcorpus



1. táblázat

Az első ciklus topikjainak címkéi

The labels of the topics in the first electoral cycle

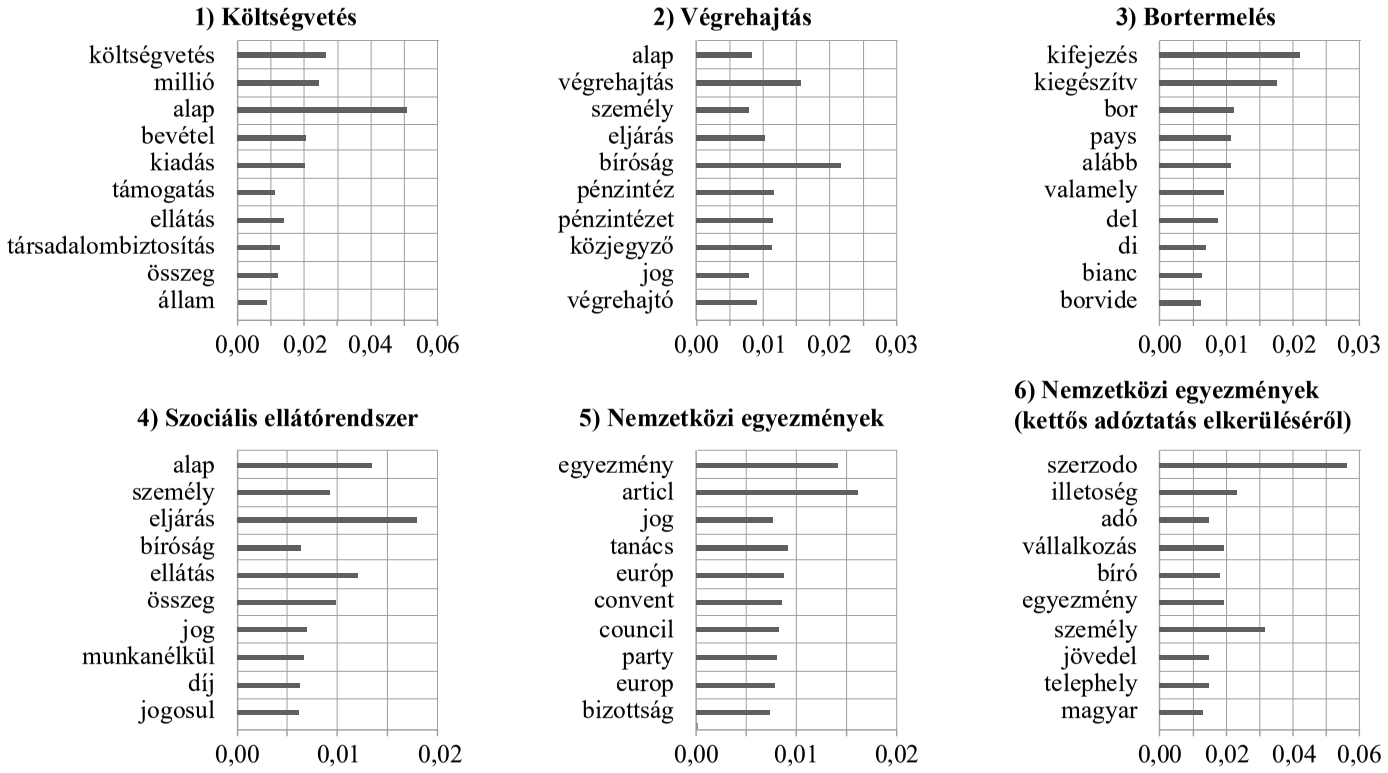
Topik száma	Topik címkéje	Idesorolt dokumentumok darabszáma
1.	Költségvetés	47
2.	Végrehajtás	32
3.	Bortermelés	1
4.	Szociális ellátórendszer	32
5.	Nemzetközi egyezmények	9
6.	Nemzetközi egyezmények (kettős adóztatás elkerüléséről)	11
7.	Privatizáció	20
8.	Önkormányzatok támogatása	16
9.	Adózás	22
10.	Pénzügyi szervezetek	24
11.	Állami koncessziók	19
12.	Költségvetés	23
13.	Budapest közigazgatási területekre felosztása	1
14.	Munkajog	26
15.	Adózás	27
16.	Külkereskedelem, vámok	3
17.	Szabadkereskedelem	1
18.	Választások	51
19.	Önkormányzatok, önszabályozó szervezetek	33
20.	Közalkalmazottak, közérdekű információk	32
	Összesen	430

Az LDA-modell futtatása után az egyes kategóriák címkézését alapvetően a kutatónak kell elvégeznie, bár a címkézésre számos különböző automatizált módszer is rendelkezésre áll *(Basave et al., 2014; Bhatia et al., 2016; Kou et al., 2015; Lau et al., 2011).* Kutatásunk során az időigényesebb, viszont megbízhatóbb kézi címkézési eljárást alkalmaztuk, mivel ez egyben azt is lehetővé tette, hogy alaposan áttekintsük az LDA segítségével kialakított modelljeinket. Az első ciklus modelljének címkézése az 1. táblázatban látható, a 6. ábrán pedig a topikokhoz tartozó legmeghatározóbb kifejezések vannak. A címkézést az egyes topikok legnagyobb bétaértékű kifejezései alapján, illetve a topikokkal leginkább összefüggésben álló dokumentumok szerint készítettük el. Mivel a hét cikluson egységesen $K = 20$ értékkel futtattuk modelljeinket, összesen 140 topikot hoztunk létre a teljes – 7 alkorpuszt tartalmazó – korpuszon. A címkézés után 65 egyedi topikot kaptunk, amelyek közül 49 csak egyszer, 16 pedig legalább kétszer jelent meg egy ciklusban. Az 1. táblázatban az első ciklus címkézésének eredményei láthatók, a 6. és 7. ábrán pedig ugyanezen topikok legnagyobb bétaértékű kifejezései.

6. ábra

Az első ciklus (1–10.) topikjait jellemző kifejezések

The *most important words of (1–10) topics in the first electoral cycle*



(Az ábra folytatása a következő oldalon)

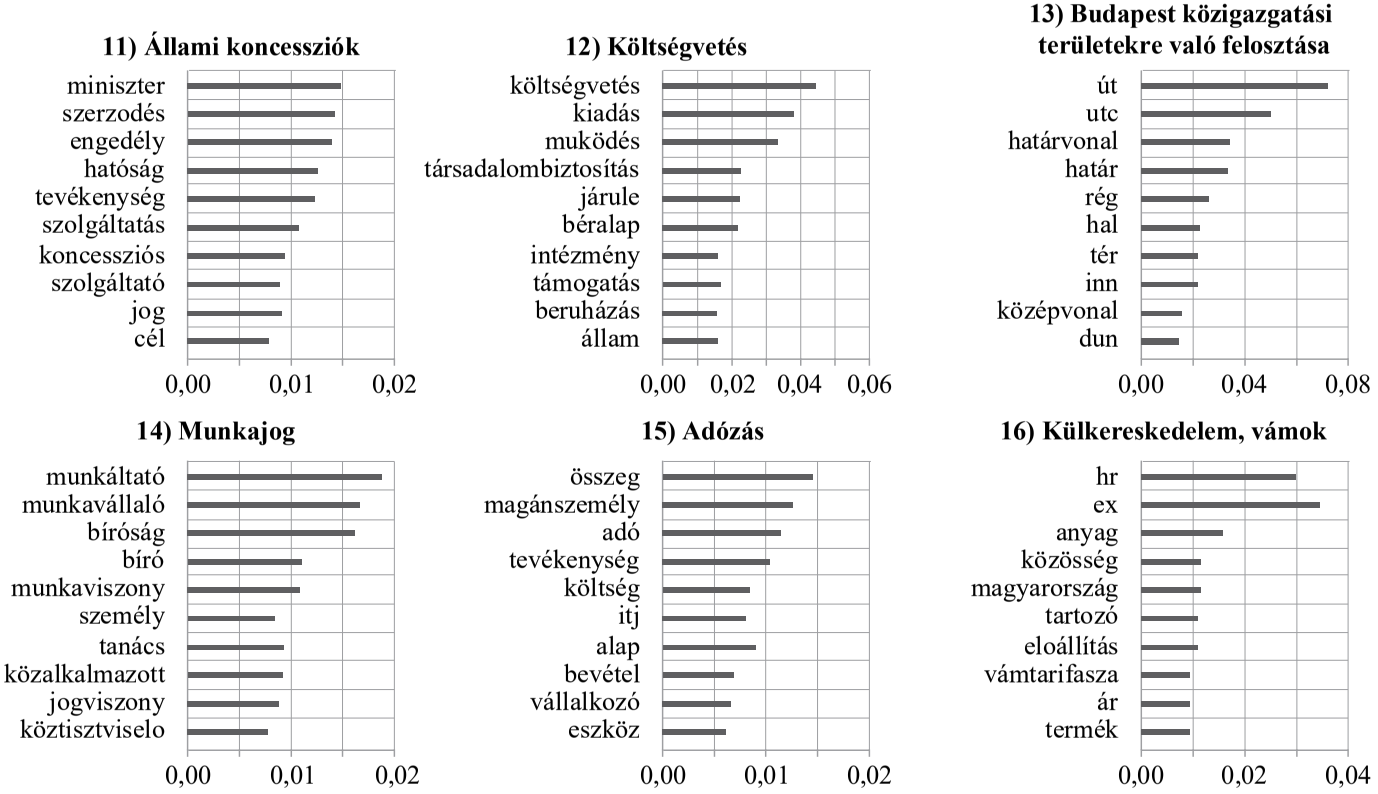
(folytatás)

9) Adózás
alap
adó
adóhatóság
lakás
összeg
adózó
jog
tevékenység
befektetés
kötelezettség
0,00 0,01 0,02 0,03
8) Önkormányzatok támogatása
megy
önkormányz
támogatás
budapes
ezer
sor
építés
önkormányzat
oszl.x
tornater
0,00 0,01 0,02 0,03
7) Privatizáció
gazdaság
társaság
iskol
intézmény
tanuló
állam
jog
szervez
vagyonügynökség
szerződés
0,00 0,01 0,02 0,03
10) Pénzügyi szervezetek
szerzodo
tag
személy
közgyulés
tertilet
pénztár
alapszabály
jog
szövetkez
megállapodás
0,00 0,01 0,02 0,03

7. ábra

Az első ciklus (11–20.) topikjait jellemző kifejezések

The *most important words of (11–20) topics in the first electoral cycle*



(Az ábra folytatása a következő oldalon)

(folytatás)

17) Szabadkereskedelem
ex
anyag
másfél
tartozó
vtsz
kivétel
eloállítás
ár
hasonló
termék
0,00
0,01
0,02
18) Választások
választás
tag
bizottság
terület
országos
kam
kamar
alap
kárpótlás
jelöl
0,00
0,01
0,02
0,03
19) Önkormányzatok, önszabályozó szervezetek
önkormányz
önkormányzat
intézmény
bizottság
állam
település
kisebbség
tanács
jog
tulajdon
0,00
0,01
0,02
0,03
20) Közalkalmazottak, közérdekű információk
adat
szolgálat
katon
országgyulés
személy
jog
alap
magyar
honvédelm
nyilvántartás
0,00
0,01
0,02

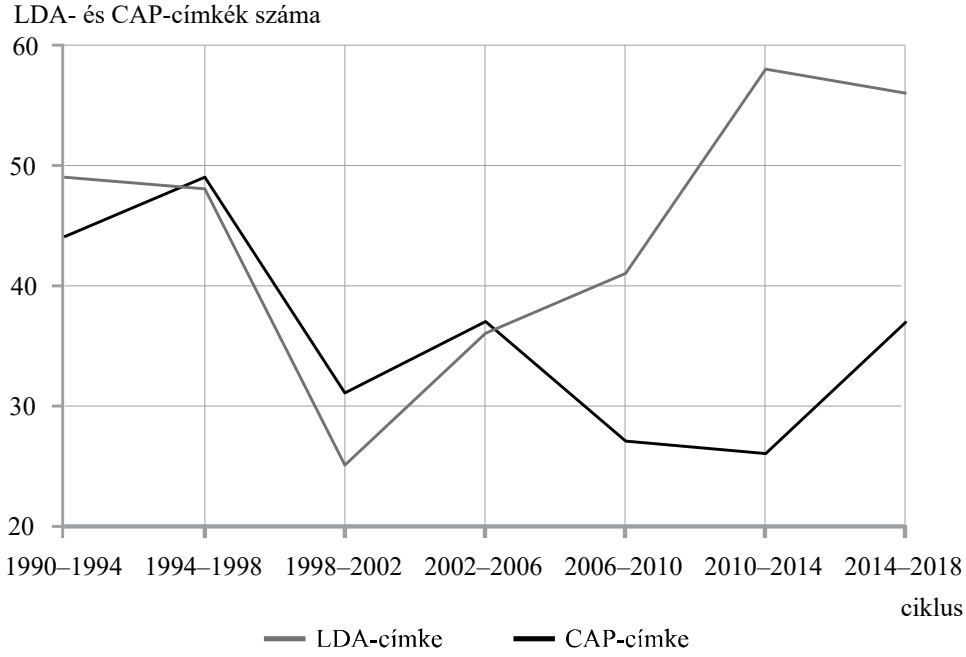
5. Validálás, értékelés

A CAP- és az LDA-csoportosítás eredményeinek összehasonlítását megnehezíti a két csoportosítás kategóriarendszereinek eltérése. Ennek ellenére vannak átfedések. Ezeket kihasználva a dokumentumok szintjén is összevetettük a két módszer eredményeit olyan esetekben, ahol a két csoportosítás egyes kategóriái látszólag azonos jelenséget írnak le (adózás, munkajog, környezetvédelem). Mivel az adózással foglalkozó topik mindegyik ciklusunkon futtatott LDA-ben jelen volt, ennek a trendjét egyszerűen összehasonlíthattuk a CAP-rendszer 107-es alcímkéjével (Adózás, adópolitika, adóreform) a 8. számú ábrán. Az eredmények alapján látható, hogy az első négy vizsgált ciklusban a két kategória mérete együtt mozog, viszont ezt követően az LDA-modelljeink jelentősen több adózással foglalkozó törvényt azonosítottak a korpuszon belül.

8. ábra

A CAP és az LDA adózással foglalkozó címkéi eredményeinek összehasonlítása a ciklusokon át

Comparison of the number of CAP and LDA labels concerning taxation throughout all subcorpi



Ennek az okát keresve elvégeztük a két csoportosítás dokumentumszintű vizsgálatát az első és a hetedik ciklusban, majd az eredményeket a 2. táblázatba rendeztük.

2. táblázat

A CAP és az LDA adózással foglalkozó címkéi eredményeinek összehasonlítása az első és a hetedik ciklusban

Comparison of CAP and LDA labels about taxation in the first and seventh electoral cycles

Ciklus	Topik	A topikhoz tartozó összes törvény	A topikhoz tartozó 107-es CAP-kódú törvények száma	Átfedés, %
1.	Adózás (9)	22	8	36,36
1.	Adózás (15)	27	19	70,37
7.	Adózás (15)	56	23	41,07

Ha csak a törvényekhez rendelt címkéket hasonlítjuk össze, akkor látható, hogy jelentős eltérések állnak fenn közöttük. Az első ciklus 15-ös számú topikjának kivételével mindegyik topiknak kevesebb, mint felét tették ki a 107-es CAP-alcímkét viselő törvények.

Viszont amikor egyenként átnéztük a törvényeket, láthattuk, hogy a címkékben lévő eltérés elsősorban nem hibás besorolások következménye. A topikok törvényei a két ciklusban mind adózással foglalkoznak, 7 törvény kivételével, amelyeket az LDA algoritmusa valóban tévesen kötött össze az adózással foglalkozó törvényekkel a tulajdonlással kapcsolatos kifejezések alapján. A többi esetben a címkék eltérésének okai egyrészt egy másik, adózáshoz kötődő CAP-címke jelenléte (a 2009-es számú altéma, azaz az adóadminisztráció), valamint az egyes CAP-kódok nehezen elválaszthatósága (pl. a Vállalatvezetés és versenyszabályozás és az Adózás, adópolitika, adóreform). Emellett az eltérések elsőszámú oka az, hogy az LDA adózással foglalkozó topikjába kerültek olyan szövegek, amelyeknél egy másik szakpolitikai területet az adóztatáson keresztül szabályozott a jogalkotó (pl. szerencsejáték, dohánytermékek, útalap).

Ezután azt is megvizsgáltuk, hogy mely 107-es CAP-alcímkével rendelkező törvények nem kerültek be az adózással foglalkozó topikokba. Az első ciklusban 44 törvényből 17 ilyen volt, amelyek a 3. táblázatban szerepelnek.

3. táblázat

A 107-es CAP-alcímkék besorolása az első ciklusban

Classification of documents with the number 107 CAP sublabel in the first electoral cycle

Topik	Törvények száma
Költségvetés	1
Választások	2
Privatizáció	1
Szociális ellátórendszer	6
Nemzetközi egyezmények (kettős adóztatás)	7

Ezeknek a legnagyobb része azért nem került be az adózással foglalkozó topikokba, mivel a kettős adóztatással foglalkozó nemzetközi egyezményeknek egy külön topikot alakított ki a modellünk. A szociális ellátórendszerrel foglalkozó topik pedig az illetékekkel foglalkozó törvények alapján került átfedésbe a 107-es CAP-címkével. A költségvetés topikba egy, az Országos Játékalapról rendelkező törvény került, így érthető az átfedés. A privatizáció topikja esetében pedig egy olyan törvény, amely egyszerre rendelkezik a gazdasági társaságokról és a társasági adóról. A választásokat érintő topik esetében beszélhetünk egyedül tényleges téves besorolásokról. A hetedik ciklus esetében hasonló trendet láthatunk, 37-ből 14 törvény nem került be az adózással foglalkozó topikba, ezeket a 4. táblázatba rendeztük.

4. táblázat

A 107-es CAP-alcímkék besorolása a hetedik ciklusban

Classification of documents with the number 107 CAP sublabel in the seventh electoral cycle

Topik	Törvények száma
Nemzetközi egyezmények (légi közlekedés)	1
Nemzetközi egyezmények (pénzügyek)	2
Nemzetközi egyezmények	1
Gazdasági megállapodások	10

Ebben az esetben az adózás kategóriái közötti átfedés hiányának oka teljes mértékben az volt, hogy az LDA konzisztensen külön topikokat alakított ki a különböző nemzetközi egyezmények törvénybe iktatásának. A két besorolás összehasonlításából látszik, hogy bár az adózás kategóriái közötti átfedés részleges, az LDA

mégis mindkét vizsgált ciklusban egy koherens besorolást alakított ki. Az eltérések oka pedig elsősorban a két csoportosítási módszer működésének eltérése.

A munkajog címkéje összesen két esetben, az első és a harmadik vizsgált ciklus modelljében jelent meg. Ha megvizsgáljuk ezeknek az átfedését az 5-ös számú CAP-címkével (amely a különböző, foglalkoztatást érintő törvényeket gyűjti össze), akkor láthatjuk, hogy az első ciklusban alacsony mértékű az átfedés, a harmadik ciklus egyik, munkajoggal foglalkozó topikja viszont jelentős átfedésben van a CAP-kódolásban jelen lévő megfelelőjével, ahogyan azt az 5. táblázat is mutatja.

5. táblázat

A CAP és az LDA munkajoggal foglalkozó címkéi eredményeinek összehasonlítása az első és a harmadik ciklusban

Comparison of CAP and LDA labels about labor law in the first and seventh electoral cycles

Ciklus	Topik	A topikhoz tartozó összes törvény	A topikhoz tartozó 5-ös CAP-kódú törvények száma	Átfedés, %
1.	Munkajog (14)	26	5	19,23
3.	Munkajog (1)	46	35	76,09
3.	Munkajog (9)	36	7	19,44

Az egyes törvények átvizsgálását követően világossá vált, hogy az első időszakban az alacsony mértékű átfedés oka elsősorban az, hogy a 14-es számú topikba azok a törvények is bekerültek, amelyek köztisztviselők, illetve bírák és ügyészek foglalkoztatását érintik, továbbá idekerültek a családi pótlékkal foglalkozó törvények is. A harmadik ciklusban, ha dokumentumok szintjén vizsgáljuk a kategóriát, láthatjuk, hogy a nem 5-ös CAP-címkét viselő törvények kamarákkal, illetve különböző közalkalmazottakat érintő szabályozásokkal foglalkoznak. Az 1-es számú topik esetében az eltérések okai a nemzetközi egyezmények, amelyek esetén 11-ből 3 foglalkozik közvetlenül munkajoggal, illetve a nemzetközi egyezmények mellett idekerült egy kultúrpolitikával foglalkozó törvény is. A 9-es számú topik esetében pedig kiderült, hogy az 5-ös CAP-kóddal nem rendelkező dokumentumok közül 9 nem foglalkozik közvetlenül munkavégzéssel, a maradék három kategória egyikébe tartozik, kamarákat szabályozó törvények, megválasztott és közigazgatási pozíciókkal kapcsolatos szabályozások, valamint mezőgazdasági munkálatokat érintő szabályozások.

Ezt követően ismét megnéztük, mely esetekben nem kerültek az 5-ös CAP-címkével rendelkező törvények a munkajoggal foglalkozó topikokba, majd topik szerint a 6. táblázatba rendeztük őket.

6. táblázat

Az 5-ös CAP-címkék besorolása az első ciklusban

Classification of documents with the number 5 CAP label in the first electoral cycle

Topik	Törvények száma
Költségvetés	1
Szociális ellátórendszer	3
Privatizáció	1
Választások	5

Egy, a Központi Ifjúsági Alapról szóló törvény a költségvetési topikba került be, mivel szövege jelentős mértékben foglalkozott az alap költségvetésének szabályozásával. A társadalombiztosítással kapcsolatos törvények a szociális ellátórendszer topikba kerültek, a szakszervezetekkel és kamarákkal foglalkozó törvények egy része pedig a választások topikba, amely az összes választási folyamatot leíró dokumentumot tartalmazza, beleértve a törvényeket, amelyek ezeket az önszabályozó szervezeteket érintik. Végezetül pedig azonosítottunk egy téves besorolást a privatizációval foglalkozó törvények topikjába. A harmadik ciklusban négyféle 5-ös CAP-címkével rendelkező törvény került a topikokba, amelyek nem a munkajoggal kapcsolatosak, ahogy az a 7. táblázat alapján is látható. Ezek közül három a bevándorlók és menekültek munkavégzését szabályozza, ezek az adatkezeléshez kerültek. A családtámogatással foglalkozó törvények és az egyes munkavégzéssel kapcsolatos állami támogatások is külön topikokba kerültek. Illetve egy, a gyermekmunka betiltásáról szóló nemzetközi egyezmény törvénybe iktatása is bekerült a nemzetközi egyezmények közé.

Ezekből az eredményekből látszik, hogy a harmadik ciklusban az eltérések oka nem elsősorban a téves besorolásban, hanem az egymással érintkező kategóriákban keresendő. Ezenkívül azt is megállapíthatjuk, hogy az első ciklus munkajog topikja alapján ugyanazt a tendenciát látjuk, mint az adózással foglalkozó topikok esetén. Ahogyan az adózásnál nem tett különbséget az LDA a kifejezetten adózással foglalkozó törvények és más szakpolitikai területek adózáson keresztül történő szabályozása között, úgy a munkajog esetében sem tudta elkülöníteni a kizárólag munkajoggal foglalkozó törvényeket és a közalkalmazottak munkakörét szabályozó törvényeket. A modelljeink tehát koherens, szemantikai-

lag értelmezhető témaköröket alakítanak ki, viszont ezek adott esetben szakpolitikák széles körét képesek lefedni.

7. táblázat

Az 5-ös CAP-címkék besorolása a harmadik ciklusban
Classification of documents with the number 5 CAP label in the third electoral cycle

Topik	Törvények száma
Adatkezelés	3
Gyermekvédelem	3
Szociális ellátórendszer	3
Nemzetközi egyezmények	1

Végezetül pedig a 8. táblázatban összevetettük a második ciklus 1-es számú LDA-topikját (környezetvédelem) is a 7-es számú CAP-címkével (amely különböző természetvédelmi témaköröket foglal magában).

8. táblázat

A CAP és az LDA környezetvédelemmel foglalkozó címkéinek összehasonlítása a második ciklusban
Comparison of CAP and LDA labels about environmental preservation in the second electoral cycle

Ciklus	Topik	A topikhoz tartozó összes törvény	A topikhoz tartozó 7-es CAP-kódú törvények száma	Átfedés, %
2.	Környezetvédelem (1)	29	6	20,69

Az eredmények alapján látható, hogy a topikba besorolt törvények döntő többsége nem rendelkezik a 7-es CAP-címkével. Ezek egyrészt a környezetvédelemmel összefüggésben álló, a vízgazdálkodásról és az erdő védelméről szóló törvények, de jelentős arányban vannak jelen a természeti erőforrások kihasználását szabályozó törvények is, amelyek az agrárpiachoz és a bányászathoz kapcsolódnak. Emellett bekerült több témakör is, amely nem kötődik szorosan a környezetvédelemhez, mint az energiapolitika, a helyi közszolgáltatások és a területfejlesztés. Ebben a ciklusban a 7-es CAP-címke csupán 8 alkalommal van jelen, és ebből 6 került be a címke topikmegfelelőjébe.

9. táblázat

A 7-es CAP-címkék besorolása a második ciklusban

Classification of documents with the number 7 CAP label in the second electoral cycle

Topik	Törvények száma
Adózás	1
Nemzetközi egyezmények	1

A maradék kettő egyike egy nemzetközi egyezmény, amely a környezetvédelemmel kapcsolatos, illetve egy környezetvédelmi termékdíjat szabályozó törvény. Tehát az LDA-besorolásunk sikeresen lefedi a CAP 7-es számú címkéjét, viszont ezenfelül további témakörök is kerültek bele, amelyek nem mindig kapcsolódnak szorosan a környezetvédelemhez. A környezetvédelmi topik esetében is egy szemantikailag értelmezhető, koherens témakört alakított ki a modellünk. Ugyanakkor a topik dokumentumszintű vizsgálata után egyértelműen látható, hogy a topik nem alkalmas a törvénykorpusz vizsgálatára, mivel túlságosan eltérő jogalkotási területekről származó törvényeket csoportosít egy kategóriába, természettel kapcsolatos kifejezések alapján.

6. Összegzés

Kutatásunk során célunk az LDA teljesítményének a kézi kódolással való összevetése volt. Eredményeink ennek a módszernek több előnyére és hátrányára is rámutattak. A kialakított topikmodellek validációja alapján elmondható, hogy az egyes topikok koherensek és szemantikailag értelmezhetőek. Ugyanakkor fontos megjegyezni, hogy ez nem feltétlenül jelenti azt, hogy ezek a csoportosítások kutatási kérdések szempontjából hasznosíthatók is, ahogy azt láttuk a második ciklus környezetvédelmi topikjának példáján is. A kialakított topikok relevanciája mellett kérdéses lehet még a kialakításuk logikájának konzisztenciája, amely kézi kódolás vagy felügyelt tanulást alkalmazó módszerek esetében nem áll fenn. Az első ciklus topikjaiban például jelen volt mind az adózás, mind az önkormányzatok támogatása topik, az előbbi egy közpolitikai terület, az utóbbi kialakítása pedig egy közigazgatási szint alapján történt.

További problémát jelentenek a topikmodellek alkalmazásánál azok az esetek, amikor a korpuszunk dokumentumainak egy jelentős része a kutatási kérdésünk szempontjából közvetlen relevanciával nem bíró tartalmi eltérésekkel rendelke-

zik. Jelen korpusz esetében ilyen problémát jelentettek a különböző nemzetközi megállapodások törvénybe iktatásai. Ezek jellemzően egyedi topikokba kerültek, és néhány esetben meg lehetett állapítani, hogy a nemzetközi egyezmények milyen specifikus típusából került kialakításra az adott topik, de többnyire nem. Ezek a problémák bizonyos mértékig kezelhetők további szövegtisztítási módszerekkel, viszont a túlzott tisztítás a rövid szövegek esetében (mint amilyenek korpuszunkban az egyes törvénymódosítások) megnehezítheti a pontos besorolásukat. Ebben a tekintetben elmondható, hogy a topikmodellek korpuszdependensek, a vizsgált korpusztól függően jelentős különbségek lehetnek a szövegtisztítás mértékében. Ahogyan azt a saját példánkon láthattuk, még egy egységes forrásból származó, egységes időtartamok szerint szegmentált korpusz esetében is szükség lehet nagyobb mértékű szövegtisztítási eljárásokra. Ennek a korpuszdependenciának egy másik oldala, hogy egyes témakörök körül, a hozzájuk kötődő egyedi kifejezésektől függően, gyakrabban és konzisztensebben történik a topikok kialakítása.

A validáció során látható volt, hogy egyes esetekben a topikok felcímkézésének bevett módszerei félrevezetőek lehetnek. Például a második ciklus 1. topikjáról, amelyről feltételeztük, hogy környezetvédelemmel foglalkozó törvényeket tartalmaz, megállapítottuk, hogy valójában számos eltérő szabályozási területről származó törvényeket tartalmaz, amelyek a természettel és a környezettel kapcsolatos kifejezések alapján kerültek egy topikba. Tehát amennyiben topikmodelleket alkalmazunk egy korpusz dokumentumainak kategorizálására, az eredményeinket validálni kell, vagy további csoportosítási módszereket kell alkalmazni a topikmodell eredményei alapján szűkített dokumentumok csoportján. Ehhez hasonlóan ugyan a ciklusonkénti modelljeink eredményei kellően konzisztensek voltak ahhoz, hogy több trendvonalat is felállítsunk, az adózás és a munkajog esete alapján látható, hogy ezek a csoportosítások nem kellően megbízhatóak ezeknek a trendekhez a felállításához. Az adózással foglalkozó topikjaink CAP-kódokkal való összevetésénél világossá vált, hogy egy ponton teljes mértékben megszűnik közöttük az összefüggés, a munkajoggal foglalkozó topikok esetében pedig kiderült, hogy inkonzisztensek a csoportosítások. Egy adott ciklusban a kamarák szabályozása bekerült a munkajoggal foglalkozó topikba, a következőben viszont már nem.

A definitív kategóriák kialakítása helyett a topikmodellek az adatok előzetes strukturálása terén rendelkeznek olyan előnyökkel, amelyek hiányoznak az alternatív csoportosítási módszerek esetében. Ezek az előnyök a költséghatékony alkalmazás, az előzetes ismeretek nélküli kategorizáció, illetve a topik és a dokumentumok szintjének elválasztása, amely lehetővé teszi, hogy ne csak azokat a dokumentumokat azonosítsuk, amelyek egy adott topikhoz a leginkább kötődnek, hanem mindazokat, amelyekben az bármilyen mértékben megjelenik. Ezek alap-

ján az egyes topikmodellek tudományos alkalmazásának lehetőségét elsősorban a kvalitatív feldolgozás, illetve más, kifinomultabb szövegbányászati módszerek augmentálásában látjuk, vagy a korpusz mintázatainak áttekintése által, vagy pedig a vizsgálandó dokumentumok azonosításában. Óva intünk azonban az alapos validáció nélküli, önálló alkalmazásuktól.

Irodalom

Aitchison, J. (1982): The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society.* Vol. 44. No. 2. pp. 139–160.

Arun, R. – Suresh, V. – Mdhavan, C. V. – Murthy, M. N. (2010): On finding the natural number of topics with latent dirichlet allocation: Some observations. In: Zaki, J. M. – Yu, X. Y. – Ravindran, B. – Pudi, V. (eds.): *Advances in Knowledge Discovery and Data Mining. Part II: 14th Pacific-Asia Conference, PAKDD.* Springer. Berlin.

Baerberá, P. – Bauer, P. C. – Ackermann, K. – Venetz, A. (2017): Is the left-right scale a valid measure of ideology? Individual-level variation in associations with „left" and „right" and left-right self-placement. *Political Behavior.* Vol. 39. No. 3. pp. 553–583. https://doi.org/10.1007/s11109-016-9368-2

Balasubrmanyan, R. – Cohen, W. W. – Pierce, D. – Redlawsk, D. P. (2012): *Modeling polarizing topics: When do different political communities respond differently to the same news?* Paper presented at the 6th International AAAI Conference on Weblogs and Social Media. 4–7 June. Dublin.

Balogh K. – Fülöp N. – Szabó M. K. (2017): A 2016-os tanártüntetések szövegeinek feldolgozása és adatvizualizációja interaktív dashboard segítségével. In: Vincze V. (szerk.): *XIII. Magyar Számítógépes Nyelvészeti Konferencia.* 299–307. old.

Basave, A. E. – Cano, E. – He, Y. – Xu, R. (2014): *Automatic Labelling of Topic Models Learned from Twitter by Summarisation.* Paper presented at the 52nd Annual Meeting of the Association for Computational Linguistics. June. Baltimore.

Bhatia, S. – Han, J. H. – Baldwin, T. (2016) *Automatic Labelling of Topics with Neural Embeddings.* Paper presented at the 26nd International Conference on Computational Linguistics. December. Osaka.

Blei, D. M. – Ng, A. Y. – Jordan, M. I. (2003): Latent Dirichlet Allocation. *The Journal of machine Learning research.* Vol. 3. pp. 993–1022.

Blei, D. M. – Lafferty, J. (2007): A Correlated Topic Model of Science. *The annals of applied statistics.* Vol. 1. No. 1. pp. 17–35. https://doi.org/10.1214/07-AOAS114

Boda Zs. – Sebők M. (szerk.) (2018): *A magyar közpolitikai napirend – Elméleti alapok, empirikus eredmények.* MTATKPTI. Budapest.

Boda Zs. – Sebők M. (2019): The Hungarian Agendas Project. In: Baumgartner, F. R. – Breunig, C. – Grossman, E. (eds.): *Comparative policy agendas*. Oxford University Press, Oxford. pp. 105–113.

Bolonyai F. – Sebők M. (2020): Kvantitatív szövegelemzés és szövegbányászat. In: Jakab A. – Sebők M. (szerk.): *Empirikus jogi tanulmányok*. Osiris, Budapest. pp. 361–380.

Browne, M. W. (2000): Cross-validation methods. *Journal of mathematical psychology.* Vol. 44. No. 1. pp. 108–132. https://doi.org/10.1006/jmps.1999.1279

Cao, J. – Xia, T. – Li, J. – Zhang, J. – Tang, S. (2009): A Density-Based Method for Adaptive LDA Model Selection. *Neurocomputing.* Vol. 72. No. 7–8. pp. 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011

Chang, J. – Blei, D. M. (2010): Hierarchical Relational Models for Document Networks. *The Annals of Applied Statistics.* Vol. 4. No. 1. pp. 124–150.

Chen, B. – Zhu, L. – Kifer, D. – Lee, D. (2010): *What Is an Opinion about? Exploring Political Standpoints Using Opinion Scoring Model*. Paper presented at the AAAI Conference on Artificial Inteligence. 11–15 July. Atlanta.

Deveaud, R. – Sanjuan, E. – Bellot, P. (2014): Accurate and Effective Latent Concept Modeling for Ad Hoc Information Retrieval. *Document numérique.* Vol. 17. No. 1. pp. 61–84. https://doi.org/10.3166/DN.17.1.61-84

Fang, Y. – Si, L. – Somasundaram, N. – Yu, Z. (2012): *Mining contrastive opinions on political texts using cross-perspective topic model.* Paper presented at the ACM international conference on Web search and data mining. 8–12 February. Seattle.

Farrell, J. (2016): Corporate funding and ideological polarization about climate change. *Proceedings of the National Academy of Sciences.* Vol. 113. No. 1. pp. 92–97. https://doi.org/10.1073/pnas.1509433112

Galántai J. – Pápay B. – Kubik B. G. – Szabó M. K. – Takács K. (2018): A pletyka a társas rend szolgálatában. Az informális kommunikáció struktúrájának mélyebb megértéséért a computational social science eszközeivel. *Magyar Tudomány.* 179. évf. 7. sz. 964–976. old.

Griffiths, T. L. – Steyvers, M. (2004): Finding Scientific Topics. *Proceedings of the National academy of Sciences.* Vol. 101. No. 1. pp. 5228–5235.

Grimmer, J. – Stewart, B. M. (2013): Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis.* Vol. 21. No. 3. pp. 267–297. https://doi.org/10.1093/pan/mps028

Günther, E. – Domahidi E. (2017): What Communication Scholars Write about: An Analysis of 80 Years of Research in High-Impact Journals. *International Journal of Communication.* Vol. 11. pp. 3051–3071.

Hajósi P. (2020): Mi van a számokon túl? Magyar blue-chip vállalatok negyedéves riportjainak kvalitatív szövegalapú elemzése. *Politikatudományi Szemle.* 29. évf. 3. sz. 107–127. old. https://doi.org/10.30718/POLTUD.HU.2020.3.107

Holecz M. – Szűcs Á. (2016): Politikai kötődés automata meghatározása az online sajtóban. In: Reményi A. – Sárdi Cs. – Tóth Zs. (szerk.): *Távlatok a Mai Magyar Alkalmazott Nyelvészetben.* Tinta Könyvkiadó. Budapest. pp. 233–240.

Hu, N. – Zhang, T. – Gao, B. – Bose, I. (2019): What do hotel customers complain about? Text analysis using structural topic model. *Tourism Management.* Vol. 72. pp. 417–426. https://doi.org/10.1016/j.tourman.2019.01.002

Jacobi, C. – Van Atteveldt, W. – Welbers, K. (2016): Quantitative Analysis of Large Amounts of Journalistic Texts Using Topic Modelling. *Digital journalism.* Vol. 4. No. 1. pp. 89–106. https://doi.org/10.1080/21670811.2015.1093271

Katona E. – Kmetty Z. – Németh R. (2021): A korrupció hazai online médiareprezentációjának vizsgálata természetes nyelvfeldolgozással. *Médiakutató.* 22. évf. 2. sz. 69–88. old.

Kerezsi D. – Bács Z. – Böcskei E. – Zéman Z. – Tarnóczi T. – Fenyves V. (2019): Magyarországi sporttevékenységet végző vállalkozások kiegészítő mellékletének összehasonlító elemzése az általános rész információi alapján. *International Journal of Engineering and Management Sciences.* 4. évf. 3. sz. 117–125. old. https://doi.org/10.21791/IJEMS.2019.3.11

Kleinberg, B. – van der Vegt, I. – Mozes, M. (2020): Measuring emotions in the Covid-19 real world worry dataset. In: Verspoor, K. – Cohen, B. K. – Dredze, M. – Ferrara, E. – May, J. – Munro, R. – Paris, C. – Wallace, B. (eds.): *Proceedings of the 1st workshop on NLP for COVID-19 at ACL 2020.* Association for Computational Linguistics.

Kollár Cs. (2020): Kína és a társadalmi kredit rendszere. *Hadtudomány: a Magyar Hadtudományi Társaság Folyóirata.* 30. évf. 2. sz. 79–97. old. https://doi.org/10.17047/HADTUD.2020.30.2.79.

Kou, W. – Li, F. – Baldwin, T. (2015): *Automatic Labelling of Topic Models Using Word Vectors and Letter Trigram Vectors.* Paper presented at the Asia Information Retrieval Symposium. 2–4 December. Bribane.

Lau, J. H. – Grieser, K. – Newman, D. – Baldwin, T. (2011): *Automatic Labelling of Topic Models.* Paper presented at the 49th Annual Meeting of the Association for Computational Linguistics.19–24 June. Portland.

Levy, K. – Franklin, M. (2014): Driving Regulation: Using Topic Models to Examine Political Contention in the US Trucking Industry. *Social Science Computer Review.* Vol. 33. No. 2. pp. 182–194. https://doi.org/10.1177/0894439313506847

Lin, W-H. – Xing, E. – Hauptmann, A. (2008): *A Joint Topic and Perspective Model for Ideological Discourse.*In Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 17–32. Springer. Berlin.

Miner, G. – Elder, J. – Fast, A. – Hill, T. – Nisbet, R. – Delen, D. (2012): *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press. Cambridge.

Nagy Gy. – Molnár Gy. (2017): Magyar Pedagógia folyóirat tudománymetriai elemzése: tendenciák, szerzők, társszerzőségi együttműködés. *Magyar Pedagógia.* 117. évf. 1. sz. 5–27. old. https://doi.org/10.17670/MPed.2017.1.5

Németh R. – Katona E. R. – Kmetty Z. (2020): Az Automatizált Szövegelemzés Perspektívája a Társadalomtudományokban. *Szociológiai Szemle.* 30. évf. 1. sz. 44–62. old. https://doi.org/10.51624/SzocSzemle.2020.1.3

Nikolenko, S. I. – Koltcov, S. – Koltsova, O. (2017): Topic Modelling for Qualitative Studies. *Journal of Information Science.* Vol. 43. No. 1. pp. 88–102. https://doi.org/10.1177/0165551515617393

Nyitrai T. (2021): A gépi tanulás módszereinek alkalmazása R-ben. *Statisztikai Szemle.* 99. évf. 2. sz. 173–198. old. https://doi.org/10.20311/stat2021.2.hu0173

Quinn, K. M. – Monroe, B. L. – Colaresi, M. – Crespin, M. H. – Radev, D. R. (2010): How to Analyze Political Attention with Minimal Assumptions and Costs. *American Journal of Political Science.* Vol. 54. No. 1. pp. 209–228.

Roberts, M. E. – Stewart, B. M. – Tingley, D. – Lucas, C. – Leder-Luis, J. – Gadarian, S. K. – Albertson, B. – Rand, D. G. (2014): Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science.* Vol. 58. No. 4. pp. 1064–1082. https://doi.org/10.1111/ajps.12103

Roberts, M. E. – Stewart, B. M. – Tingley, D. (2019): Stm: An R Package for Structural Topic Models. *Journal of Statistical Software.* Vol. 91. No. 1. pp. 1–40. https://doi.org/10.18637/jss.v091.i02

Rosen-Zvi, M. – Griffiths, T. – Steyvers, M. – Smyth, P. (2012): *The Author-Topic Model for Authors and Documents.* Paper presented at the Conference on Uncertainty in Artificial Intelligence. 7–11 July. Banff.

Rosen-Zvi, M. – Chemudugunta, C. – Griffiths, T. – Smyth, P. – Steyvers, M. (2010): Learning author-topic models from text corpora. *ACM Transactions on Information Systems.* Vol. 28. No. 1. pp. 1–38. https://doi.org/10.1145/1658377.1658381

Sebők M. – Kacsuk Z. (2020): The Multiclass Classification of Newspaper Articles with Machine Learning. *Political Analysis.* Vol. 29. No. 2. pp. 236–249. https://doi.org 10.1017/pan.2020.27

Sebők M. – Máté Á. – Orsolya R. (2021): *Szövegbányászat és mesterséges intelligencia R-ben.* Typotex. Budapest.

Sebők M. – Mészáros E. – Kis Gy. M. (2018): A politikai elit médiareprezentációja a rendszerváltás után: Egy napilap-címlapokra épülő szövegbányászati elemzés. *Politikatudományi Szemle.* 27. évf. 1. sz. 91–112. old.

Sebők M. (szerk.) (2016): *Kvantitatív szövegelemzés és szövegbányászat a politikatudományban.* L'Harmattan Kiadó. Budapest.

Tikk D. (2017): *Szövegbányászat.* Typotex. Budapest.

Wang, X. – Mccallum, A. (2006): *Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends.* Paper presented at the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 20–23 August. Philadelphia.

Yau, C – Porter, A – Newman, N. – Suominen, A. (2014): Clustering Scientific Documents with Topic Modeling. *Scientometrics.* Vol. 100. No. 3. pp. 767–786. https://doi.org/10.1007/s11192-014-1321-8

Zhao, Z. – Chen, J. C. – Perkins, R. – Liu, Z. – Ge, W. – Ding, Y. – Zou, W. (2015): A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC bioinformatics.* Vol. 16. No. 13. pp. 1471–2105.