Statisztikai Szemle

Közzététel: 2022. június 14.

A tanulmány címe:
A kvantitatív szövegelemzés lehetőségei a menedzsmenttudományban

Szerző:

STRELICZ ANDREA,
a Pannon Egyetem PhD-hallgatója
E-mail: strelicz.andrea@gmail.com

DOI: https://doi.org/10.20311/stat2022.6.hu0551

Strelicz Andrea

A kvantitatív szövegelemzés lehetőségei a menedzsmenttudományban

Possibilities of quantitative text analysis in management science

STRELICZ ANDREA,
a Pannon Egyetem
PhD-hallgatója
E-mail: strelicz.andrea@gmail.com

A kvantitatív szövegelemzés egyfajta dokumentum, valamint a dokumentumok tartalmát alkotó szavak, kifejezések csoportosítására szolgáló, statisztikai módszerekkel és eszközökkel megtámogatott eljárások halmaza. Legfőbb előnye, hogy a numerikus adatok által kinyerhetőkön túl lehetővé teszi a további, eddig elérhetetlennek tűnő információkhoz való hozzáférést. A kreatív szövegelemzést tehát leginkább akkor használják, amikor a szöveg több információt tartalmaz, mint amennyi a számszerűsített adatok alapján kinyerhető lenne. A szöveg- és tartalomelemzés mára az egyik legizgalmasabb és leginkább fókuszált területe az adatbányászatnak és az adatelemzésnek, mind matematikai, statisztikai, módszertani, mind szoftveres illeszkedés szempontjából. Ez a tanulmány egy konkrét példán keresztül a szövegelemző eljárások közül a látens Dirichlet-allokációt (latent Dirichlet allocation, LDA) mint egy lehetséges módszertani megközelítést mutatja be. Elsődleges cél azt bebizonyítani, hogy a szövegelemzés hosszú távú és újszerű perspektívákat jelenthet a menedzsmenttudomány területén.

TÁRGYSZÓ: kvantitatív szövegelemzés, topikmodell, látens Dirichlet-allokáció,

Quantitative text analysis is a set of procedures for grouping documents and the words and expressions that make up the content of documents, supported by statistical methods and tools. Its main advantage is that in addition to the information that can be extracted from numerical data, it provides a solution for accessing additional information that previously seems unavailable. Thus, text analysis is most commonly used where the text contains more information than could be obtained from quantified data. Text and content analysis is today one of the most exciting and focused areas of data mining and data analysis, both in terms of mathematical, statistical, methodological, and software fit. In this study, latent Dirichlet allocation (LDA) is presented as a possible methodological approach through a concrete example. The primary aim of the article is to show that text analysis can provide long-term and novel perspectives in the fields of management science.

KEYWORD: quantitative text analysis, topic model, latent Dirichlet allocation

A kvantitatív szövegelemzés többtényezős (vagyis a modellalkotáshoz szükséges, több beállítási paramétert értelmező), iterációs folyamat, amelynek használatakor nem kell az első beállítás szerinti eredményeket elfogadni. A szövegelemzés során az adott beállítások mellett a programozott algoritmusok kiadnak egy olyan eredményt, amelyet nem szükséges, de nem is érdemes elfogadni. A feldolgozandó adathalmaz (szövegelemzés esetében a korpusz) általánosságban rendelkezik olyan korlátokkal, amelyek determinálják az eredményeket, de vannak olyan közös metszetek is az adathalmazon belül, amelyek megváltoztatása esetén az adathalmazra kiadott eredmény összességében képes módosulni. Mindkét irány kezelhető és értelmezhető az ún. hiperparaméter-optimalizálással, amelynek segítségével keresni és találni lehet egy olyan eredményt, ami a kutatás, vagy szándék szempontjából a legjobban illeszkedik (*Feurer–Springenberg–Hutter* [2015]). Jellemzően a gépi tanulás során használnak hiperparaméter-optimalizálást, a feldolgozandó adatok homogenizálása érdekében, ugyanakkor maga az elv és a megvalósítás alkalmazható a nem gépi tanulásos eljárások esetében is. Általánosságban hiperparaméternek tekinthető minden olyan paraméter, amelynek értékét a tanulási folyamat vezérlésére kell használni. Minden egyéb nem hiperparaméter-érték (általában a csomóponti súlyok) megtanulható. A kvantitatív szövegelemzés során az eredménymodell a hiperparaméterek módosításával, hangolásával változtatható (*Sayak* [2018]). Hiperparaméter lehet bármely szövegelemző eljárás esetében a karakterkódolás, a szövegtisztítás jellege és mértéke, a korpuszba besorolandó szavak, kifejezések minimumgyakoriságának meghatározása, vagy maguk a feldolgozandó dokumentumok. Ennek az az oka, hogy a dokumentumok tömeges feldolgozása nem teszi lehetővé azoknak a kutatás szempontjából értelmetlen, zavaró vagy fölösleges szavaknak, jeleknek, karaktereknek a manuális kiszűrését, amiket a standard szövegtisztítási eljárások nem fednek le. Így az eredmények alapos vizsgálatát követően lehet megállapítani további műveletek szükségességét a szöveg homogenitását, illetve értelmezhetőségét illetően.

A kvantitatív szövegelemző eljárások közül a topikmodell iránt közel 20 éve mutatkozik egyre növekvő érdeklődés a statisztikai szakértők körében. Ez idő alatt mind a matematikai háttere, mind az alkalmazási területek lehetősége komoly fejlődésen ment keresztül. Utóbbiakat illetően még bőséges lehetőség áll a kutatók és fejlesztők előtt, hiszen a programok és az algoritmusok folyamatos fejlesztése teremt lehetőséget az alkalmazhatóság kiterjesztésére. Jó példák erre a különböző menedzsmentkérdésekre adható válaszok, vagy a szakmaspecifikus tudás halmazolásának lehetősége.

A topikmodell-eljárások közül az egyik legjelentősebb előrelépés az LDA-eljárás kidolgozása volt, ami *Blei* nevéhez fűződik, és 2003-ra sikerült publikálható eredményként közzétenni (*Blei–Ng–Jordan* [2003]). Az azóta eltelt 19 év alatt az LDA-ből számtalan származtatott topikmodell-eljárás készült, attól függően, hogy az eljárás mely részét kívánták hangsúlyosan hatékonyabbá vagy megbízhatóbbá tenni. Ilyen a dinamikus topikmodell (dynamic topic model, DTM), amely a topikok időbeni változására fekteti a hangsúlyt (*Blei–Lafferty* [2006]), a korrelált topikmodell (correlated topic model, CTM), amelynek legfőbb célja a topikok közötti interakciók vizsgálata és modellezése (*Blei–Lafferty* [2007]), vagy a szintaktikus topikmodellezés (syntactic topic model, STM), amely a korpuszban levő szavak szintaktikai korlátjaira épül (*Byord-Graber–Blei* [2009]). *Balogh* [2015]-ben közreadott tanulmányában összegyűjtötte a létező összes eddig kifejlesztett topikmodell-megközelítést, amelyek a különböző feltételeket hangsúlyozzák, vagy elhagyják. A származtatott módszerek mellett természetesen kritikák is megfogalmazódtak a módszert illetően, többek között a topikok számának a meghatározása mögötti módszertan hiánya miatt (*Gerlach–Peixoto–Altmann* [2018]). Bár ez a kritika helyénvaló, személyes tapasztalat alapján elmondhatom, hogy nem feltétlenül segíti a kutatót, ha nem találkozik az iteráció adta változásokkal. Az iterációs folyamat adta eredmények lehetővé teszik a kutató számára, hogy anélkül ismerkedjen meg a korpusszal, hogy végig kellene olvasnia valamennyi dokumentumot, ugyanakkor összegző rálátást tesz lehetővé. Módszertani integráció tekintetében említésre méltók az idősoros elemzéssel (*Blei–Lafferty* [2006], *Pruteanu-Malinici et al*. [2010]) vagy külső kovariátoktól függő beállításokkal (*Mimno–McCallum* [2012]) kiterjesztett topikmodell-eljárások.

A „kollektív tudat” megismerésére mind a felhasználók, mind a szakértők szerint az egyik legígéretesebb módszer a topikmodell-eljárás (*Barde–Bainwad* [2017]). Ez a dokumentumcsoportosító módszerek közül az összegző vagy klaszterező módszerek közé tartozik. Alapvetően három alkotóelemből dolgozik: a dokumentum a dokumentumot alkotó szavak, és a dokumentumokból álló korpusz. Az e három változó egymáshoz fűződő valószínűségi viszonya mögötti rejtett változókat tárja fel az eljárás – vagyis kiküszöböli a humán kódolás adta szubjektív korlátokat (*Sebők* [2016]). A topikmodell iránti érdeklődés lényege, hogy a létrejött csoportokat az eljárás címkékkel látja el, vagyis a kutatónak nem kell a csoportokban levő kifejezéseket szubjektíven értelmezni, mert a módszer az értelmezést elvégzi a kutató helyett, a teljes topik szemantikai értelemben vett összefüggését (a rejtett változót) megadva.

1. Látens Dirichlet-allokáció (LDA)

A topikmodellezés olyan aktív kutatási terület, amelyet a legnagyobb és legismertebb keresőmotorban, a Google-ban is meg lehet találni (*Miner et al*. [2012]). Ebben a fejezetben bemutatom a módszer célját, mélyebb matematikai megalapozás nélkül a statisztikai hátterét, illetve a működési elvét. Továbbá szó lesz az alkalmazási területek széleskörűségéről, valamint azokról a fejlődési irányokról is, amelyek magát a topikmodell-eljárást helyezik a fókuszba.

A dokumentumcsoportosító módszerek kétfelé választhatók. Az egyik az összegző vagy klaszterező, a másik az osztályozó módszerek csoportja. Az összegző módszerek alá tartozik a topikmodellezés is (*Ashish–Paul* [2016]), amelynek egyik legismertebb fajtája a LDA. A topikmodell-eljárás tehát a klaszterező eljárások közé tartozik, és a dokumentum- vagy szövegklaszterrel foglalkozó fejlesztő kutatók nagy része is a topikmodell-eljárásokat tekinti az egyik legígéretesebb fejlesztési iránynak (*Barde–Bainwad* [2017]).

„Az LDA-módszer előnye, hogy az azonosított témák olyan szavakat is tartalmazhatnak, amelyek elsőre nem feltétlen jutottak volna a humán kódoló eszébe, így vélhetően – ha osztályozással próbáltuk volna megoldani ezt a feladatot – nagy eséllyel kimaradtak volna." (*Sebők* [2016] 96. oldal). A topikmodell a klaszterező eljárások között azért érdemel kiemelt figyelmet, mert a létrejött „puha klasztereket" (*Miner et al*. [2012]) címkékkel, ún. „topic tag"-ekkel látja el, ellentétben az egyéb klaszterező eljárásokkal. A címke egy összefoglaló kifejezés, a leginkább meghatározza a klasztert, illetve a benne lévő szavak csoportját. Vagyis a topik a csoportban található szavak szemantikai lényege. A puha klaszter ebben a kontextusban annyit tesz, hogy egy kifejezés akár több klaszter alá is tartozhat (*Miner et al*. [2012]).

Számos topikmodellezésre alkalmas módszert dolgoztak már ki, de mindegyiknek ugyanaz a legfőbb célja: a rejtett változók (szemantikai összefüggések) megtalálása adott dokumentum halmazban. A topikmodellek olyan strukturált eloszlásból állnak, amelyekben a megfigyelt adatok (vagyis a korpuszban lévő szavak) kölcsönhatásba lépnek a rejtett, véletlen változókkal. Rejtettvalószínűségiváltozómodellel a kutató rejtett struktúrát helyez el a rendelkezésre álló adatokban, és ezt hátsó valószínűségi következtetések felhasználásával megtámogatva hozza létre a témákat (*Blei–Lafferty* [2009]). A legismertebb módszerek a rejtett szemantikus allokáció (latent semantic allocation, LSA), a rejtett szemantikus indexelés (latent semantic indexing, LSI) a rejtett vagy LDA (*Ashish–Paul* [2016]). A jelen kutatás szempontjából utóbbival foglalkozik részletesebben ez a fejezet.

Az LDA *Blei*, *Ng* és *Jordan* nevéhez fűződik. (*Blei–Ng–Jordan* [2003]). Bleiről annyit érdemes tudni, hogy az LDA mellett más topikmodell-eljárásokkal is dol-

gozik, de az összegző vagy a klaszterező eljárások kapcsán is több publikációja jelent már meg.

Az elmúlt 19 év alatt számtalan, az LDA-ből származtatott topikmodell-eljárás látott napvilágot, attól függően, hogy az eljárás mely részét kívánták hangsúlyosan hatékonyabbá, vagy megbízhatóbbá tenni. Ilyenek az alábbiak is:

– DTM, amely a topikok időbeni változására fekteti a hangsúlyt (*Blei–Lafferty* [2006]);

– CTM, amelynek legfőbb célja a topikok közötti interakciók vizsgálata és modellezése (*Blei–Lafferty* [2007]);

– STM, amely a korpuszban lévő szavak szintaktikai korlátjaira épül (*Byord-Graber–Blei* [2009]).

A felsorolt példákon túl léteznek további eljárások, amelyek legfőbb célja a feltételek szűkítése vagy kiterjesztése. Erre vonatkozóan *Balogh* [2015]-ben kiadott tanulmányában összegyűjtötte a létező összes eddig kifejlesztett olyan topikmodell-megközelítést, amelyek a különböző feltételeket hangsúlyozzák, vagy elhagyják.

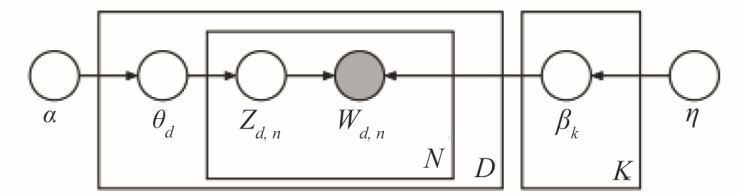
A származtatott módszerek mellett természetesen megfogalmazódtak kritikák is, többek között a topikok számának meghatározása mögötti módszertan hiánya miatt (*Gerlach–Peixoto–Altmann* [2018]). Bár a kritika helyénvalónak látszik, amennyiben a topikmodelleket egyfajta emeltebb szintű klaszteréljárásnak tekinti a szakirodalom, személyes tapasztalat alapján elmondhatom, hogy nem feltétlen segíti a kutatót, ha nem tapasztalja meg az iteráció okozta változásokat. Az iterációs folyamat adta eredmények lehetővé teszik a kutató számára, hogy anélkül ismerkedjen meg a korpusszal, hogy végig kellene olvasnia valamennyi dokumentumot, ugyanakkor egy összegző rálátást is lehetővé tesz.

A topikmodell-eljárások a korpuszt szózsákként kezelik, miáltal a szó a mondatban levő helyét elveszti, illetve a dokumentumok sorrendjének már nem lesz jelentősége. A szózsák előnye, hogy a szavak és a dokumentumok halmazban történő kezelése jobban segíti a témák szemantikájának a megértését (*Ashish–Paul* [2016]).

Az LDA általánosan megfogalmazva egy generatív valószínűségi modell, a mögötte lévő alapfeltételezés az, hogy a k számú topik minden (a korpuszban lévő) dokumentumot valamilyen arányban reprezentál. Ez a legáltalánosabb feltételezés, mivel a korpuszban található dokumentumok általában heterogének (vagyis önállóan nézve a tartalmuk eltérő), viszont egyesítve, azaz korpuszként kezelve olyan fő vagy részelképzeléseket, témákat adnak ki (*Blei–Lafferty* 2009), amelyek esetében előfordulhat, hogy sem a kulcsszavas keresés, sem a humán annotátorok által megadott tematikus besorolás nem ad ki eredményt. Az 1. ábrán látható az LDA, bayesi hierarchikus modellként ábrázolva, amelyben az LDA keretei között egy dokumentumot a különböző témák közötti elosztásnak kell tekinteni (*Liu–Niculescu-Mizil–Gryc* [2009]). A modellalkotás során kiszámítjuk az egyes szavak súlyértékeit, ezek alapján jönnek létre a topikok, majd a topikok-

ban külön meghatározzuk az alájuk tartozó szavakat. A bavesi hierarchikus modell kör nélküli gráfként mutatja be a valószínűségi változók eloszlását, ezzel általában a topik-modell, de azon belül a különböző LDA-változatok és -adaptációk modellfüggőségének magyarázatát szemléltetik. Az ábra három szintet azonosít: korpuszszint, dokumentumok szintje, szavak szintje (*Blei–Ng–Jordan* [2003]). Az élek a valószínűségi változók közötti függőséget jelölik. A körök vagy csomópontok a dirichletelsőbbségi paraméterek, vagyis az árnyékolt csomópontok a megfigyelt változókat jelölik; az üres csomópontok a rejtett változókat. A téglalapok a változók replikációit (többszörös értékeit) jelentik (*Blei–Lafferty* [2009]). Az 1. ábra magyarázatát az 1. táblázat tartalmazza.

1. ábra. A látens Dirichlet-allokáció (LDA) grafikus modellábrázolása
(Graphical model of Latent Dirichlet Allocation)



Forrás: *Ble* [2010].

1. táblázat

Az 1. ábra jelmagyarázata
(Legend of Figure 1)

Jelölés	Jelentés	Szint	Megjegyzés
K	Témák	–	–
D	Dokumentumok	–	–
N	Szavak	–	A különböző származtatott ábrákon esetenként a szakemberek indexelést alkalmaznak, jelölve, hogy ez az érték dokumentumonként eltérhet
α és β	A Dirichlet elsőbbségeinek (priori valószínűségek) paraméterei	Korpusz	Egyszeri mintavétellel kerül a korpuszba (*Blei* [2003])
θ	A dokumentumonkénti témák aránya	Dokumentum	Dokumentumokonként egyszer mintázva
$Z_{d,n}$	–	Szavak	Mindegyik dokumentumban minden egyes szóhoz egyszeri mintát jelent (ezek a megfigyelt változók), emellett arra utal, hogy az LDA a vegyes tagsági modellek közé tartozik (*Blei* [2010], *Blei–Carin–Dunson* [2010])
$W_{d,n}$			

Általánosságban kijelenthető, hogy Magyarországon a topikmodellel összefüggő kutatások és tudományos publikáció száma kevés. Az eddig vizsgált témák között figyelemre méltó a magyarországi korrupcióval kapcsolatos kutatás (*Katona* [2018]), a romaellenes megnyilvánulások tematikus elemzését vizsgáló kutatás (*Balogh* [2015]), valamint a politikai szövegek társadalomtudományi elemzését tartalmazó tanulmány (*Sebők* [2016]). A külföldi publikációk tekintetében folyamatos az új ismeretekhez, lehetőségekhez és megoldásokhoz való hozzáférés. Ezen a ponton már nem érdemes az LDA-módszert hangsúlyozni, hanem általánosan kell a topikmodellre vonatkozó eredményeket figyelni, hiszen az elsődleges cél mindegyik esetben ugyanaz.

Bár a csoportosító eljárások többsége, vagyis a topikmodell, azon belül pedig az LDA-módszer sem tartozik a felügyelt gépi tanulási folyamatok közé, léteznek olyan fejlesztett változatok, amelyek igen (*Blei–McAuliffe* [2007]). A fejlődési irányok pedig már inkább módszertani integrációk felé haladnak, nem úgy, mint az idősoros elemzéssel (*Blei–Lafferty* [2006], *Pruteanu-Malinici et al*. [2010]) vagy külső bementi változóktól függő beállításokkal (*Mimno–McCallum* [2012]) kiterjesztett topikmodell-eljárások.

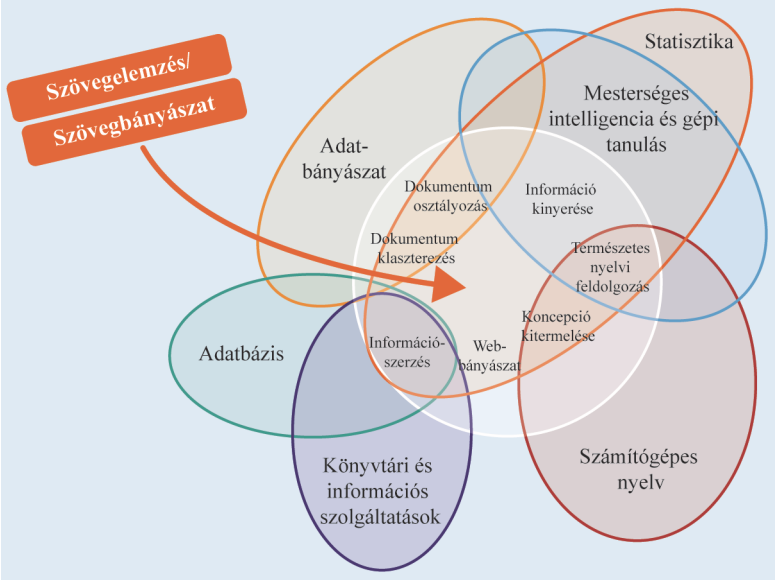
Érdemes szót ejteni néhány érdekesebb területről, ahol topikmodell-eljárással értelmeztek szöveget és/vagy válaszokat, megoldásokat kerestek; például dalok szövegeinek az elemzésével (*Hoffman–Blei–Cook* [2008], [2009]), vagy régészeti ásatásokon fellelt szövegek elemzésével, amelyek alapján kultúrasajátosságokat kerestek (*Mimno* [2012]) továbbá a teológia területén vallási szövegek (*Sperling* [2020]) értelmezésével, esetleg a pénzügyi szektorban az ügyfélmegtartás lehetőségével (*Moro* [2015]).

A topikmodell-eljárás nemcsak a matematikusok és informatikusok érdeklődésének a központja, hanem az alkalmazhatóság értelmében vett olyan területeké is, ahol a lényeglátás, összegzés igénye felmerül. Magyarországon a publikált korrupciós ügyek (*Katona* [2018]), a romaellenes megnyilvánulások (*Balogh* [2015]), vagy a politikai szövegek társadalomtudományi tartalma (*Sebők* [2016]) is topikmodell-eljárással lett vizsgálva, különböző kutatások keretein belül. Továbbá széles körben alkalmazható a topikmodell-eljárás olyan területeken, mint az internetes bányászat, a keresőmotoros optimalizálás (search engine optimization, SEO), az információkeresés és a topológiai elemzés. Lehetőséget kínál a szöveges adatokból történő információ kinyerésre, numerikus adattranszformáció nélkül, és az ezzel járó nem kívánatos adatvesztés elkerülésével. De a topikmodell eljárás alkalmas szakirodalmi áttekintésre adott témában online vagy fix szöveges tartalmak elemzésére, ha az adatmennyiség túl nagy és átfogó, tartalomösszegzésre irányuló információra van szükség, akár modellredukciós, akár egyszerű szógyakoriság vizsgálati eljárások segítségével. A topikmodell-eljárás alkalmazásának legizgalmasabb területe jelenleg a szöveg- és tartalomelemzésnek a közösségi médiában való aktivitás letükrözése a

felhasználó számára a keresések és a hangulatreakciók (emoticonok használata) által, hiszen ennek a kettőnek a segítségével a Facebook, a Twitter és a többi közösségi-média-felület a felhasználó érdeklődési körét adja vissza, kiegészítve a témáihoz kapcsolódó lehetőségek felajánlásával (*Settanni–Marengo* [2015], *Munoz-Garcia–Navarro* [2012], *Wang et al*. [2014], *Zou–Song* [2016]).

Számos iparágban alkalmaznak még szövegelemző módszereket a szöveges adatkészletek mennyiségi és minőségi megértése érdekében (*Chen* [2020]). A szövegelemzést több területen lehet értelmezni, amihez áttekintést ad a 2. ábra. A Venn-diagram segítségével láthatók a területek, valamint a metszetük, amelyek a területek közötti átfedésekre utalnak. A 6 terület a statisztika, a mesterséges intelligencia és a gépi tanulás, a számítási nyelvészet, a könyvtári és információs szolgáltatások, az adatbázisok és adatbányászat.

2. ábra. A szövegelemzés metszéspontjának ábrázolása Venn-diagram segítségével
(The intersection of text analysis using a Venn diagram)



Forrás: *Chen* [2020].

Érdemes megemlíteni a szövegelemző eljárások egy komplex megoldását, amihez külön szoftverek készülnek, ez pedig a chatbot (*Chen* [2020]), amely képes az emberihez hasonló beszélgetésre és interaktív kommunikációra, valós személy beavatkozása nélkül (Creative Site, év nélkül).

Ami a szoftveres támogatást illeti, a szöveg- és tartalomelemzésre számos kiváló szoftver áll rendelkezésre, a felhasználók kompetenciaszintjének és igényeinek megfelelően. Alapesetben négyféle megközelítés létezik: a programnyelves, a Windows-felülethez hasonló felépítésű, és a folyamatdiagram-jellegű, továbbá ezek kombinációi. Mindegyiknek megvannak az előnyei és a hátrányai, így a célnak és a felhasználó ismereteinek megfelelő megközelítés kiválasztására széleskörben érhetők el fizetős és ingyenes lehetőségek.

Számos olyan, kimondottan szövegelemzésre specifikált szoftver van forgalomban, ingyenes vagy fizetős változatokban, amelyek nem igényelnek programozónyelv-tudást. A szoftverek közül a textmineR-csomag Rstudioban még ma sem egy hétköznapi statisztikai eszköz Magyarországon. A textmineR-csomag legegyszerűbb parancssora a 3. ábrán látható.

3. ábra. Topikmodell R-parancssor
(Topic model R script)

```
set.seed(12345)

model <- FitLdaModel(dtm = dtm,
                     k = 30,
                     iterations = 1000,
                     burnin = 180,
                     alpha = 0.1,
                     beta = 0.05,
                     optimize_alpha = TRUE,
                     calc_likelihood = TRUE,
                     calc_coherence = TRUE,
                     calc_r2 = TRUE,
                     cpus = 4)
```

Megjegyzés. dtm = dtm: két ugyanazon *dtm* transzformációja, vagyis ugyanannak a korpusznak vizsgálja a dokumentumait és a szavait – a modell tanítására szolgál; *k:* a kutató által megadott, kívánt topikszám; *iterations:* a kutató által megadott, kívánt iterációs szám; *alpha és beta:* a dirichlet elsőbbsége, vagyis a témákhoz kapcsolódó dokumentumokhoz és a szavakhoz kapcsolódó témákhoz; *calc-likelihood:* ez számolja ki a log-likelihood *p*-értéket (szavak/topik) minden iterációnál; *calc-coherence:* számoljon a topikokon belül a kifejezések között valószínűségikoherencia-értéket; *calc-r2:* számolja a modell R^2 értékét; *cpus:* a használatban levő számítógép futási idejét befolyásoló teljesítményadat, vagyis hány központi feldolgozóegységet (central processing unit, CPU) használjon a gép a teszt futtatása során.[1]

Forrás: *Jones* [2014].

[1] https://www.rtextminer.com/articles/c topic modeling.html

2. A szövegelemzés kutatási keresztmetszete: az üzletmenetfolytonosság-menedzsment

Az üzletmenetfolytonosság-menedzsment olyan koncepcióból indult ki, amelynek legfőbb célja általános értelemben a veszteségcsökkentés volt. Utóbbi azonban ebben az esetben globális kontextusban értelmezhető, az alábbi peremszempontokat alapul véve:

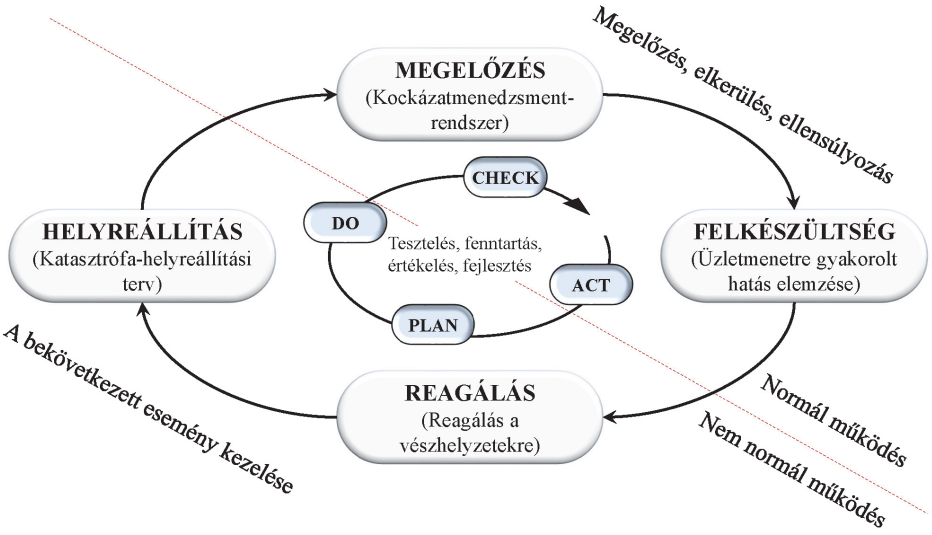
> – az emberiség jólétét, biztonságát, vagy eszmei értékeit érintő, visszafordíthatatlan károk,
> – olyan károk és veszteségek, amelyek egy vagy több nemzetgazdaság fejlődését hátráltatják a helyreállítási költségek által,
> – olyan mértékű emberáldozat, amely hatást gyakorol egy nemzet összetételére, biztonságára, nemzeti és kulturális bővülésére (*Tucker* [2015]).

Az üzletmenetfolytonosság-menedzsment ennek a globális törekvésnek az a szűkített keresztmetszete, amely szerint egy nemzet fejlődését és fenntartható gazdasági növekedését a benne működő gazdasági társulások léte és folyamatosan biztonságos működése garantálja. Gazdasági társulás kifejezés alatt pedig az egy nemzetet alkotó valamennyi gazdasági szereplőt kell érteni, a kormányzati és a nonprofit társulásokkal együtt.

A 4. ábrán látható a FEMA (Federal Emergency Management Agency) vészhelyzetmenedzsment-stratégiája, amely egyaránt értelmezhető mikro- és makroszintre. A stratégia nem ígéri, hogy minden lehetséges veszélyt el tud kerülni egy gazdasági társulás, de olyan felkészültségre kényszerít, amely segítségével a válságos állapotokra és helyzetekre létezik kész válasz. A stratégiának van egy proaktív és egy tervezetten reaktív része, ami egyben elkülönített és folyamatosan rendelkezésre álló erőforrásokat, valamint begyakorolt módszereket, eszközöket, illetve akciókat feltételez a válságok kezelésére. A négypilléres stratégia egyes elemeiben, legalább analógiás értelmezésben, mondhatni, közismertek, kivéve az üzletmenetre gyakorolt hatások elemzését. Noha a kockázatok és a hatások módszertani szinten gyakran kéz a kézben járnak, értelmezésük és viselkedésük esetlegesen össze is mosódik, az üzletmenet-folytonosság aspektusában egy egyedi, de szintén nem ismeretlen, viszont nem begyakorlott megközelítés húzódik. Ma már léteznek olyan módszertani támogatóeszközök, amelyekkel az üzletmenetre gyakorolt hatások elemzése relatíve eredményesen megvalósítható, de magának a kifejezésnek még nincs önálló definíciója, ami jelentős mértékben nehezíti a kifejezés megértését. Az üzletmenetre gyakorolt hatások elemzése során azokat a működésbeli együttállásokat kell feltárni,

elemezni és értékelni, amelyek az üzleti áramlás tényezőinek a folytonosságára (információ-, pénz- és termék/szolgáltatás) nézve akadályozó hatással lehetnek. Az elemzés során ezeket az együttállásokat a jól ismert projektháromszög (idő, pénz, kibocsájtás) mint lehetséges és számszerűsíthető veszteségek tényezői alapján kell vizsgálni.

4. ábra. Az üzletmenet-folytonosság stratégiai modellje
(A strategic model of business continuity)



Forrás: *Tucker* [2015]: E-governance, e-goverrnment.

Általánosan elmondható, hogy a Covid19-járványidőszak igényt hozott létre világszinten a gazdasági társulások körében az üzletmenetre gyakorolt hatások azonosítása és értékelése iránt, hiszen a még ma is begyűrűző válságjelenségek folyamatos és változatos kihívásokat jelentenek az üzleti áramlások tényezői számára a folytonosság fenntartása tekintetében. A szabvány begyakorlottságának hiánya miatt a szakértők a tapasztalatok segítségével arra keresték és keresik a választ, hogy mik lehetnek azok a tipikus és a leginkább előforduló, az üzletmenetre gyakorolt hatások, amelyek veszélyeztetik egy gazdasági társulás üzleti áramlásait, vagy a biztonságos működését. Az univerzális válasz igénye az aktuális globális egészségügyi és gazdasági válsághelyzetek miatt folyamatos, azt remélve, hogy az egyedszintű, időigényes önvizsgálat rövidíthető és egyszerűsíthető, illetve instant válaszok és gyakorlatok generálhatók általa. Figyelembe véve az eddigi, az üzletmenetre gyakorolt hatásokra jellemző ismérveket, ez a szakértői törekvés

érzékelhetően egy többlépcsős, nagy mintájú és időben elhúzódó kutatás. Azonban létezik megoldás olyan halmaz létrehozására, amely általános érvényűnek mondható. Jó megközelítésnek bizonyul, ha megtörténik azoknak a menedzsmentterületek feltárása, amelyek a gazdasági társulások számára önmagukban kihívást jelentenek. Ezeken a területeken ugyanis nagy valószínűséggel esetleg koncentráltan jelennek meg az üzleti áramlások (pénz, információ, termék/szolgáltatás) zavarait okozó együttállások.

Az ISO 22301:2012 szabvánnyal az International Organization for Standardization (*ISO* [2012]) kiadott egy tool kitet, amelyben egy kérdőíves keretbe foglalt üzletmenetre gyakorolt hatások elemzési megoldást kínált. A kérdőívről hamar kiderült, hogy az általa nyújtott eredmények sok szempontból nem kielégítők, így ennek áthidalására a szabványalkotó kiadta az *ISO/TS 22317:2015 Societal Security – Business Continuity Management Systems – Guidelines For Business Impact Analysis (BIA)* szabványt (*ISO* [2015]). Ezáltal az üzletmenetre gyakorolt hatások köre önálló keretrendszert és értelmezési tartományt kapott a négypilléres vészhelyzetmenedzsment-stratégiában, amely megőrizte az ISO irányítási rendszerek sajátos értelmezési tartományát. A szabvány segítségével az egész üzletmenetfolytonosságmenedzsmentet kizárólag a gazdasági társulás fizikailag, működésileg és hatókörileg lehatárolt egységére értelmezhető. Ebből következik, hogy az üzletmenetre gyakorolt hatások elemzéséhez szükséges értelmezési tartomány is a belső működési funkciókra és az ott rendelkezésre álló erőforrásokra korlátozódik, az egymásra gyakorolt hatások szintjén. Hasonlóan más ISO irányítási rendszerekhez, a működési közeg, vagyis a gazdasági társulás környezete és az azzal való interakciók és kölcsönhatások bizonyos mértékig az üzletmenetre gyakorolt hatások esetében is elválnak a belső rendszertől. A lényegi különbség mégis az, hogy az üzletmenetfolytonosságmenedzsment az erőforrás-gazdálkodás és a működésszabályozás hangsúlya mellett mindezeket egy alternatív vészhelyzeti kontextusban is értelmezi a hatás, a reagálás és a helyreállítás hármasságában. Vagyis azért nincs az üzletmenetre gyakorolt hatás mögé definíció rendelve, mert az bármi lehet, ami az elemzés módszertanából kijön (*Okolita* [2010], *Bowman* [2008]). Az üzletmenetre gyakorolt hatások elemzése során első lépésként az üzletmenet-folytonosságot érintő kulcsterületek (lényegében valamennyi a szervezetben jelenlevő működési funkció) és a hozzájuk rendelt erőforrásellátottság-keresztmetszetet kell jól meghatározni. Ezt követően ezeket a kapcsolódásokat kell helyettesíthetőség, reakcióidő, rendelkezésre bocsájtható alternatív erőforrások, biztosítások, beépített védelmi rendszerek, esetleg presztízs szempontjából számszerűen vizsgálni. A terület–erőforrás keresztmetszet szemléltetésében az NHS Organizations in Scotland (a skót egészségügyi kormányhivatal) által értelmezett 5P-megközelítés segít, amelyet az 2. táblázat szemléltet.

2. táblázat

Egyszerűsített üzletmenetre gyakorolt hatások elemzése
(Analysis of the effects on simplified business operations)

PEOPLE	PREMISES	PROCESSES	PROVIDERS	PROFILE
Key Staff: What staff do you require to carry out your key functions?	**Buildings:** What locations do your department's key functions operate from? (Primary site, alternative premises)	**IT:** What IT is essential to carry out your key functions?	**Reciprocal Arrangements:** Do you have any reciprocal agreements with other organisations?	**Reputation:** Who are your key stakeholders?
Skills/ Expertise/ Training: What skills/ level of expertise is required to undertake key functions?	**Facilities:** What facilities are essential to carry out your key functions?	**Documentation:** What documentation/ records are essential to carry out your key functions, and how are these stored?	**Contractors/ External Providers:** Do you tender critical services out to another organisation, to whom and for what?	**Legal Considerations:** What are your legal, statutory and regulatory requirements?
Minimum Staffing Levels: What is the minimum staffing level with which you could provide some sort of service?	**Equipment / Resources:** What equipment / resources are required to carry out your key functions?	**Systems & Communications** What systems and means of communication are required to carry out your key functions?	**Suppliers:** Who are your priority suppliers and whom do you depend on to undertake your key functions?	**Vulnerable Groups:** Which vulnerable groups might be affected by failing to carry out key functions?

Forrás: *NHS Organisations in Scotland* [n. a.].

A lehetséges üzletmenetre gyakorolt hatások kockázatként történő értelmezése 2 dimenzió alapján történik: a hatás (mint veszély vagy veszteség érintettsége) és a súlyosság (jelentőség), amelyre magyarázó táblázatok szolgáltatnak támpontokat. Az egyszerűsített elemzés az alábbi üzletmenetre gyakorolt hatásokat vizsgálja:

– a jelenség vagy tünet súlyosságát (patient experience);
– a megjelenés érintettségét (objectives/project);
– a humán erőforrás érintettségét (injury /physical and psychological/ to patient/visitor/staff);
– a követelmények megfelelőségének érintettségét (complaints claims);
– az üzemi folytonosság érintettségét (service/business interruption);
– az erőforrások érintettségét (staffing and competence);
– a pénzügyi érintettséget (financial [including damage/loss/ fraud]);
– az észlelhetőséget (inspection/audit);
– a jó hírnév érintettségét (adverse publicity/reputation).

Annak ellenére, hogy az NHS a kormányzati szektorhoz tartozik, az önálló vagyongazdálkodás okán a gazdasági társulások körébe sorolható. Többek között ennek köszönhetően is az üzletmenetfolytonosság-menedzsment szakértői a kormányzati szektort szintén gazdasági tárulásként értelmezik, aminek egyik legfőbb ismérve jogszabályban megfogalmazott útmutatás. Magyarországon 2021. június 13-án összesen 42, üzletmenet-folytonossággal összefüggő jogszabály létezett, ezek jellemzően az információbiztonsághoz kötődnek szorosan.

3. Problémafelvetés, a kutatás tárgya

Az üzletmenetre gyakorolt hatások köre a hasznosság és a hatékonyság növelése érdekében modellértékű determinációt igényel. A szakértői kör – különös tekintettel a Covid19-járványidőszak tapasztalataiból merítve – azt vizsgálja, hogy az elemzés módszertana hogyan viselkedik, és milyen eredményeket azonosít a különböző gazdasági társulások körében. Vagyis a világ nem azonosított olyan, modellértékűen tipikus, a többségre jellemző üzletmenetre gyakorolt hatásokat vagy működési területeket, funkciókat, amelyek a jelenkorra értelmezve egy tudatosabb menedzsmenttel egy jobban célzó elemzést tehetnének lehetővé.

2017-ben, amikor a kutatás elindult, a témával kapcsolatos tudásanyag elérhetősége Magyarországon nehezebb volt. Az elérhető tudásanyagból és szakirodalomból azt az általános következtetést lehetett levonni, hogy az üzletmenetre gyakorolt hatások köre felhasználónként, szakértőként, esetleg vállalati profilonként vagy alkalmazhatóság szempontjából változó és eltérő. A későbbiekben bemutatandó kutatás ennek a sokrétűségnek köszönhetően azt tűzte ki kezdeti célnak, hogy a menedzsmentterület-publikációból szövegelemzéssel megvizsgálja azokat a kulcsténye zőket, kulcsterületek, funkciókat, vagy erőforrásokat, értékeket vagy kompetenciákat, amelyek nagy valószínűséggel alkotói vagy fókuszpontjai lehetnek az üzletmenetre gyakorolt hatások körének. A kutatás elvi megközelítése a következő. Minden ismert menedzsmentmodellt elhanyagolva, kizárólag a tudományos „kollektív tudatból” azt kinyerni, ami a tudományos világot menedzsmentszempontból a leginkább foglalkoztatja. Azzal a feltételezéssel élünk, hogy a szövegelemzés adta eredmények megmutatják azokat a területeket, amelyekkel a legtöbben foglalkoznak, vagyis amelyek vállalati környezetben aktuálisan fejlődésre, esetlegesen megoldásra szorulnak.

4. A szövegelemzés előkészítése és a megfelelő modell kiválasztása

Annak érdekében, hogy valóban a kollektív tudatból lehessen kinyerni eredményeket, a topikmodell-eljárások közül *Bley* LDA-módszerével lett vizsgálva 178 szakértői cikk. Ezek – együttes néven korpusz – általános jellemzői:

– a cikkek szókereséssel lettek kiemelve, ami esetenként szinonimákkal is élt;

– a Science Directen megjelent cikkek lettek legyűjtve az egységes formátum érdekében;

– a szókeresés eredményeképpen egy 374 cikkből álló listából történt válogatás, aminek elsődleges szempontja az volt, hogy a cikkek kutatási eredményekről számoljanak be.

A kutatás R-stúdióban, egész pontosan a textmineR-csomag topikmodell parancssora lett használva (https://www.rtextminer.com/articles/c topic modeling.html).

A szövegtisztítás lépései:

1. Valamennyi cikkből kizárólag az értelmi rész maradt, tehát az absztrakt, az irodalomjegyzék, valamint a publikáció metaadatai el lettek távolítva.

2. A szövegtisztítás írásjelekre, számokra, betűméretre és egyéb karakterkövetelményekre tökéletesen hozta az eredményeket, azonban a stopszavak esetében további manuális bővítés bizonyult szükségesnek. Mivel a nemzetközi szinten megjelenő tudományos cikkek többségében angolul íródnak, de a szerzők számos nemzetből kerülnek ki, a beépített unicode funkció nem tudta hozni az eredményt. Ennek oka az volt, hogy 2020. márciusban világszintre kiterjesztett unicode-változtatás történt, ezért a megfelelő unicode kiválasztása külön lépést igényelt.

3. További tisztításként kikerült az összes olyan szó, amelyik tízszer vagy annál kevesebbszer fordul elő a teljes korpuszban.

4. Végül a kutatás szempontjából zavaró, de értelmes kifejezések el lettek távolítva manuálisan (case study, literature, review, science, procedia, harvard, cc, aalst, abic, abstract, conclusions, introduction, studies, references, contents, lists, revised, received, łódzkie, voivodeship), mert a topikmodell-eredményekben megjelentek, torzítva a korpuszmenedzsment-fókuszt.

5. A megfelelő topikmodell kiválasztása

A topikmodell-eljárással való munka során nincsenek egzakt szabályok a megfelelő modell kiválasztásához (*Jones* [2014]), hiszen szöveges eredményeket ad ki, így a kutatónak lehetősége van arra, hogy a kutatás szempontjából az általa leginkább elfogadható modell segítségével haladjon tovább. Valamennyi statisztikai mutatóról és ábráról el lehet mondani, hogy inkább tájékoztató erejűek a gyakorlatban. A megismételhetőség érdekében érdemes egy kétlépcsős vizsgálati és döntési megközelítést használni. Az első lépcsőben a topikmodell textmineR-be épített eredményei és kimenetei alapján történik a döntés, a második lépcsőben hiperparaméter-optimalizálás során változott eredmények alapján, egzakt ökölszabályok segítségével.

Az *első lépcső* szerinti értékelésben a topikmodell-parancs mögé beépített kimenetek követelményei a következők:

– *A modell jóságára vonatkozó* R^2: a topikmodell esetében el lehet tekinteni a numerikus alapú valószínűségszámításban ismert modell jóságára vonatkozó szabályoktól. Topikmodell esetében a 0,25-os érték ugyanolyan erősnek tekinthető, mint a numerikus adatok esetében a 0,5 (*Jones* [2019]) – tájékoztató információ.

– *Iterációs görbe*: log-likelihood számítással ideális esetben folytonos és szabályos logaritmikus függvény. Ennek segítségével az optimális iterációs szám is láthatóvá válhat – tájékoztató információ.

– *Valószínűségikoherencia-hisztogram*: az egyes topikok koherenciaeloszlását szemlélteti a gyakorisághisztogram. Vagyis azt ábrázolja, hogy adott koherenciatartományokon belül hány topik található. Ebben az esetben a koherenciaértékek között nem cél a normál eloszlás, sem bármilyen regressziós képesség. A mínusz előjeles koherenciaérték úgynevezett „junk" topikra utal, amibe olyan kifejezések kerültek, amelyek egymással nincsenek kapcsolatban, így együttesen nem egy téma köré csoportosulnak, hanem inkább kilógó értékként értelmezhetők. Mindettől független a textmineR felcímkézi a junk topikokat is, és ha a kutatás értelmezési tartományába illeszkedik a címke elnevezés, akkor a modell felhasználható kutatásra. A hisztogramtól nem elvárás az sem, hogy folytonos legyen, és minden egyes koherenciatartományba kerüljön topik – tájékoztató információ.

– *Súlyosság- vagy sűrűségmátrix*: az alfaértékkel vizsgálva azt mutatja meg, hogy egy topikban mennyi kifejezés került. A sűrűségmátrix két dimenzióban mutatja meg azt a sokdimenziós teret, ahol a korpusz működik, vagyis az, hogy két topik azonos sűrűségű és a mát-

rixban egymáson látható, nem jelenti azt, hogy egyformák. Ahhoz, hogy az egyezőséget ki lehessen zárni, a topszavak áttekintésére van szükség. Ha két topik egymásba ér, a topszavak egy része megegyezhet. Ebben az esetben nem ajánlott a modellt felhasználni, mert egy információra két topik mutat – tájékoztató információ.

– *Címkék elnevezése*: a textmineR-be a bigramok (kétszavas kifejezések) segítségével egy naivtopik-címkézés van beépítve, azaz minden egyes topiknak képes címkét adni (ezért kaphat a junk topik is címkét). Azt azonban nem zárja ki a rendszer, hogy két különböző topik ugyanazt a címkét kapja. Két, egyező címkével rendelkező topik esetében szintén a topszavak segítenek új, vagy jobban közelítő címkét adni a topiknak, ami egyértelműen és kizárólag a kutató szubjektív látásmódján és a kutatáskontextusán alapul. Tehát a kutatás érdekében történő átcímkézés elfogadott és bevált gyakorlat, így a címkék elnevezése is tájékoztató információ.

– *Topkifejezések*: a topkifejezések kiszámítása egyértelműen az adott topikban megjelenő szógyakoriság mentén történik. A topkifejezések együttesen utalnak arra a témára, amelyik topik alá bekerültek, de nem feltétlen adják ki a címkeelnevezést. A címke elnevezése a topkifejezések segítségével változtatható meg. A topkifejezések száma szintén szubjektív, a kutató dönti el, hogy az első hány kifejezést tekinti topnak, ezért ez döntést elősegítő információ.

– *Technikai beállítások és verzióváltások*: az R-ben több lehetőség is van a működésre vagy az adat értelmezésére, illetve az R-érzékeny az egyéb érintett szoftverek változásaira is, mint például egy vírusvédelem, vagy egy Windows-frissítés (ez esetben a korábbi verzió állapota nem hívható vissza). A kutatás megismételhetősége és reprodukálhatósága érdekében ezeket a beállításokat a modellalkotás során folytonosan és állandóan érdemes tartani.

Mindezek alapján belátható, hogy valamennyi eredmény annyira tájékoztató erejű, amennyire az a kutató számára elfogadható, és egyiknek sincs önállóan döntési ereje, de ezeknek az eredményeknek az együttes értelmezése sem segít egyértelműen a megfelelő modell kiválasztásában. Ennek a szubjektív megközelítésnek az elkerülése érdekében válik szükségessé a *második lépcsős* döntési szempontrendszer a koherenciaértékekre helyezve a hangsúlyt.[2] Ennek megfelelően az alábbi szempontok szigorú sorrendje szerint meg lehet állapítani a megfelelő modellt:

[2] *Varjú Zoltán*nal (Complytron Kft.) folytatott konzultációk, 2020

1. Nincs a topikmodellben 0 alatti koherenciaérték, vagyis mindegyik topik koherenciaértéke nagyobb, mint 0. Ez azt jelenti, hogy nincs olyan topik, amelyben kilógó kifejezések gyűltek össze.

2. Amelyik modell esetében – szintén a koherenciaértékeket vizsgálva – a

$$\frac{Medián–Min}{Terjedelem} \cong 0{,}5$$ a legközelebb teljesül.

3. Az iterációs görbe folytonos, szabályos logaritmikus függvény. Az iterációs görbéket egy 1–5-ig terjedő skálán értékeltem szubjektíven, ahol 5 a legjobb, 1 a legrosszabb értéket jelenti. Ökölszabálynak tekintettem, hogy amelyik iterációs görbében hullám látható, az egyértelműen 1-es értéket kap.

4. A prevalenciamátrixban a topikok nem úgy szóródnak, hogy betöltsék a teljes teret, az látható, hogy valamerre tartanak. Ez arra utal, hogy a topikoknak van egy közös, de nem ismert metszete.

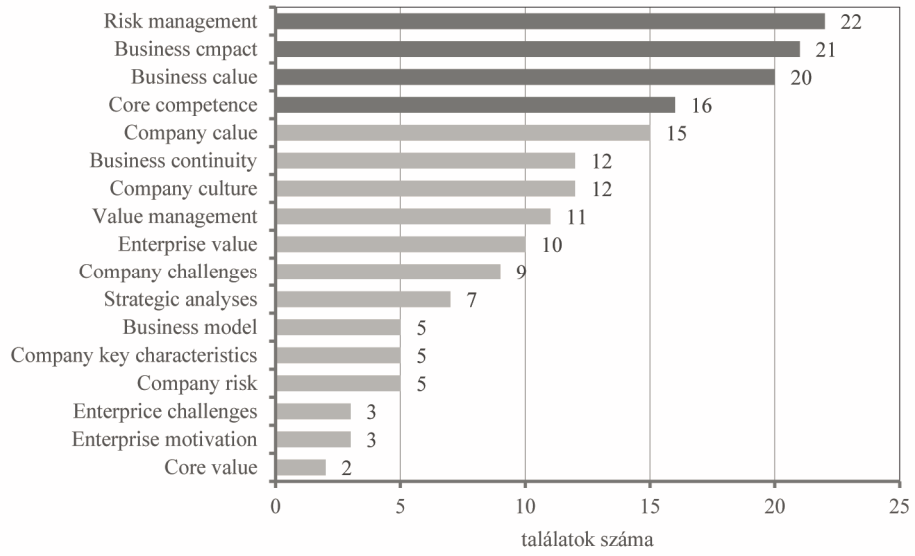
5. A koherenciatartomány legyen a legnagyobb a többi modell koherenciatartományához képest, úgy, hogy az első 4 feltétel teljesült.

Ennek az 5 szempontnak a mentén kutatói szubjektivitás nélkül választható ki az a megfelelő modell, amellyel a döntési mechanizmus reprodukálhatóvá válik.

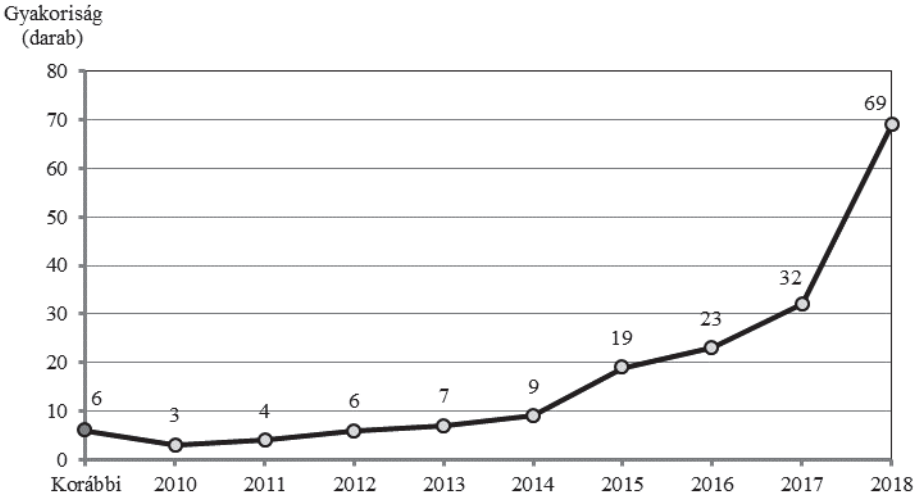
6. Néhány eredmény a korpuszt illetően

A korpuszba 17 kulcskifejezés alapján lettek szűrve a cikkek, amelyek között szükségesnek látszott a szinonimák alkalmazása. A 178 cikkre nézve az 5. ábrán látható, hogy a top 3 találat között szerencsére megtalálható az üzletmenetre gyakorolt hatás, valamint az üzleti érték, amelyek elsődlegesen érintik a kutatást és a témát.

5. ábra. A korpusz kulcsszavas (tételes) keresés szerinti összetétele
(Corpus composition by keyword search [itemized])

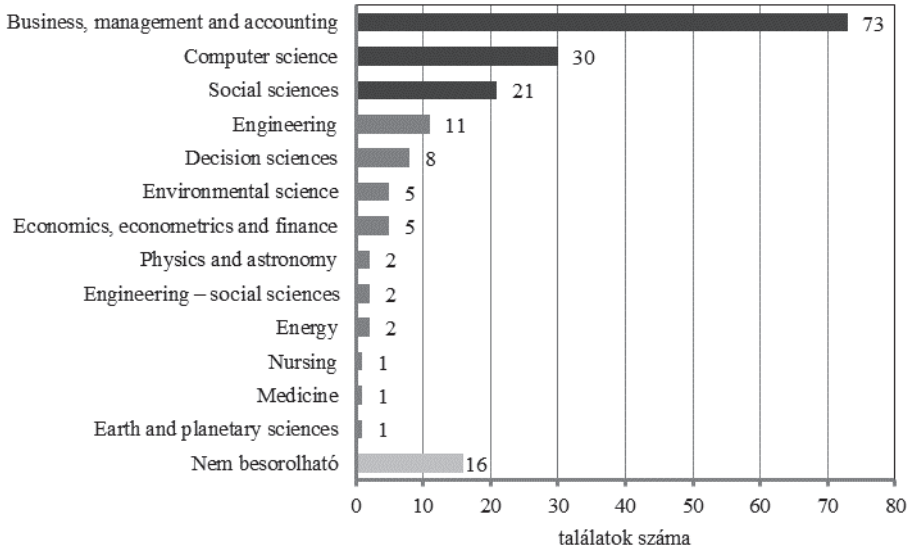


6. ábra. A korpuszt alkotó cikkek megjelenésének időrendi gyakorisága
(The chronological frequency of the articles that make up the corpus)



A korpusz szerencsésnek mondható a cikkek megjelenését tekintve, mert a 6. ábra mutatja azt, hogy a régebbi cikkekből kevesebb került be, vagyis 8 év távlatában jeleníti meg a menedzsmentterületek egyre növekvő jelentőségét.

7. ábra. A korpusz SC1 tematikus rangsora
(Thematic ranking of the corpus SC1)



A folyóiratok szerinti tematikus besorolás is jelzi a korpusz összetételének a megfelelőségét, ugyanis a 7. ábrán látható Business, management and accounting tökéletes tematikája az üzletmenetfolytonosság-menedzsmentnek. Ami a leíró statisztikákat illeti, a kiválasztott topikmodellhez tartozó korpuszra vonatkozó információk is még a korpusz megfelelőségére utalnak a következők szerint:

– 178 cikkhez maradt 840 640 egyszavas és 835 014 kétszavas kifejezés.
– A legrövidebb cikk 735 szavas.
– A leghosszabb cikk 21 072 szavas.
– Átlagos cikkhossz: 8 890 szó.
– A medián szerinti cikkhossz: 8 626 szó.

A 3. és a 4. táblázat tartalmazza a top egy- és kétszavas kifejezések rangsorát és adatait.

3. táblázat

Top 10 egyszavas kifejezések gyakorisága
(Top 10 one-word-phrase frequency)

Kummulált gyakoriság				Dokumentumalapú szógyakoriság			
term (kifejezés)	*term_ freq*	*doc_ freq*	*idf*	term (kifejezés)	*term_ freq*	*doc_ freq*	*idf*
business (üzlet)	8 563	166	0,069796	based (alapozott)	1 734	175	0,016998
risk (kockázat)	5 798	122	0,377763	research (kutatás)	3 361	174	0,022728
management (irányítás)	5 032	172	0,034289	management (irányítás)	5 032	172	0,034289
research (kutatás)	3 361	174	0,022728	business (üzlet)	8 563	166	0,069796
model (modell)	3 291	158	0,119189	important (fontos)	1 154	165	0,075838
data (adat)	2 680	155	0,138358	level (szint)	1 309	165	0,075838
social (társadalom)	2 545	134	0,283944	analysis (elemzés)	1 791	165	0,075838
information (információ)	2 387	161	0,100379	development (fejlesztés)	1 396	163	0,088033
companies (vállalatok)	2 300	136	0,269129	online (online)	394	162	0,094187

A *term freq* azt mutatja meg, hogy az adott kifejezés a tejes korpuszban hányszor jelenik meg.

– A *doc freq* azt mutatja meg, hogy az adott kifejezés hány cikkben fordul elő.

– Az *idf*, vagy más néven *invers term frequency* egy súlyozott érték, a kifejezés jelentőségét számolja, annak ellenére, hogy a sima gyakorisági érték esetleg alacsony.

A tapasztalat alapján elmondható, hogy az *idf* értéke nem megfelelő szövegtisztítás és metaadat-kezelés következtében képes olyan kifejezések, betűkombinációk mellett magas értéket felvenni, aminek semmi értelme. Ezért, bár logikusnak tűnik az *idf*-értékkel dolgozni, az eredmények tükrében érdemes volt ezt az értéket figyelmen kívül hagyni.

4. táblázat

Top 10 kétszavas kifejezések gyakorisága
(Top 10 two-word-phrase frequency)

Kummulált gyakoriság				Dokumentumalapú szógyakoriság			
term (kifejezés)	*term_ freq*	*doc_ freq*	*idf*	term (kifejezés)	*term_ freq*	*doc_ freq*	*idf*
future_ research	147	77	0,837978	risk_ management	1 125	48	1,310583
data_ collection	100	59	1,104246	business_ model	1 062	39	1,518222
competitive_ advantage	180	58	1,121341	supply_ chain	712	56	1,156432
decision_ making	105	57	1,138732	business_ models	569	39	1,518222
supply_ chain	712	56	1,156432	social_ media	502	10	2,879198
business_ business	135	55	1,174450	big_data	293	17	2,348570
strategic_ management	165	54	1,192800	internatio- nal_business	280	30	1,780586
international_ conference	127	53	1,211492	organizati- onal_culture	276	23	2,046289
table_ shows	118	53	1,211492	core_ competence	257	12	2,696877
inter- nal_external	111	53	1,211492	information_ systems	271	43	1,420583

7. A topikmodell eredményei

A topikmodell-eljárás nem hosszú parancssora az előkészítő műveletek után a 8. ábrán látható. A megfelelő modellhez összesen 3 × 63 = 189 kísérlet zajlott le: három szövegtisztítási eljárásra volt szükség a korpuszon, 63 = 10-től 30-ig = 21 topikra, 250, 500, és 750 iterációs számra lettek futtatva a kísérletek.

A megfelelő topikmodell kiválasztásához megadott lehetőségek száma az alábbiak szerint lett meghatározva:

1. A topikok száma (beállítástól függően) 10-től 30-ig terjedt, aminek az eredménye összesen 21 lehetséges topikmodell akkor, ha egy, azonos iterációs számon van futtatva a program.

2. Mind a 21 lehetséges topikmodell esetére 3 iterációs szám lett beállítva, annak érdekében, hogy az így kapott topikmodell-mennyiség biztos, hogy megfelelő mennyiségű legyen az összehasonlítás és a megfelelő modell kiválasztása érdekében. A megadott iterációs számok a 250, az 500 és a 750. Így a 21x3 beállítás eredménye 63 lehetséges topikmodell, ami már kellően biztonságos alapot képes adni a megfelelő modell kiválasztásához.

3. Egy körös futtatás esetén tehát 63 lehetséges modellt jelent a topikszámok és az iterációs számok variálása. Több körös futtatásra akkor van szükség, ha a topikok alatti top kifejezések további szövegtisztítás vagy egyéb beállítás szükségességére mutatnak rá.

A kiválasztott topikmodell leíró statisztikája az 5. táblázatban látható, a 0,28-os R^2-érték szövegelemzés esetén jónak számít.

8. ábra. A választott topikmodell parancssora R-ben
(Script for the selected topic model in R.)

```
set.seed(12345)

model <- FitLdaModel(dtm = dtm,
                     k = 19,
                     iterations = 250,
                     burnin = 180,
                     alpha = 0.1,
                     beta = 0.05,
                     optimize_alpha = TRUE,
                     calc_likelihood = TRUE,
                     calc_coherence = TRUE,
                     calc_r2 = TRUE,
                     cpus = 4)
```

5. táblázat

A választott modellre jellemző leíró statisztika
(Descriptive statistics specific to the chosen model)

k	19
iter	250
R^2	0,2847814
Min	0,0088890
1st,Qu,	0,0241470
Median	0,0304730
Mean	0,0309810
3rdQu,	0,0846700
Max	0,1373510

A topikmodell eredménye a 6. táblázatban látható, a topik számával, a címkéjével – ami a topikmodell-eljárás egyik legnagyobb előnye –, a koherencia- és a prevalenciaértékekkel. A topikok száma a koherenciaértékek sorba rendezésével generálódik, a legkisebb koherenciaértékű topikkal kezdve.

6. táblázat

A választott modell címkéi
(Labels for the selected model)

Téma	Címke	Koherencia	Elterjedtség	Téma	Címke	Koherencia	Elterjedtség
*t*1	business_continuity	0,13735	6,21549	*t*11	risk_management	0,08354	5,44366
*t*2	human_resource	0,02246	5,66995	*t*12	intellectual_capital	0,03638	8,02514
*t*3	international_business	0,06047	3,81966	*t*13	international_business	0,01103	4,91318
*t*4	business_risk	0,07635	4,23966	*t*14	core_competence	0,02293	9,31122
*t*5	business_model	0,10652	5,04932	*t*15	business_education	0,06828	4,90150
*t*6	organizational_culture	0,02459	5,07255	*t*16	renewable_energy	0,07543	5,44473
*t*7	enterprise_risk	0,02989	5,64945	*t*17	risk_factors	0,00889	3,79179
*t*8	big_data	0,12587	4,52991	*t*18	risk_management	0,08580	4,13105
*t*9	big_data	0,02370	3,90830	*t*19	supply_chain	0,11230	4,85555
*t*10	social_media	0,04686	5,02789				

A táblázatban látható, hogy az algoritmusok ismétlést hajtottak végre 3 topik esetében, annak ellenére, hogy a topikokba csoportosított kifejezések teljesen eltérők. A topikmodell ebben az esetben ugyanúgy működik, mint a klaszterező eljárások, vagyis a csoportok címkéje, azaz értelmezése egyedileg (újra)elnevezhető. Ezt a lehetőséget kihasználva a 7. táblázat szerint lettek megváltoztatva a *t*3, a *t*8 és a *t*18 topikok címkéi.

7. táblázat

A választott modell átcímkézve
(The selected model has been relabelled)

Téma	Címke	Koherencia	Elterjedtség	Téma	Címke	Koherencia	Elterjedtség
*t*1	business_continuity	0,13735	6,21549	*t*11	risk_management	0,08354	5,44366
*t*2	human_resource	0,02246	5,66995	*t*12	intellectual_capital	0,03638	8,02514
*t*3	business_network	0,06047	3,81966	*t*13	international_business	0,01103	4,91318
*t*4	business_risk	0,07635	4,23966	*t*14	core_competence	0,02293	9,31122
*t*5	business_model	0,10652	5,04932	*t*15	business_education	0,06828	4,90150
*t*6	organizational_culture	0,02459	5,07255	*t*16	renewable_energy	0,07543	5,44473
*t*7	enterprise_risk	0,02989	5,64945	*t*17	risk_factors	0,00889	3,79179
*t*8	market_positioning	0,12587	4,52991	*t*18	business_aptness	0,08580	4,13105
*t*9	big_data	0,02370	3,90830	*t*19	supply_chain	0,11230	4,85555
*t*10	social_media	0,04686	5,02789				

Megjegyzés: A 6. táblázatban szürke háttérrel jelölve a legmagasabb koherencia- (business_continuity) és a prevalencia- (core_competence) értékű topikok. A legmagasabb koherenciaérték azt jelzi, hogy az ez alá a topik alá csoportosított kifejezések ebben a legkoherensebbek egymással. A legerősebb prevalenciaérték azt jelzi, hogy ebben a topikban található a legtöbb kifejezés, vagyis ez a legsűrűbb.

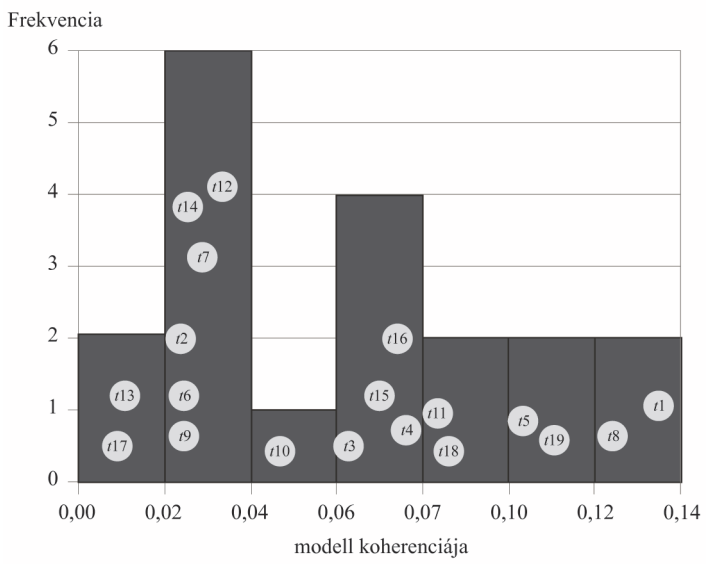
A 9. ábrán látható koherencia-hisztogram értéktartománya a többi topikmodellhez képest is megfelelőnek mondható. A topikok a koherenciaértékek alapján az alábbiak szerint állnak közel egymáshoz:

0,00–0,02: risk_factors (*t*17), international_business (*t*13);
0,02–0,04: human_resource (*t*2), organization_culture (*t*6), enterprise_risk (*t*7), big_data (*t9*), intellectual_capital (*t*12), core_competence (*t*14);
0,04–0,06: social_media (*t*10);
0,06–0,08: business_network (*t*3), business_risk (*t*4), business_education (*t*15), renewable_energy (*t*16);

0,08–0,10: risk_management (*t*11), business aptness (*t*18);
0,10–0,12: business_model (*t*5), supply_chain (*t*19);
0,12–0,14: business_continuity (*t*1), market_positioning (*t*8);

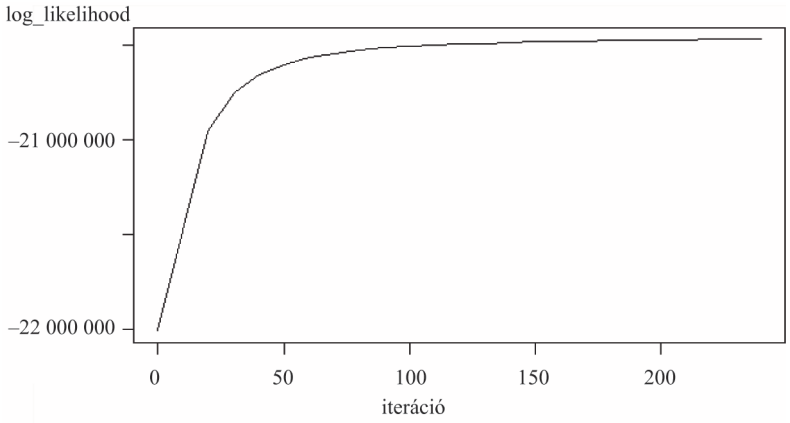
Ez az értelmezés tekintetében nem jelenti azt, hogy a hasonló koherenciaértékekkel rendelkező topikok kizárólag a koherenciaértékek alapján bármilyen értelemben is összevonhatók, ám – különösen az egyes topikokat topkifejezésekkel együtt vizsgálva – értelmezhetővé válik a korpusz korlátossága, leginkább a kontextusokat, illetve a látens változót illetően. Az alacsony koherenciaértékkel rendelkező topikok alacsonyabb topikszám esetében feltehetően beolvadnának más, erősebb topikokba.

9. ábra. A kiválasztott topikmodell koherencia-hisztogramja
(Coherence histogram of the selected topic model)



k	19
iter	250
*t*1	0,1373500
*t*2	0,0224600
*t*3	0,0604700
*t*4	0,0763500
*t*5	0,1065200
*t*6	0,0245900
*t*7	0,0298900
*t*8	0,1258700
*t*9	0,0237000
*t*10	0,0468600
*t*11	0,0835400
*t*12	0,0363800
*t*13	0,0110300
*t*14	0,0229300
*t*15	0,0682800
*t*16	0,0754300
*t*17	0,0088900
*t*18	0,0858000
*t*19	0,1123000

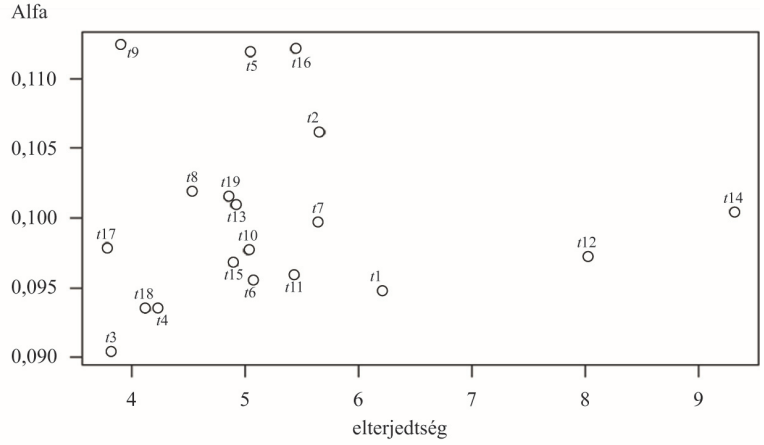
10. ábra. A kiválasztott topikmodell iterációs görbéje
(Iteration curve of the selected topic model)



Ugyan a 10. ábrán az iterációs görbében van két enyhe törés, azoktól eltekintve a függvény a logaritmikus alakot jól tartja, és hullámmentes. Ennek megfelelően az iterációs görbe is teljesíti az elvárást. A 19 topik esetében elmondható, hogy minél magasabb az iterációs szám, annál meredekebb az iterációs görbe emelkedő ága.

A 11. ábrán látható prevalenciamátrix, szintén teljesíti az elvárásokat, mivel a topikok a tér egyik pontján se fedik egymást, és nem is érnek egymásba – vagyis nem tartalmaznak közös kifejezéseket, elmondható, hogy függetlenek egymástól. Mégis, látható, hogy két topik kivételével mindegyik a tér bal oldali része felé tart, amiből arra lehet következtetni, hogy a topikoknak együttesen is létezik egy közös mondanivalója. A két kilógó topik pedig nem bontja meg az egységet, mert a távolság a többséghez képest nem tekinthető nagynak. A topikok erősségére, súlyára vonatkozó értékeket az ábra melletti táblázat tartalmazza.

11. ábra. A kiválasztott modell prevalenciamátrixa
(The prevalence matrix of the selected model)



k	19
iter	250
*t*1	6,2154900
*t*2	5,6699500
*t*3	3,8196600
*t*4	4,2396600
*t*5	5,0493200
*t*6	5,0725500
*t*7	5,6494500
*t*8	4,5299100
*t*9	3,9083000
*t*10	5,0278900
*t*11	5,4436600
*t*12	8,0251400
*t*13	4,9131800
*t*14	9,3112200
*t*15	4,9015000
*t*16	5,4447300
*t*17	3,7917900
*t*18	4,1310500
*t*19	4,8555500

8. Az eredmények értelmezése

A topikmodell-eljárás segítségével 19 olyan területet lehetett azonosítani a menedzsment-szakirodalomból, ahol az üzletmenet-folytonosságot érintő negatív hatások koncentráltan, vagy nagy valószínűséggel megjelenhetnek. Ez két megállapítást tesz lehetővé. Az egyik, hogy a topikmodell alkalmas a menedzsment-fókuszterületek megállapítására. A másik, hogy további fejlesztéseket érdemes ezeken a szervezeti területeken, funkciókban hangsúlyosan végezni, mert a nemzetközi menedzsmentszakértői kör szerint ezek azok a menedzsmentterületek, ahol a legtöbb kihívás megjelenik.

12. ábra. A gazdasági társulások kulcstényezői
(Key factors in economic associations)

Az eredmények bemutatásának zárásaként érdemes megnézni azt a lehetőséget, amikor a 19 topik tovább van szűkítve, immár kutatói szubjektivitás alapján. A sikeres gazdasági társulás a jelenkor kihívásai mellett 4 bázispontra emelhető, amit a 12. ábra és a 8. táblázat szemléltet. Ez a 4 bázis egyfajta evolúciót is jelent, miszerint a III. ipari forradalomból tovább lépve, a digitalizációs forradalom korszakának megfelelően a hangsúly átkerült egy magasabb perspektívára és egy tágabb gondolkodásmódra.

8. táblázat

A gazdasági társulások kulcstényezői a kritikus területek felosztásával
(The key factors in economic associations are the division of critical areas)

Kockázatok	Hálózatok	Humán tényezők	Üzleti környezet
risk management	big data	human resources	market pozitioning
risk factors	supply chain	intellectual capital	business model
business risk	international business	organization culture	business aptness
enterprise risk	business network	core competence	business continuity
	social media		business education

9. Konklúzió

A topikmodell-eljárásról összességében elmondható, hogy alkalmas a menedzsmenttudomány fókuszkérdéseinek a vizsgálatára is. Blei LDA-módszere a szakirodalom alapján is egy olyan generikus módszerré vált, amely jó kiindulópont a származtatott topikmodell-eljárások kifejlesztésére, stabilitás, egyszerűsítés, vagy egyéb statisztikai eljárásokkal való integráció szempontjából. Ugyanakkor a kutatótól és a kutatás kérdéseitől függően az alapmódszer is képes a menedzsmentkérdések és -problémák előremozdítását elősegíteni.

Ahhoz, hogy a korpusz valóban megismerhető legyen a kutató számára, mind a hiperparaméter-optimalizálás, mind az iterációs kísérletek szükségesek, vagyis némileg időigényes a valóban jó modell megállapítása, de a kutatót magabiztossá teszi az eredményeket illetően. Az iterációknak köszönhetően a kutató számára beláthatóvá válik a korpusz azon korlátja, amely a szavakból ered. Ez annyit jelent, hogy különböző iterációs számok és egyéb paraméter-beállítások mellett is átláthatóvá válik a korpuszban koncentrált legfőbb mondanivaló, és a kutatón csak az múlik, hogy ezt mennyire részletezve kívánja megjeleníteni, vagyis hány topikban részletezve tartja a kutatás szempontjából kielégítőnek a lényeget. Egyértelmű összefüggés van a szöveg tisztítás mértéke, az iterációs szám, a topikok száma, és a kísérletek lefutási ideje között.

Ami a menedzsmenteredményeket illeti, a menedzsmentszakértők publikációiból azonosított területek és rendszerelemek a gyakorlati tendenciákhoz viszonyítva elfogadhatónak tekinthetők. A teljes kutatásban sem a koherencia-, sem a prevalenciaértékek tekintetében nem történt különbségtétel, vagy hierarchikus összerendezés. Ennek oka az, hogy teljesen valószerű a mai globális tendenciák figyelembevétele mellett ez a 19 kritikus terület, vagyis feltételezhető, hogy ezeken a menedzsmentterületeken nagy valószínűséggel és akár koncentráltan azonosíthatók az üzleti áramlásokat akadályozó együttállások.

Megállapítható, hogy a topikmodell alkalmazhatóságát tekintve jelentős előrelépésnek tekinthető a menedzsment szakértők számára, hiszen a módszer képes arra a valósághű összegzésre, amelyet már egyéb alkalmazási területeken bizonyítottak. A topikmodell segítségével történő tudományos összegzések várható előnyei többek között az alábbiak lehetnek:

– a legkritikusabb és legproblémásabb területek meghatározásában segít;

– hozzájárul azon fókuszterületek azonosításához, amelyek magának a tudományterületnek a fejlődési irányait jelölik ki;

– hasonlóságokat, különbségeket, esetleg közös metszeteket segít feltárni több tudományterület között, vagyis szinergiát és átjárhatóságot teremt, ami egy további elmozdulás a holisztikus működési szemlélet felé.

A topikmodell-eljárás a megfelelő körültekintéssel képes olyan irányokat megmutatni, amelyek az adott tudományterületen előremutatók lehetnek.

Irodalom

ASHISH, K. – PAUL, A. [2016]: *Mastering Text Mining with R*. Packt Publishing Ltd. Birmingham.

BALOGH K. [2015]: *A látens Dirichlet-allokáció társadalomtudományi alkalmazása A kuruc.info romaellenes megnyilvánulásainak tematikus elemzése*. Szakdolgozat. Eötvös Loránd Tudományegyetem. Budapest.

BARDE, B. V. – BAINWAD, A. M. [2017]: An overview of topic modeling methods and tools. In: *Institute of Electrical and Electronics Engineers*: *2017 International Conference on Intelligent Computing and Control Systems (ICICCSI)*. Piscataway. pp. 745–750. https://doi.org/10.1109/ICCONS.2017.8250563

BLEI, D. M. – NG, A. Y. – JORDAN, M. I. [2003]: Latent Dirichlet allocation. *Journal of Machine Learning Research*. Vol. 3. January. pp. 993–1022.

BLEI, D. M. – LAFFERTY, J. D. [2006]: Dynamic topic models. In: *Cohen, W. W. – Moore, A. W.* (eds.): *Proceedings of the 23rd International Conference on Machine Learning – Pittsburgh, Pennsylvania, June 25–29, 2006*. Association for Computing Machinery, International Machine Learning Society. New York.

BLEI, D. M. – LAFFERTY, J. D. [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

BLEI, D. M. – LAFFERTY, J. D. [2009]: Topic models. In: *Srivastava, A. N. – Sahami, M.* (eds.): *Text Mining: Classification, Clustering, and Applications*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. Taylor and Francis Group. London.

BLEI, D. M. [2010]: *Introduction to Probabilistic Topic Models.* Semantic Scolar Computer Science. https://www.semanticscholar.org/paper/Introduction-to-Probabilistic-Topic-Models-Blei/5f1038ad42ed8a4428e395c96d57f83d201ef3b3

BLEI, D. M. – CARIN, L. – DUNSON, D. [2010]: Probabilistic topic models. A focus on graphical model design and applications to document and image analysis. *IEEE Signal Process Magazine*. Vol. 27. Issue 6. pp. 55–65. https://doi.org/10.1109/MSP.2010.938079

BLEI, D. M. – MCAULIFFE J. D. [2007]: *Supervised Topic Models.* Advances in Neural Information Processing Systems 20 (NIPS 2007) http://papers.nips.cc/paper/3328-supervised-topic-models

BOWMAN, R. H. JR. [2008]: *Business Continuity Planning for Data Centers and Systems: A strategic Implementation Guide*. John Wiley & Sons, Inc. Hoboken.

BYORD-GRABER, J. – BLEI, D. M. [2009]: Syntactic topic models. In: *Koller, D. – Schuurmans, D. – Bengio, Y. – Bottou, L.* (eds.): *Advances in Neural Information Processing Systems 21: 22nd Annual Conference on Neural Information Processing Systems, 2008*. Curran Associates Inc. Red Hook. pp. 185–192. https://papers.nips.cc/paper/3398-syntactic-topic-models

CHEN, M. [2020]: *A Guide: Text Analysis, Text Analytics & Text Mining*. 21 October. https://towardsdatascience.com/a-guide-text-analysis-text-analytics-text-mining-f62df7b78747

FEURER, M. – SPRINGENBERG, T. J. – HUTTER, F. [2015]: Initializing Bayesian hyperparameter optimization via meta-learning. In: *Bonet, B. – Koenig, S.* (eds.): *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence and the Twenty-Seventh Innovative Applications of Artificial Intelligence Conference*. Association for the Advancement of Artificial Intelligence. Pablo Alto. pp. 1128–1135. https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/10029/9349

GERLACH, M. – PEIXOTO, T. P. – ALTMANN, E. G. [2018]: A network approach to topic models. *Science Advances*. Vol. 4. No. 7. https://doi.org/10.1126/sciadv.aaq1360

HOFFMAN, M. – BLEI, D. – COOK, P. [2008]: Content-based musical similarity computation using the hierarchical Dirichlet process. In: *Bello, J. P. – Chew, E. – Turnbull, D*. (eds.): *ISMIR 2008: Proceedings of the 9th International Conference on Music Information Retrieval*. Lulu.com. Research Triangle. pp. 349–354.

HOFFMAN, M. D. – BLEI, D. M. – COOK, P. R. [2009]: Finding latent sources in recorded music with a shift-invariant HDP. In: *Proceedings of the 12th International Conference on Digital Audio Effects (DAFx-09), Como, Italy, September 1–4, 2009*. https://dafx09.como.polimi.it/proceedings/data/DAFx09_Proceedings.pdf

IOS (INTERNATIONAL ORGANIZATION FOR STANDARDS) [2012]: *ISO 22301:2012 Security and Resilience – Business Continuity Management Systems – Requirements*. (MSZ EN ISO 22301:2014 Szociális biztonság. Üzletvitel-irányítási rendszerek. Követelmények.)

IOS [2015]: *ISO/TS 22317:2015 Societal Security – Business Continuity Management Systems – Guidelines for Business Impact Analysis (BIA)*.

JONES, T. W. [2014]: *Introduction to Topic Modeling with LDA and More*. https://www.jonesingfordata.com/talk/2014_11_12_dcnlp/

JONES, T. W. [2019]: *A coefficient of Determination for Topic Models*. 26 November. https://doi.org/10.48550/arXiv.1911.11061

KATONA, E. R. [2018]: *Látens topikok és látható címkék Az Author-Topic Model gyakorlati alkalmazása a korrupció témában.* PhD-dolgozat Eötvös Loránd Tudományi Egyetem. Budapest

LIU, Y. – NICULESCU-MIZIL, A. – GRYC, W. [2009]: Topic-link LDA: Joint models of topic and author community. In: *Association for Computing Machinery*: *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning, June 2009*. New York. pp. 665–672. https://doi.org/10.1145/1553374.1553460

MIMNO, D. – MCCALLUM, A. [2012]: *Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression*. 13 June. https://doi.org/10.48550/arXiv.1206.3278

MIMNO, D. [2012]: *Reconstructing Pompeian households*. 14 February. https://doi.org/10.48550/arXiv.1202.3747

MINER, G. – DELEN, D. – ELDER, J. – FAST, A. – HILL, T. – NISBER, R. A. [2012]: *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Elsevier Science Publishing Co. Inc. San Diego.

MUNOZ-GARCIA, O. – NAVARRO, C. [2012]: Comparing user generated content published in different social media sources. In: *Calzolari, N. – Choukri, Kh. – Declerck, Th. – Doğan, M. U. – Maegaard, B. – Mariani, J. – Moreno, A. – Odijk, J. – Piperidis, S.* (eds.): *LREC 2012, Eighth International Conference on Language Resources and Evaluation, @NLP can u tag #user_generated_content?! via lrec-conf.org. Workshop*. 26 May. http://www.lrec-conf.org/proceedings/lrec2012/index.html

NHS ORGANISATIONS IN SCOTLAND [n. a.]: *Business Continuity. A Framework for NHS Scotland. Strategic Guidance for NHS Organisations in Scotland*. https://www.sehd.scot.nhs.uk/EmergencyPlanning/Documents/BusinessContinuity.pdf

OKOLITA, K. [2010]: *Building an Enterprise-Wide Business Continuity Program*. Taylor and Francis Group. London.

PRUTEANU-MALINICI, J. – REN, L. – PAISLEY, J. – WANG, E. – CARIN, L. [2010]: Hierarchical Bayesian modeling of topics in time-stamped documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 32. Issue 6. pp. 996–1011. https://doi.org/10.1109/TPAMI.2009.125

SAYAK, P. [2018]: Hyperparameter optimization in machine learning models. *Datacamp*. 15 August. https://www.datacamp.com/tutorial/parameter-optimization-machine-learning-models

SEBŐK M. [2016]: *Kvantitatív szövegelemzés és szövegbányászat a politikatudományban*. L'Harmattan Kiadó. Budapest.

SETTANNI, M. – MARENGO, D. [2015]: Sharing feelings online: Studying emotional well-being via automated text analysis of Facebook posts. *Frontiers in Psychology*. 23 July. https://doi.org/10.3389/fpsyg.2015.01045

SPERLING, M. [2020]: *Topic Modeling of Church Father Writings*. https://www.academia.edu/43960210/Topic_Modeling_of_Church_Father_Writings

TUCKER, E. [2015]: *Business Continuity from Preparedness to Recovery – A Standards-Based Approach*. Butterworth-Heinemann Inc. Woburn.

WANG, N. – KOSINSKI, M. – STILLWELL, D. J. – RUST, J. [2014]: Can well-being be measured using Facebook status updates? Validation of Facebook's gross national happiness index. *Social Indicators Research*. Vol. 115. 3 February. pp. 483–491. https://doi.org/10.1007/s11205-012-9996-9

ZOU, L. – SONG, W. [2016]: LDA-TM: A two-step approach to Twitter topic data clustering. In: *Institute of Electrical and Electronics Engineers*: *Proceedings of 2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA 2016): July 5–7, 2016, Chengdu, China*. Curran Associates, Inc. Red Hook. pp. 342–347. http://dx.doi.org/10.1109/ICCCBDA.2016.7529581