Fesztiválállátogatók véleményeinek számítógéppel támogatott tematikus modellezése – egy kísérlet eredményei

Computer-aided topic modelling based on festival-goers' opinions – results of an experiment

Szerző: Hinek Mátyás[1]

Tanulmányunkban arra teszünk kísérletet, hogy egy számítógépes algoritmus, a rejtett Dirichlet eloszlást alkalmazó strukturált témamodell (stm) segítségével meghatározzuk a Sziget Fesztivál látogatói által a Facebookon írt vélemények jellemző témáit, és ezeket összevessük egy korábbi kutatásunkban körvonalazott témákkal. A Sziget Fesztivál látogatóinak az elmúlt hét évben angol nyelven írt szöveges véleményei alapján az algoritmus segítségével kilenc témát modelleztünk, melyek tartalma és köre csak részben egyezett meg a korábbi, kvalitatív kutatásunkban azonosított témákkal. Vizsgálatunk legfontosabb eredménye, hogy számítógépes eszközökkel eredményesen vizsgálhatók a látogatói vélemények, ugyanakkor az eredmények minőségét meghatározza a korpusz nagysága, vagyis az elemzett hozzászólások száma és terjedelme.

In our study, we attempt to determine the typical topics of opinions written by Sziget Festival visitors on Facebook using structured topic model (stm) computer algorithm and latent Dirichlet allocation, and compare the results with our previous research. Based on written opinions of the visitors of the Sziget Festival in the last seven years, we modelled nine topics. Their content and scope partly matched the topics identified in our previous qualitative research. The most important result of our study is that visitor opinions can be successfully examined with computer tools, but the quality of the results is determined by the size of the corpus, i.e. the number and scope of the analysed posts.

Kulcsszavak: közösségi média elemzés, számítógépes szövegfeldolgozás, témamodellezés, látens Dirichlet allokáció.
Keywords: social media analysis, natural language processing, topic modelling, latent Dirichlet allocation.

1. Bevezetés

Az információk elektronikussá és interneten hozzáférhetővé válásával óriási mennyiségű írott, képi és egyéb formátumú adat áll rendelkezésünkre. A rendezett formában elérhető információk mellett nap mint nap hatalmas tömegű strukturálatlan információ is generálódik, amelyek megtalálhatósága, visszakereshetősége, feldolgozhatósága gyakran korlátokba ütközik, különös tekintettel a közösségi média és egyéb felületeken megjelenő bejegyezésekre.

A szöveges, képi és egyéb formátumú úgynevezett *big data* gyűjtésére, kezelésére és ebből hasznos információ kinyerésére az elmúlt években számítógépes technológiák egész sora jött létre, többek között internetes keresőmotorok, az azokat támogató statisztikai módszertan, gépi tanulási algoritmusok, illetve a mesterséges intelligencia megoldások, amelyek forradalmasítják az információ feldolgozását.

[1] főiskolai tanár, Budapesti Metropolitan Egyetem, mhinek@metropolitan.hu

Jelen tanulmány ennek a problémakörnek egy speciális részterületével foglalkozik. Azt vizsgáljuk, hogy a közösségi média felületen elérhető írásos vendégvélemények milyen módon és minőségben elemezhetők számítógépes módszerek segítségével. A vizsgálat során a Sziget Festival Official Facebook oldalán olvasható angol nyelvű látogatói véleményeket dolgoztuk fel, arra keresve a választ, hogy milyen témák azonosíthatóak a bejegyzésekben, illetve hogy az azonosított témák mennyiben relevánsak, ha ugyanezen információk egy részhalmazának emberi (kézi) feldolgozásával hasonlítjuk össze.

2. A számítógépes témamodellezés tudományos háttere

A számítógépes témamodellezéshez a látens Dirichlet allokációt (LDA) alkalmaztuk. Az LDA egy úgynevezett nem felügyelt gépi tanulási algoritmus, amely a témákat, mint látens (rejtett) információkat azonosítja nagy dokumentumgyűjteményekben. Nem felügyelt abban az értelemben, hogy a témák azonosításához nem alkalmaz előre elkészített szó- vagy fogalomtárat, a kutató nem *tanítja be* előzetesen az algoritmust.

A látens Dirichlet allokáció (LDA) azon a hipotézisen alapul, hogy a szerző bizonyos témákat szem előtt tartva ír meg egy dokumentumot. Egy adott témához a témára jellemző szavakból bizonyos valószínűséggel választ ki egy adott szót. A dokumentum egésze különféle témák keverékeként jellemezhető.

Az LDA a témák *visszafejtéseként* értelmezhető, modellezési folyamata úgy írható le, hogy a dokumentumokban (jelen esetben a közösségi médiában megjelenő hozzászólásokban) meghatározza az abban előforduló témák keverékét, mint valószínűségi eloszlást, majd mindegyik témához hozzárendeli a témára jellemző kifejezéseket (szavakat), mely hozzárendelés szintén valószínűségi eloszlást követ (KRESTEL et al. 2009). Az LDA tehát azt feltételezi, hogy egy dokumentumgyűjtemény (korpusz) minden dokumentuma K számú rejtett téma valószínűségi eloszlásaként reprezentálható, illetve minden téma a dokumentumok szókincsét alkotó szavak multinomiális eloszlásaként határozható meg (BLEI et al. 2003).

Az LDA grafikus reprezentációját az *1. ábra* mutatja. A szöveggyűjtemény (korpusz) *M* darab dokumentumot tartalmaz. Minden dokumentum N_i számú szóból áll, $w_{d,n}$ jelöli a *d.* dokumentum *n.* szavát. A korpusz összes szava *K* számú témába csoportosítható. A korpusz egyedi szavait a *V* jelöli, míg a *Z* azt, hogy a dokumentumot alkotó szavak melyik témához tartoznak. Minden egyes téma multinomális eloszlásként értelmezhető a dokumentumokat alkotó szavak halmazán. Az α és β priori Dirichlet hiperparaméterek. Értékük jellemzően alacsony, ami arra a feltételezésre épül, hogy a témák száma és az egyes témákhoz tartozó szavak száma is korlátozott, azaz *ritka*. A modellben egyetlen megfigyelt változó van, a korpuszt alkotó szavak. A modellből kikövetkeztetni kívánt rejtett változók:

- $\theta_{d,}$ a témák valószínűségi eloszlása a *d* dokumentum felett,
- φ_k a *k* témához tartozó szavak valószínűségi eloszlása,
- $z_{dn,}$ *témacímke*, ami azt jelöli, hogy a dokumentumok egyes szavai melyik témába sorolhatók.

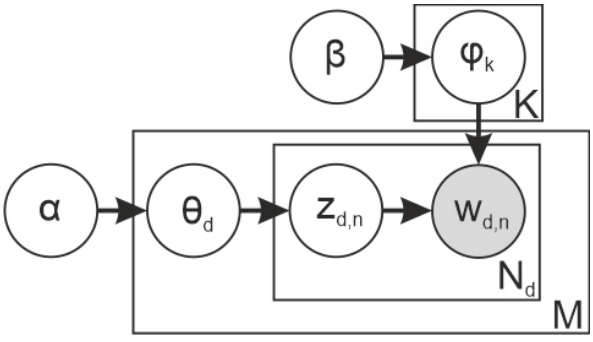
Az LDA generatív folyamata a következő (BALOGH 2015):

1. Minden k téma esetében válasszunk egy $\varphi_k \sim Dir(\beta)$ szóeloszlást.
2. Minden dokumentum esetében
   a. válasszunk egy $\theta_d \sim Dir(\alpha)$ témaeloszlást.
   b. Minden egyes szó esetében
      i. válasszunk egy $z_{dn} \sim Mult(\theta_d)$ témahozzárendelést, ahol a $z_{dn} \epsilon\ 1....K$;
      ii. válasszunk egy $w_{d,n} \sim Mult(\varphi_{z_{d,n}})$ szót, ahol $w_{d,n} \epsilon\ 1, \ldots, V$.

1. ábra

Az LDA modell változói között fennálló kapcsolat, *tálca* reprezentáció segítségével

(A külső tálca a dokumentumokat, míg a belső tálca a témák és szavak ismételt kiválasztását jelenti a dokumentumban)



Forrás: BLEI et al. 2003, BALOGH 2015

A modell különféle eloszlásainak poszteriorai, azaz a témák halmazának, a hozzájuk kapcsolódó szavak valószínűségeinek, az egyes szavakhoz tartozó témáknak, valamint az egyes dokumentumok specifikus téma-mixeinek megismerése statisztikai következtetéssel történik. Többféle módszer is alkalmazható, de leggyakrabban a Gibbs-mintavételt alkalmazzák, amely jól algoritmizálható.

Az LDA további jellemzője, hogy nem veszi figyelembe a szavak pozícióját, sorrendjét a dokumentumokban, és a dokumentumok sorrendjét sem. Emiatt az LDA-t gyakran *szózsák* (*bag of words*) modellként is jellemzik (BLEI et al. 2003). Szintén fontos megjegyezni, hogy az LDA vegyes tagságú modell, azaz nem hagyományos klasszifikációról van szó, a modellezés végeredményeként a témákhoz azonosított szavak nem lesznek kizárólagosak. Ugyanaz a szó több téma esetében is felbukkanhat más-más súllyal, illetve valószínűséggel (AIROLDI et al. 2014).

Több kutató az eredeti modell továbbfejlesztését javasolta, javítva a témamodellek predikciós képességét, lásd például PAUL–DREDZE 2015, GERRISH–BLEI 2012, WEINSHALL et al. 2013, WILSON–CHEW 2010, EISENSTEIN et al. 2011, BLEI–LAFFERTY 2006 (stb.) munkáit. Jelen vizsgálatban az eredeti LDA modell továbbfejlesztését javasló ROBERTS és szerzőtársai (2019) munkájára támaszkodtunk. Az általuk kidolgozott strukturált témamodell (stm) olyan szövegek feldolgozására alkalmas, amelyekhez rendelkezésre állnak dokumentumszintű metaadatok. Ezek a metaadatok tetszőlegesek. Lehet például a dokumentum keletkezési időpontja, a hozzászóló nemzetisége, esetleg a kedvelések (lájkok) száma, stb. De lehet folytonos (például járművek adatlapjain a gyorsulás mérőszáma) vagy diszkrét jellegű (például minősítés egy ötfokozatú skálán). A modell kimenetei felhasználhatóak a metaadatok és a témák kapcsolatait leíró hipotézisek tesztelésére is (ROBERTS et al. 2019).

3. Kutatási előzmények

A látens Dirichlet allokáció társadalomtudományi és egyéb alkalmazására magyar példák is fellelhetők az elmúlt évekből. Balogh az ELTE mesterképzésén íródott szakdolgozatában a kuruc.info romaellenes megnyilvánulásait vizsgálta (BALOGH 2015). Bíró PhD értekezésében a rejtett Dirichlet allokáció dokumentumosztályozási lehetőségeit vizsgálta, az eredeti LDA modellt multi-korpusz (MLDA) és linkalapú LDA modellekké továbbfejlesztve (BÍRÓ 2009). A linkalapú LDA modellt Bíró és társai a webes spamszűrésben alkalmazták, ami a konkurens megoldásokhoz képest valamivel magasabb hatékonyságúnak bizonyult a tesztelés során (BÍRÓ et al. 2009a). Ugyancsak Bíró és társai vizsgálták, hogy hogyan alkalmazható a látens Dirichlet allokáció a szöveges dokumentumok felügyelt szemantikai kategorizálása során (BÍRÓ et al. 2009b).

A nemzetközi szakirodalomban számos példa található az LDA, illetve a strukturált témamodell (stm) alkalmazására, ám jellemzően nem a turizmus területén. Curry és Fix azt vizsgálták, hogy az amerikai államok fellebbviteli bíróságain tevékenykedő bírók hogyan használják a Twittert (CURRY–FIX 2019). Fischer-Preßler és szerzőtársai 51 ezer *tweet* alapján vizsgálták az emberek tipikus reakcióit a 2016-os berlini adventi vásár elleni terrortámadásra, rámutatva arra, hogy a bejegyzések témái hogyan változtak a támadást követő napokban (FISCHER-PREßLER et al. 2019). Rodriguez és Storer ugyancsak Twitter bejegyzéseket elemezve azt vizsgálta, hogy az emberek miért maradnak bántalmazó kapcsolatokban, vagy miért távoznak azokból. Módszertani megállapításuk az, hogy a témamodellezés hasznos módszer a nem strukturált közösségi média adatkészletek leíró elemzésére, emellett a témamodell eredményei vezethetik be a mélyebb kvalitatív elemzést (RODRIGUEZ–STORER 2020).

Jelen kutatás közvetlen előzményeként Hinek és Kulcsár a Sziget Fesztivál élménydimenzióit vizsgálta. Kutatásuk során kvalitatív szövegelemzést végeztek a Sziget Festival Official honlapján megjelenő vendégvélemények körében. Ennek során számos témát (a fesztivállal kapcsolatos elégedettség különböző tényezőit) azonosítottak a látogatók által írt hozzászólásokban, amelyeket hierarchikus gondolattérkép segítségével rendszereztek (HINEK–KULCSÁR 2019).

Jelen vizsgálat arra keresi a választ, hogy a számítógépes témamodellezés is képes-e azonosítani azokat a témákat, amelyeket korábban emberi közreműködéssel detektáltunk. Tágabban értelmezve pedig azt vizsgáljuk, hogy automatizálható-e, és ha igen, milyen pontossággal (részletezettséggel), az emberi munka folyamata egy speciális szövegtesten, a Facebookon, megjelenő vendégvélemények esetében.

4. Módszertan

Az adatokat a Facebookon a Sziget Festival Official oldalán nyilvánosan elérhető vendégvélemények rovatából gyűjtöttük. A vizsgálatot a 2014. január 1. és 2020. július 31. között íródott vélemények körében hajtottuk végre. Ebben az időszakban több ezer bejegyzést találtunk, de ezek döntő többsége semmiféle szöveges információt, véleményt nem tartalmazott, mivel csak a fesztivál 1-5 skálán történő értékelését adták meg a látogatók. A bejegyzések körülbelül 8-10%-a tartalmazott szöveges információt, melyek gyakorisága évről évre eltérően alakult. A legtöbb szöveges bejegyzést 2017-ben regisztráltuk, ezt követte a 2016-os, majd a 2015-ös és a 2014-es évek. 2018 volt az utolsó év, amikor szöveges bejegyzést találtunk a vélemények rovatban. Az azt követő években már csak számszerű értékelések születtek.

A szöveges véleményeket kézzel másoltuk ki a Facebookról. A vélemények mellett rögzítettük a véleményt író nevét (profiljának nevét), ha azono-

sítható volt, akkor származási országát, a jelölések (lájkok és egyéb emotikonok) összesített számát, a posztoló 1-5 skálán adott értékelését, valamint a bejegyzés időpontját. A posztok alá írt hozzászólásokat szintén kigyűjtöttük, tekintettel arra, hogy elemzésünk szempontjából értékes információkat tartalmaztak, de itt már nem állt rendelkezésre a bejegyzés dátuma.

A következő lépésben az elsődleges adattisztítást végeztük el. A legfontosabb döntésünk az volt, hogy csak az angol nyelvű bejegyzéseket tartjuk meg. Ugyan a témamodellezés bármely nyelven lefolytatható, a módszertan statisztikai jellegéből adódóan azonban egyidejűleg csak azonos nyelvű korpuszon végezhető el hatékonyan az elemzés, máskülönben az ugyanolyan értelmű, de eltérő nyelven megjelenő kifejezések és az ezekből modellezett témák nem lesznek koherensek. A bejegyzések körülbelül 60-70%-a egyébként is angol nyelvű volt, különösen azok, amelyek a fesztivál *érdemi* tényezőivel (zenei kínálat, catering, kiegészítő szolgáltatások, látogatói élmény) foglalkoztak, míg a magyar nyelvű bejegyzéseket gyakran nem a fesztivál látogatói írták.

A további adattisztítás során kihagytuk a bejegyzések közül azokat, amelyek spam jelleggel kerültek be a korpuszba (például a fesztivál zajosságával kapcsolatban, többször egymás után, pontosan ugyanazzal a tartalommal írt bejegyzéseket), valamint kiszűrtük a Sziget Festival Official által írt válaszokat is.

A nyelvi jellegű előkészítést megelőzően a 752 bejegyzést tartalmazó korpuszunk összesen mintegy 37 ezer szót és 3800 egyedi szót (kifejezést) tartalmazott, összesen 175 ezer karakter terjedelemben. A kutatási előzményekhez képest kicsi korpusz jött létre, ami nem feltétlenül vezet kielégítő eredményekhez a géppel végrehajtott statisztikai-valószínűségi feldolgozás során, mivel az eljárás együttesen előforduló kifejezések/szavak alapján határoz meg témákat, így a nagyobb korpusz előnyösebb (CROSSLEY et al. 2017, SYED–SPRUIT 2017). Az ilyen kis korpusz emberi közreműködéssel is elemezhető, ám igen munkaigényes feladat, miközben az ember által végzett témaallokáció sem feltétlenül koherens.

Voltak olyan vélemények is, amelyek több száz szavasak voltak, sok bejegyzés azonban rövid volt, sőt előfordult olyan bejegyzés is, amely csak egyetlen szót tartalmazott (például „Super!"). Mivel az LDA minden dokumentumot témák mixeként kezel, és a témák a dokumentumokat alkotó szavak eloszlásaiból alakulnak ki, a nagyon rövid bejegyzések nem ideálisak a témamodellezéshez. Vannak olyan vizsgálatok, amelyek kizárják a 7-8 szónál rövidebb bejegyzéseket a témamodellezésből (WANG et al. 2013), míg más vizsgálatok erre nem fektetnek hangsúlyt (lásd a korábban bemutatott Twitter alapú témamodelleket, ahol a bejegyzések maximum 140 karakter hosszúságúak). A korpusz méretét mérlegelve úgy döntöttünk, hogy az első lépésben megtartjuk a rövid bejegyzéseket is, hiszen a következő lépésben, a nyelvi előfeldolgozás során számos szó, illetve dokumentum esett ki a kiinduló korpuszból.

A feldolgozáshoz az R (R CORE TEAM 2020) stm (ROBERTS et al. 2019) csomagját alkalmaztuk, az elemzést teljes egészében az R-el végeztük. A nyers adatok előfeldolgozása során a korpuszból eltávolítottuk az írásjeleket, a számokat, valamint a gyakran előforduló, de a szövegben jelentést nem hordozó szavakat (az úgynevezett *stopszavakat*, például a névelőket és a segédigéket). Ugyanannak a szónak többféle alakja is előfordulhat, ami torzítaná az elemzést, így a szavakat lemmatizáltuk, azaz a ragozott és képzett alakok szótövét hagytuk meg (például amazing → amaz, beautiful → beauti).

A következő fázisban a ritkán előforduló szavakat is kizártuk az elemzésből a pontosság javítása érdekében. Csak azokat tartottuk meg, amelyek legalább három dokumentumban előfordultak. Hasonlóképpen, a sok dokumentumban előforduló igen gyakori szavak is elhagyhatók (a mi esetünkben ilyen a „sziget" és a „festival" szó), ám a próbafuttatások után úgy döntöttünk, hogy ezeket megtartjuk, mert nem torzítják az elemzést.

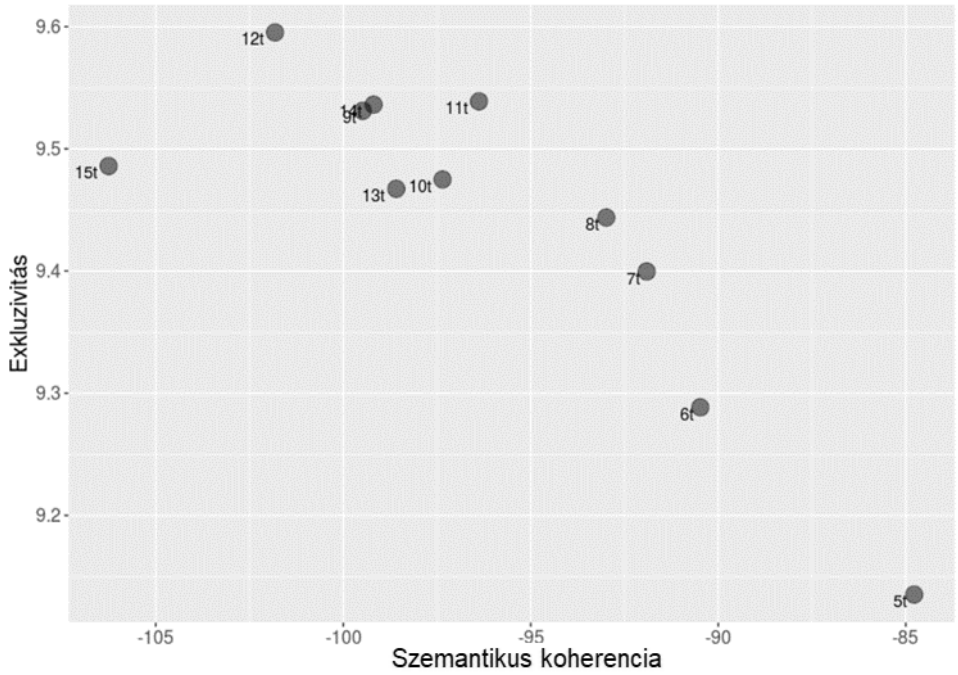
Az stm R csomag lehetőségeit kihasználva két dokumentumszintű metaadatot vontunk be az elemzésbe: a bejegyzés évét, valamint a jelölések számát. A további dokumentumszintű metaadatokat (például a hozzászóló nemzetiségét, az értékelést az ötfokozatú skálán, stb.) ki kellett hagynunk az elemzésből, mert nem álltak rendelkezésre minden bejegyzés esetében. Az előkészítési lépéseket követően 736 dokumentumunk és 751 egyedi kifejezést tartalmazó szótárunk maradt a témamodellezés elvégzésére.

A témamodellezés fontos jellemzője, hogy a témák számát (K) előre kell meghatározni. Korábbi kutatásunkban 9 fő témát, majd ezeket tovább bontva közel 30 altémát azonosítottunk. Ezt tapasztalati (empirikus) témaszámként kezeltük. Emellett az stm csomag lehetőségei segítségével két további mutatót is vizsgáltunk: a szemantikus koherenciát, ami azt jelzi, hogy a témákat mennyiben alkotják egymáshoz kapcsolódó szavak, amelyek megkönnyítik a témák egyértelmű azonosíthatóságát, illetve az exkluzivitást, ami azt fejezi ki, hogy milyen mértékben eltérő szavak alkotják az egyes témákat *(2. ábra)*. Amennyiben modellünket két témával futtatjuk le (lásd a 2. ábra jobb alsó sarkát!), akkor lenne a legmagasabb a szemantikus koherencia értéke, de

a két témát alkotó kifejezések exkluzivitása alacsony lenne. Ebből a szempontból 11 vagy 12 téma teljesítene a legjobban, ám ebben az esetben a szemantikus koherencia értéke alacsony. A két mutató szerinti (valamennyire) kiegyensúlyozott témaszám a 8, 9, 10 11, 14 lenne, bár egyik mutató szempontjából sem találunk valóban kedvező értékeket. A kutatási előzményeket is figyelembe véve úgy döntöttünk, hogy 9 témaszámmal folytatjuk le a vizsgálatot.

2. ábra

Exkluzivitás és szemantikus koherencia különböző számú témák esetében a Sziget Fesztivál látogatói véleményeinek témamodellezése során (Mindkét mutató esetén a magasabb érték kedvezőbb, az ábrán ez a jobb felső sarokhoz közelebbi pontokat jelenti)



Forrás: saját szerkesztés az stm és a ggplot2 R programcsomagok segítségével

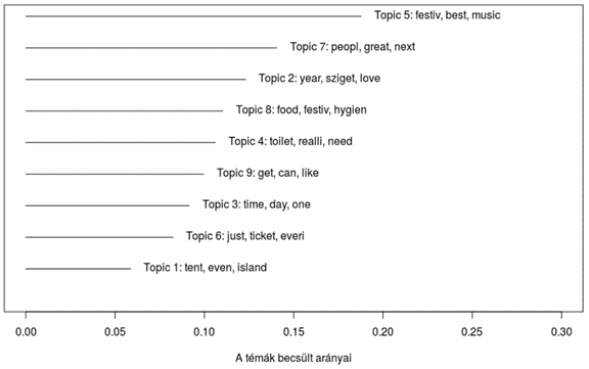
5. A vizsgálat eredményei

A statisztikai valószínűségek alapján az algoritmus megadja, hogy melyek az egyes témákhoz tartozó legvalószínűbb szavak és hogy mekkora a témák aránya az egyes dokumentumokban (hozzászólásokban), illetve a korpusz egészében (lásd a *3. ábrát*!). A látogatói véleményekben az 5. téma jelenik meg a legnagyobb becsült gyakorisággal, részaránya meghaladja a 18%-ot, míg a további témák az algoritmus szerint 6-14%-os valószínűséggel fordulnak elő.

Az stm csomag megadja azokat a szavakat is, amelyek egyszerre gyakoriak és kizárólagosak (exkluzívak). Ezek azok a szavak, amelyek megkülönböztetik egymástól a témákat (FREX). A *lift* és *score* mutatók egy szó súlyozott gyakoriságát mérik az adott témában. A súlyok úgy állnak elő, hogy a más témákban ritkábban előforduló szavak magasabb súlyt kapnak. A témák azonosítása a szavak alapján már kutatói feladat, lásd az *1. táblázatot*!

3. ábra

A témák (Topic) főbb szavai és becsült arányai a hozzászólásokban



Forrás: saját szerkesztés az stm R programcsomag segítségével

1. táblázat

A témákat reprezentáló szavak

Témák	*Lemmatizált szavak a különféle mutatók szerint*	*A téma azonosítása*
1.	*Legvalószínűbb: tent, even, island, peopl, use, complet, mani* *FREX: past, tent, worth, horribl, complet, even, use* *Lift: direct, gone, leak, worth, yet, across, auchan* *Score: tent, refund, wash, station, often, continu, leak*	*A Sziget kempinggel kapcsolatos tapasztalatok*
2.	*Legvalószínűbb szavak: year, sziget, love, experi, thank, come, still* *FREX: love, thank, year, experi, stay, come, absolut* *Lift: air, heaven, unbeliev, altern, blast, contact, epic* *Score: year, love, thank, absolut, welcom, fun, experi*	*Nagyszerű fesztiválélmény a Szigeten*
3.	*Legvalószínűbb: time, day, one, thing, think, line, first* *FREX: agre, think, line, total, guy, idea, thing* *Lift: agre, comparison, total, breakfast, except, front, guest* *Score: guest, time, one, thing, day, line, think*	*Vélemények, gondolatok a fesztivál különféle tényezőivel kapcsolatban (például elhozhatók-e gyerekek a Szigetre?)*
4.	*Legvalószínűbb: toilet, realli, need, water, price, camp, last* *FREX: water, queue, price, arriv, disappoint, eat, fest* *Lift: bottl, min, queue, tap, add, arriv, belie* *Score: toilet, water, suck, enough, need, problem, price*	*Az alapvető higiéniai és egyéb szolgáltatásokkal kapcsolatos vélemények (például a palackos víz árazása, nyilvános csapok)*
5.	*Legvalószínűbb: festiv, best, music, amaz, back, ever, life* *FREX: best, music, ever, life, beer, world, amaz* *Lift: beer, cultur, die, england, ever, excit, greatest* *Score: best, music, amaz, festiv, ever, life, world*	*A legjobb fesztivál, a legnagyobb élmény*
6.	*Legvalószínűbb: just, ticket, everi, way, stage, feel, crowd* *FREX: crowd, act, budapest, play, push, way, everi* *Lift: act, anniversari, budapest, corner, bare, beach, beat* *Score: ticket, crowd, lineup, budapest, play, way, broke*	*A tömeggel és a zenei felhozatallal kapcsolatos vélemények*
7.	*Legvalószínűbb: peopl, great, next, will, see, place, much* *FREX: great, awesom, look, next, everyon, will, far* *Lift: awesom, alcohol, bin, control, dirti, drunk, environ* *Score: great, next, awesom, peopl, will, simpl, look*	*A nagyszerű fesztiválozókkal, helyszínnel, a kiváló szervezéssel és szolgáltatásokkal kapcsolatos vélemények*
8.	*Legvalószínűbb: food, festiv, hygien, staff, got, sziget, realli* *FREX: hygien, staff, dust, visitor, trash, right, food* *Lift: age, bacteria, drive, favourit, feedback, mate, result* *Score: food, hygien, poison, sick, staff, favourit, bar*	*A cateringgel, valamint élelmiszerhigiéniával kapcsolatos tapasztalatok*
9.	*Legvalószínűbb: get, can, like, sziget, stage, dont, mani* *FREX: concert, sound, live, perform, phone, band, get* *Lift: move, park, phone, sound, steve, aoki, assum* *Score: sound, get, concert, telekom, assum, aoki, chainsmok*	*A koncertekkel, fellépőkkel, előadókkal, a Sziget színpadaival és helyszíneivel kapcsolatos vélemények*

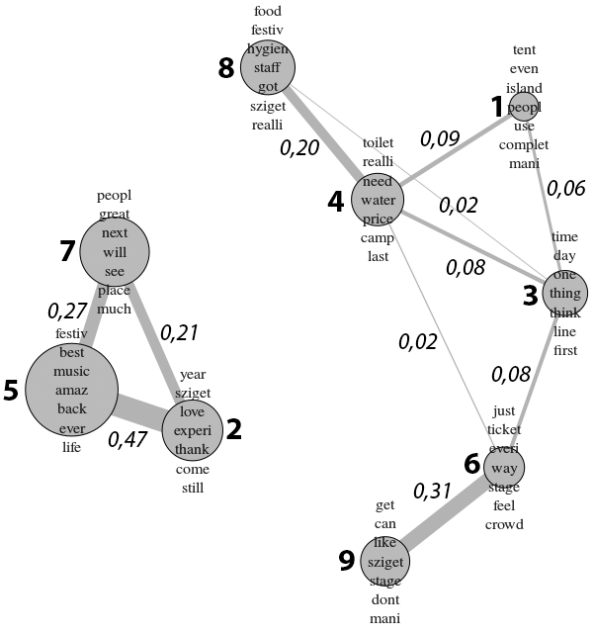
Forrás: saját szerkesztés az stm R programcsomag segítségével

Esetenként nehéz csak a szavak alapján azonosítani az egyes témák valós tartalmát, különösen, ha a szavak között gyenge a szemantikai kapcsolat. Így több téma esetében megvizsgáltuk, hogy mely konkrét véleményekben jelentek meg a legnagyobb valószínűséggel. Például a 3. téma két egymással szorosan összefüggő hozzászólásban jelent meg, amelyek arról szóltak, hogy érdemes-e gyerekekkel elmenni a Sziget Fesztiválra, illetve, hogy ezzel a hozzászólók mennyiben értettek egyet. A további, témára jellemző hozzászólásokban a posztolók ugyancsak valamilyen tényezővel kapcsolatos véleményt fogalmaztak meg („I think"), esetleg támogatták más hozzászólók véleményét („I agree").

A pozitív Sziget-élmény több témában is megjelenik. Ilyen a 2., az 5. és a 7. téma, melyek másmás szavakkal fejezik ki (majdnem) ugyanazt, nevezetesen, hogy a Sziget Fesztivál nagyszerű élmény. Ezek a témák korrelálnak is egymással, az algoritmus a témák korrelációs mátrixát is kiszámolta *(4. ábra)*. Az ábrán a csúcsok a témákat képviselik, nagyságuk az egyes témák becsült arányát jelzik a korpuszban. A csúcsokat összekötő élek vastagsága a témák közti korreláció erősségét mutatja. A legjelentősebb – de erősségét tekintve csak közepes – korrelációs kapcsolat a 2. és az 5. téma között figyelhető meg (a Pearson-féle korrelációs együttható értéke 0,47). Az 5. és 7. téma között 0,27, illetve a 7. és 2. téma között 0,21 a korreláció mérőszáma, ami gyenge-közepes kapcsolatra utal.

4. ábra

A témák súlya és a közöttük lévő korrelációs kapcsolatok



Forrás: saját szerkesztés az stm és igraph R programcsomagok segítségével

Az ábrán jól látható, hogy az élményekkel kapcsolatos témák nem korrelálnak a fesztivál különféle szolgáltatási tényezőivel kapcsolatos témákkal. Ezek közül a 8. téma súlya átlagosan 11% a hozzászólásokban, itt jelennek meg a cateringgel, valamint az élelmiszerhigiéniával kapcsolatos tapasztalatok. Ugyanakkor a 8. téma enyhén korrelál a 4. témával, ami az alapvető higiéniai és egyéb szolgáltatásokkal kapcsolatos véleményeket képviseli. Ebben a *bokorban* találhatók továbbá a kempinggel kapcsolatos tapasztalatok, a fellépőkkel és a line-uppal, illetve a tömeggel kapcsolatos vélemények, lásd a 6. és 9. témát, amelyek között közepes korreláció figyelhető meg. Együttesen főként a kiegészítő jellegű szolgáltatási elemekkel kapcsolatos témák halmazát képviselik, amelyeket a korábbi vizsgálatunkban is megkülönböztettünk a fesztivál alapszolgáltatásaitól.

Az előzmény kutatásunkban azonosított fő témák az alapszolgáltatások (line-up, előadók, hangosítás), a támogató szolgáltatások (catering, beléptetés, személyzet, fizetési rendszer, szaniteregységek és higiénia, kemping stb.) és a kiegészítő élményelemek (környezet, kiegészítő programok) voltak, a fesztiválatmoszféra, a látogatók, a zsúfoltság és az egyéb tényezők (például az időjárás) mellett, melyeket további 25-30 résztényezőre bontottunk. Ha ezt összevetjük jelen számítógépes témamodell eredményeivel, látható, hogy jelentős átfedések vannak (catering, élelmiszerbiztonság, higiéniai szolgáltatások, kemping, fesztiválozók és atmoszféra), de a fő témák tartalma csak részben egyezik meg. Különösen a 2. és az 5. téma különbözik jelentős mértékben a kvalitatív vizsgálatban azonosítottaktól. Ennek legfontosabb oka az, hogy a hozzászólások nagy hányadában az elégedettség és a különféle tetszésnyilvánítások jelentek meg, amelyeket a fesztivál *megfogható*, önállóan is értékelhető tényezőit (*élményelemeit*) azonosító kvalitatív elemzés során nem tudtunk jól megragadni, mivel egy tetszésnyilvánítás nem élményt alakító tényező. A 2., az 5. és a 7. téma együttes részaránya több mint 45%, így a kvantitatív elemzés jobban kiemeli a hozzászólások legfontosabb dimenzióját, a véleményt írók által megélt egyedi, felejthetetlen élményt.

Végül, de nem utolsó sorban, a kovariánsok szerepét is megvizsgáltuk. A strukturált témamodellben két metaadatot, a vélemény évét és a bejegyzéshez tartozó lájkok, egyéb megjelölt emotikonok összesített számát építettük be. Az stm csomag segítségével megvizsgáltuk, hogy a kovariánsok hatására hogyan alakul az egyes témák eloszlása a hozzászólásokban (lásd az *5. ábrát*!).

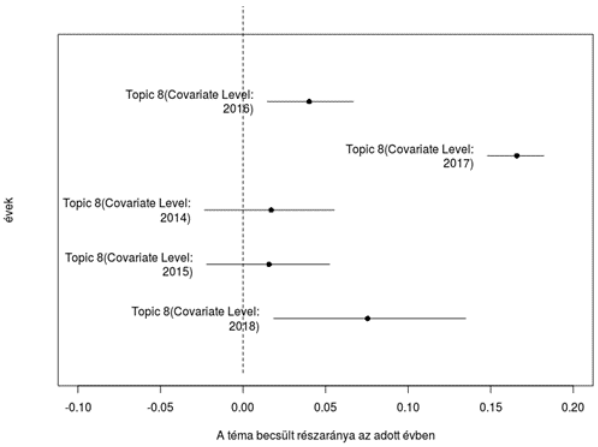
A bejegyzés éve, mint diszkrét kovariáns, különösen a 8. téma megjelenésére volt hatással. A catering és élelmiszerhigiénia a 2017-es évben jelent meg legnagyobb arányban, 15%-ot meghaladó mértékben, a hozzászólásokban. (Az ábrán a pont mutatja a becslés értékét, míg a jobb és bal oldalán látható vonalak a becslés konfidenciaintervallumát jelölik.) A hozzászólásokat áttekintve ebben az évben találtunk egy hosszú hozzászólásfolyamot, amelyben többen panaszkodtak hasmenésre, ételmérgezésre. Ennek az lehetett az oka, hogy a nyári hőségben megromlottak a sokszor a tűző napon tárolt élelmiszerek.

A lájkok (jelölések) száma és a témák várható részarányai közti kapcsolat a *6. ábrán* látható, amely megmutatja, hogy a lájkok számának függvényében hogyan alakul a kempingszolgáltatásokkal kapcsolatos téma megjelenése a hozzászólásokban. Az ábrán jól látható, hogy növekvő lájk-számok mellett nő(ne) a téma részaránya a hozzászólásokban, egyre szélesedő alsó és felső konfidenciaintervallumok mellett, azaz a becslés bizonytalansága nagy. Ennek a kovariánsnak megítélésünk szerint nincs nagy prediktív ereje. Ugyan vannak olyan témák, amelyek nagyobb mennyiségű lájkot generálnak, de a lájkok számának növekedése kevéssé magyarázza egy adott téma várható részarányának növekedését a hozzászólásokban.

5. ábra

Diszkrét kovariáns (a hozzászólás évének) hatása a 8. téma (catering és élelmiszerhigiéniával kapcsolatos vélemények) részarányára

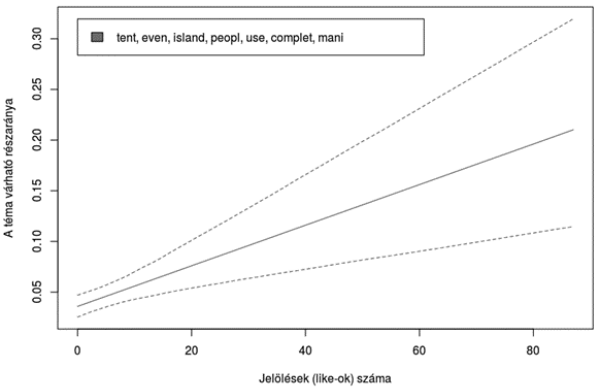
(Covariate level: a kovariáns szintje, amit a különböző évek reprezentálnak)



Forrás: saját szerkesztés az stm R programcsomag segítségével

6. ábra

Folytonos kovariáns (a bejegyzések lájkjainak száma) hatása az 1. téma (kempingszolgáltatásokkal kapcsolatos vélemények) részarányára



Forrás: saját szerkesztés az stm R programcsomag segítségével

6. Összefoglalás

Tanulmányunkban arra tettünk kísérletet, hogy egy számítógépes algoritmus, a rejtett Dirichlet-eloszlást alkalmazó strukturált témamodell (stm) segítségével határozzuk meg a Sziget Fesztivál látogatói által a Facebookon írt vélemények jellemző témáit, és ezeket összevetettükük egy korábbi kutatásunkban körvonalazott témákkal. A Sziget Fesztivál látogatóinak az elmúlt hét évben angol nyelven írt szöveges véleményei alapján az algoritmus segítségével kilenc témát modelleztünk, amelyek tartalma részben egyezett meg a korábbi, kvalitatív kutatásunkban azonosított témákkal.

A vizsgálat egyértelmű korlátját a korpusz, a látogatók által írt bejegyzések számossága jelentette. A számítógépes témamodellezés szempontjából egy kis korpuszt kívántunk gépi módszerekkel elemezni, azonban úgy tűnik, hogy ez részben a szemantikus koherencia rovására ment. Akárhány témával kísérleteztünk a futtatások során, az eredmények minősége, azaz az algoritmus által létrehozott témák (*szózsákok*) érthetősége keveset javult. A jövőben az ilyen jellegű vizsgálatokhoz valószínűleg lényegesen nagyobb korpuszra lesz szükség, azzal együtt is, hogy a kutatási előzmények azt jelzik, hogy az LDA módszertan esetében nem annyira az egyes dokumentumok hossza a kulcskérdés, hanem a korpusz nagysága: néhány száz közösségi média bejegyzés nem tekinthető jelentős adatállománynak. Ebben az értelemben vizsgálatunk csak részlegesen tekinthető sikeresnek, miközben más szempontból jól látható, hogy a látogatók közösségi média bejegyzései számítógépes módszerekkel eredményesen vizsgálhatóak, így kvantitatív módszerrel is rá lehet mutatni arra, hogy melyek azok a tényezők, amelyek meghatározzák az elégedettségüket.

Felhasznált irodalom

AIROLDI, E. M. – BLEI, D. M. – EROSHEVA, E. A. – FIENBERG, S. E. (2014): Introduction to Mixed Membership Models and Methods. *Handbook of mixed membership models and their applications.* 100. pp. 3–14.

BALOGH K. (2015). *A látens Dirichlet allokáció társadalomtudományi alkalmazása. A kuruc.info romaellenes megnyilvánulásainak tematikus elemzése.* Szakdolgozat, ELTE Társadalomtudományi Kar, mesterképzés. https://tas.precognox.com/labs/kuruc-info-visualization/A_latens_Dirichlet_allokacio_tarsadalomtudomanyi_alkalmazasa_Balogh_Kitti.pdf

BÍRÓ I. (2009): *Dokumentum osztályozás rejtett Dirichlet-allokációval.* PhD dolgozat. Eötvös Lóránt Tudományegyetem, Informatikai Kar, Információtudományi Tanszék, Informatikai Doktori Iskola. http://www.tnkcs.inf.elte.hu/vedes/Biro_Istvan_Tezisek_hu.pdf

BÍRÓ, I. – SIKLÓSI, D. – SZABÓ, J. – BENCZÚR, A. (2009a): Linked latent dirichlet allocation in web spam filtering. *AIRWeb '09: Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web.* pp. 37–40. https://doi.org/10.1145/1531914.1531922

BÍRÓ, I. – SZABÓ, J. (2009b): Latent dirichlet allocation for automatic document categorization. In: Buntine, W. – Grobelnik, M. – Mladenić, D. – Shawe-Taylor, J. (eds): *Machine Learning and Knowledge Discovery in Databases.* ECML PKDD 2009. Lecture Notes in Computer Science. 5782. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04174-7_28

BLEI, D. M. – LAFFERTY, J. D. (2006): Correlated topic models. *Advances in neural information processing systems.* NIPS 18. pp. 147–154.

BLEI, D. M. – NG, A. Y. – JORDAN, M. I. (2003): Latent dirichlet allocation. *Journal of Machine Learning Research*. 3(Jan). pp. 993–1022.

CROSSLEY, S. – DASCALU, M. – McNAMARA, D. (2017): How important is size? An investigation of corpus size and meaning in both latent semantic analysis and latent Dirichlet allocation. *Proceedings of the 30th International Florida Artificial Intelligence Research Society (FLAIRS) Conference*.

CURRY, T. A. – FIX, M. P. (2019): May it please the twitterverse: The use of Twitter by state high court judges. *Journal of Information Technology & Politics*. 16(4). pp. 379–393. https://doi.org/10.1080/19331681.2019.1657048

EISENSTEIN, J. – AHMED, A. – XING, E. P. (2011): Sparse additive generative models of text. *Proceedings of the 28th International Conference on Machine Learning.* June 2011. Bellevue, WA, USA. pp. 1041–1048.

FISCHER-PREßLER, D. – SCHWEMMER, C. – FISCHBACH, K. (2019): Collective sense-making in times of crisis: Connecting terror management theory with Twitter user reactions to the Berlin terrorist attack. *Computers in Human Behavior.* 100. pp. 138–151. https://doi.org/10.1016/j.chb.2019.05.012

GERRISH, S. – BLEI, D. M. (2012): How they vote: Issue-adjusted models of legislative behavior. *Advances in neural information processing systems 25.* (NIPS 2012). pp. 2753–2761.

HINEK M. – KULCSÁR N. (2019): Fesztiválélmény a közösségi médiában: a Sziget Fesztivál példája. *Turizmus Bulletin.* 19(3). pp. 4–12.

KRESTEL, R. – FANKHAUSER, P. – NEJDL, W. (2009): Latent dirichlet allocation for tag recommendation. In: *Proceedings of the third ACM conference on Recommender systems.* pp. 61–68. https://doi.org/10.1145/1639714.1639726

PAUL, M. J. – DREDZE, M. (2015): SPRITE: Generalizing topic models with structured priors. *Transactions of the Association for Computational Linguistics*. 3. pp. 43–57. https://doi.org/10.1162/tacl_a_00121

R CORE TEAM (2020): *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

ROBERTS, M. E. – STEWART, B. M. – TINGLEY, D. (2019): stm: An R package for structural topic models. *Journal of Statistical Software*. 91(2). pp. 1–40. https://doi.org/10.18637/jss.v091.i02

RODRIGUEZ, M. Y. – STORER, H. (2020): A computational social science perspective on qualitative data exploration: Using topic models for the descriptive analysis of social media data. *Journal of Technology in Human Services*. 38(1). pp. 54–86. https://doi.org/10.1080/15228835.2019.1616350

SYED, S. – SPRUIT, M. (2017): Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation. *2017 IEEE International conference on data science and advanced analytics (DSAA).* Tokyo. pp. 165–174. https://doi.org/10.1109/DSAA.2017.61

WANG, Y. C. – BURKE, M. – KRAUT, R. E. (2013): Gender, topic, and audience response: an analysis of user-generated content on Facebook. *Proceedings of the SIGCHI Conference on Human Factors in Computing* Systems. Association for Computing Machinery, New York, NY, USA. pp. 31–34. https://doi.org/10.1145/2470654.2470659

WEINSHALL, D. – LEVI, G. – HANUKAEV, D. (2013): LDA topic model with soft assignment of descriptors to words. *Proceedings of the 30th International Conference on Machine Learning.* Atlanta, Georgia, USA. JMLR: W&CP 28. pp. 711–719.

WILSON, A. – CHEW, P. A. (2010): Term weighting schemes for latent dirichlet allocation. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.* Los Angeles, California. pp. 465–473.