Fürdőlátogatók véleményeinek témamodellezése a Gellért Gyógyfürdő és Uszoda példáján

Hinek Mátyás
Budapesti Metropolitan Egyetem

A TANULMÁNY CÉLJA

A tanulmány a Gellért Gyógyfürdő és Uszoda 2004 és 2021 között a fürdő látogatói által írt vendégvélemények számítógépes topikmodellezésének eredményeit ismerteti. A turizmusmarketing szempontjából kiemelten fontos a vendégvélemények elemzése, különös tekintettel a jelentős forgalmat bonyolító attrakciókra. A Gellért Budapest, illetve Magyarország ikonikus műemlékfürdője, olyan egészségturisztikai attrakció, amely nagyon sok Budapestre látogató számára jelent kihagyhatatlan élményt. Ezt jól tükrözi az elmúlt másfél évtizedben a Tripadvisoron íródott több mint tízezer, közel harminc nyelven írt vendégbejegyzés, amely olyan óriási tömegű, hogy átfogó érétkelésére csak gépi módszerekkel van lehetőség.

ALKALMAZOTT MÓDSZERTAN

A Tripadvisoron 2005 és 2021 között írt összes véleményt egy erre a célra írt kisalkalmazás segítségével töltöttük le. A nem angol nyelven íródott véleményeket a Google Fordító segítségével angolra fordítottuk. Az így kapott korpuszt a szövegbányászat egyik gyorsan fejlődő módszerével, a látens Dirichlet allokációt alkalmazó strukturált témamodellezés segítségével vizsgáltuk meg, hogy melyek a több vendégvéleményben jellemzően előforduló témák. Ehhez az R statisztikai szoftvert, illetve az R környezetben futó STM strukturált témamodellező alkalmazást alkalmaztuk.

LEGFONTOSABB EREDMÉNYEK

A modellezés során a vendégvéleményekben 12 jellemző témát azonosítottuk. Ezeket összevetettük egy korábbi, ugyanezt a korpuszt vizsgáló, a szavak gyakorisági vizsgálatára épülő elemzéssel is, amelyből kiderült, hogy mindkét módszertan alapján hasonló témák azonosíthatók. Külön is elemeztük a vendégek által jórészt negatívan értékelt szolgáltatási jellemzőkkel kapcsolatos vélemények reprezentációját a vizsgált időhorizonton. Eszerint a higiéniával, tisztasággal kapcsolatos téma aránya nőtt, míg a fürdő vendégkommunikációjával kapcsolatos, szintén jórészt negatív megítélésű téma részaránya csökkent az írott a Gellért fürdőről írt vendégvéleményekben.

Kulcsszavak: számítógépes szövegfeldolgozás, témamodellezés, látens Dirichlet eloszlás, Gellért fürdő

DOI: 10.15170/MM.2021.55.04.03

BEVEZETÉS
INTRODUCTION

A látens Dirchlet eloszlásra (latent Dirichlet allocation – LDA) épülő témamodellezés egy dokumentumgyűjtemény, például egy turisztikai attrakció vendégvéleményeinek egyes dokumentumaira (bejegyzéseire) jellemző valószínűsíthető témák, illetve a témákat alkotó legvalószínűbb szavak gépi azonosítását jelenti. Az LDA arra hipotézisre épül, hogy a szerzők az egyes dokumentumokat konkrét témákat szem előtt tartva írják meg. Mivel a témákat specifikus szavak jellemzik, így a rájuk jellemző szókincs alapján a korpuszt és az egyes dokumentumokat alkotó témák visszafejthetők.

Az LDA lényegében egy gépi tanulási algoritmus, amely a témákat, illetve azok keverékét, mint látens információkat azonosítja a dokumentumokban, meghatározva, hogy az egyes témákhoz milyen szavak, mekkora valószínűséggel tartoznak. Az LDA nem felügyelt algoritmus, a témák azonosításához nem alkalmaz előre elkészített szó- vagy fogalomtárat. Az azonosítási eljárás többszöri iterációval történik, ennek során a kiinduló témaeloszlások és az azokat alkotó szavak eloszlásai folyamatosan változnak.

Az LDA algoritmusa nem veszi figyelembe a szavak pozícióját a dokumentumokban, és a dokumentumok sorrendjét sem. Emiatt az LDA-t gyakran „szózsák" (bag of word) modellként is jellemzik. Emellett 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 (Blei 2012, Blei *et al*. 2003).

AZ LDA ALGORITMUS
THE LDA ALGORITHM

Az LDA algoritmus a korpuszt alkotó dokumentumokban előforduló témák keverékét, mint valószínűségi eloszlást határozza meg, majd mindegyik témához hozzárendeli a témára jellemző, ugyancsak valószínűségi eloszlást követő kifejezéseket (szavakat).

Az LDA azt feltételezi, hogy egy dokumentumgyűjtemény (korpusz) minden dokumentuma *K* számú rejtett téma valószínűségi eloszlásaként reprezentálható, illetve minden egyes téma a dokumentumok szókincsét alkotó szavak multinomiális eloszlásaként határozható meg. A modellalkotás során előre megadjuk *K* értékét, amelyet korábbi futtatások eredményei alapján vagy különféle illeszkedési mutatók segítségével választunk ki.

A vizsgált korpusz *M* darab dokumentumot tartalmaz. Minden dokumentum N_i számú szóból áll, $w_{d,n}$ a *d.* dokumentum *n.* szava. Az egyes témák multinomális eloszlásként értelmezhetők a dokumentumokat alkotó szavak halmazán. V (vocabulary – szótár) jelöli a korpusz egyedi szavait, *Z* jelöli azt, hogy dokumentumot alkotó szavak (szópozíciók a dokumentumokban) melyik témához tartoznak, így $z_{d,n}$ mutatja, hogy a *d.* dokumentum *n.* szava melyik témához tartozik (Blei *et al.* 2003).

A modellben egyetlen megfigyelt változó van, a korpuszt és a dokumentumokat alkotó szavak. A modellből kikövetkeztetni kívánt rejtett változók:

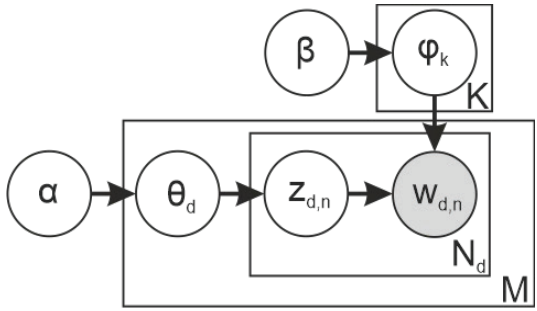
- Θ_d, a *d* dokumentumot alkotó témák valószínűségi eloszlása,
- φ_k a *k* témához tartozó szavak valószínűségi eloszlása,
- $z_{d,n}$, „témacímke", ami azt jelöli, hogy a dokumentumok egyes szavai melyik témába sorolhatók.

Az LDA generatív folyamata a következő (Balogh 2015):

1. Minden *k* téma esetében válasszunk egy $\varphi_k \sim Dir(\alpha)$ szóeloszlást.
2. Minden dokumentum esetében
   a. válasszunk egy $\Theta_d \sim Dir(\alpha)$ témaeloszlást.
   b. Minden egyes szó esetében
      i. válasszunk egy $Z_{d,n} \sim Mult(\Theta_d)$ téma hozzárendelést, ahol a $Z_{d,n} \in 1, \ldots, K$;
      ii. válasszunk egy $w_{d,n} \sim Mult(\varphi_{Zd,n})$ szót, ahol $w_{d,n} \in 1, \ldots, V$.

Az α és a β priori (kezdeti) Dirichlet hiperparaméterek, értéküket jellemzően alacsonyan adjuk meg, arra a feltételezésre építve, hogy a témák száma és az egyes témákhoz tartozó szavak száma is korlátozott. Az LDA grafikus reprezentációját az 1. ábra mutatja.

1. ábra: Az LDA témamodell változói között kapcsolat tálca reprezentáció segítségével
Figure 1. Relationship between the variables in the LDA topic model using plate representation



Megjegyzés: A külső tálca a korpuszt alkotó dokumentumok ismétlődő kiválasztását, a belső tálca a szavak ismétlődő kiválasztását jelenti a dokumentumban.
Forrás: Balogh 2015, Blei et al. 2003

Az 1. ábra tálcái által szemléltetett ismétlődő kiválasztások során az egyes dokumentumok témamixe (θ_d) alapján a dokumentumot alkotó szavak ($w_{d,n}$) besorolásra kerülnek valamelyik témába. A zd,n jelzi, hogy a dokumentumokat alkotó szavak mely témacímkéket kapják. Ezzel párhuzamosan a dokumentumokban megfigyelhető szavak témákba (K) sorolása is megtörténik, ahol φk az adott témákhoz tartozó szavak valószínűségi vektora. Utóbbi besorolás vegyes természetű, egy konkrét szó több témához is tartozhat, eltérő valószínűséggel. A becslési eljárás során meg kell adni a kezdeti téma- (α), és szóeloszlások (β) paramétereit.

Az optimalizációs eljárás elvégzéséhez többféle módszer is alkalmazható, de a leggyakrabban a Gibbs-mintavételt alkalmazzák, mivel jól algoritmizálható. (Ismertetésétől jelen vizsgálat során eltekintünk.) Többszöri iterációt követően az egyes valószínűségi eloszlások poszteriorait kapjuk eredményül, azaz az egyes témákat alkotó szavak valószínűségeit, valamint az egyes dokumentumok téma-mixeit és a dokumentumok egyes szavainak (a korpusz szótárát alkotó szavak szópozícióinak) témacímkéit.

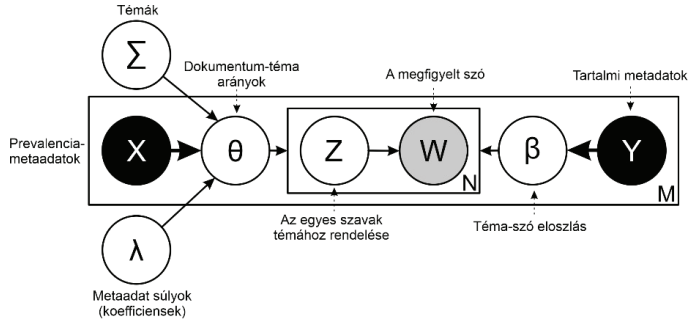
A STRUKTURÁLIS TÉMAMODELL (STM)
STRUCTURAL TOPIC MODEL (STM)

A témamodellek teljesítménynek javítása érdekében, többen javasolták az eredeti modell továbbfejlesztését lásd pl. Gerrish & Blei 2012, Paul & Dredze 2012, Weinshall *et al.* 2013 stb. munkáit. Roberts és szerzőtársai (2013) dolgozták ki a strukturált témamodellt (STM), mely olyan szövegek feldolgozására is alkalmas, amelyhez a rendelkezésre állnak különféle dokumentumszintű metaadatok. Ezek a metaadatok tetszőlegesek lehetnek, pl. a vendégvélemények esetén a vélemény keletkezésének időpontja, a véleményt író nemzetisége, a vélemény nyelve, esetleg a szolgáltatás minősítésének számszerű értéke. A metaadatok (az STM terminológia szerint: kovariánsok) lehetnek folytonosak (pl. a hozzászólás dátuma), vagy diszkrét jellegűek (pl. a szolgáltatás minősítése az ötfokozatú skálán).

A dokumentumszintű kovariánsok egy egyszerű általánosított lineáris modell révén épülhetnek be a modell előzetes téma és szóeloszlásaiba. Ezzel az STM lehetővé teszi a kutató számára, hogy a dokumentum-téma arányok (θd) és az érdeklődésére számot tartó kovariáns (X), összefüggéseit vizsgálja. További tartalmi kovariáns(ok) (Y) beépítése révén, a téma-szó prior eloszlása is módosítható, azaz az is vizsgálható, hogy egyes kovariánsok hogyan befolyásolják a témák szóhasználatát (2. ábra).

A strukturált témamodell segítségével a kovariánsok együttes hatása is vizsgálható, az így kapott eredmények felhasználhatók a metaadatok és a témák kapcsolatait vizsgáló hipotézisek tesztelésére (Roberts *et al.* 2013, 2016).

2. ábra: A strukturált témamodell (STM) tálcadiagrammja
Figure 2. Plate diagram of the Structured Topic Model (STM)



Forrás: Park & Ha, 2017, Roberts et al. 2013, 2016

A VIZSGÁLAT ELŐZMÉNYEI
RESEARCH BACKGROUND

A nemzetközi szakirodalomban az LDA, illetve strukturált témamodell (STM) alkalmazására számos példa található. Ugyanakkor turizmuskutatás területén az LDA-n alapuló témamodellek csak az elmúlt években nyertek nagyobb népszerűséget.

- Park és Ha (2017) egy előkelő Las Vegas-i szálloda Tripadvisoron olvasható negatív vendégvéleményeit elemezték. Először szentimentelemzést végeztek, és erre támaszkodva kiemelték a negatív véleményeket a vélemények teljes halmazából, amelyeket témamodellezéssel vizsgáltak tovább. Összesen 11 negatív töltetű témát azonosítottak, köztük a vendégeket zavaró zajt, a kommunikáció hiányosságait, a személyzet által nyújtott szolgáltatások hiányosságait, a dohányfüstöt, az étkezéssel kapcsolatos problémákat stb. Azt is megvizsgálták, hogy az azonosított témáknak hogyan alakultak a valószínűsíthető részarányai az idő múlásával.
- Calheiros és szerzőtársai (2017) egy portugál ökoszálloda 400 látogatójának vendégbejegyzéseit értékelték szótár alapú szentimentelemzéssel, ezt követően a vélemények jellemző témáit látens Dirichlet allokációt alkalmazó témamodellezés segítségével azonosították. A strukturálatlan adatokból a szálloda menedzsmentje számára is hasznosítható információkat sikerült kinyerni.
- Hu és szerzőtársai (2019) 315 New York-i szálloda Tripadvisoron olvasható 28 ezer vendégértékelése alapján strukturált témamodellel vizsgálták a szállodai vendégek elégedetlenségének forrásait, illetve azt, hogy a különböző kategóriájú szállodák között van-e különbség a panaszok okai között. A szállodák minőségi különbségeit, valamint a negatív véleményét jelző változókat kovariánsokként építették be modellükbe. Eredményeik azt mutatták, hogy az alacsony kategóriájú szállodák esetében a létesítménnyel kapcsolatos problémák, míg a magas kategóriájú szállodák esetében a szolgáltatással kapcsolatos problémák és az árképzés a fogyasztói elégedetlenség fő forrásai.
- Korfiatis és szerzőtársai (2019) európai légitársaságok Tripadvisoron olvasható több mint félmillió utasvéleménye alapján végzett strukturált témamodellezést, amelyet további statisztikai eljárásokkal elemeztek tovább, így egyebek mellett vizsgálták a témaeloszlást a véleményezők által adott számszerű értékelések függvényében, és vizsgálták az egyes témák prevalenciájának időbeli változását. Ezen felül korrespondenciaanalízis segítségével több dimenzióban vizsgálták a kontinensen belüli és az interkontinentális járatokat üzemeltető légitársaságok vendégvélemények alapján kirajzolódó profiljait, rámutatva egymáshoz viszonyított versenypozíciójukra.
- Sutherland és szerzőtársai (2020) dél-korai szálláshelyek 104 ezer vendégvéleménye alapján végeztek el strukturált tematikus témamodellezést. A vizsgálat során 14 témát azono-

sítottak, amelyek a szálláshely elhelyezkedése (vidéki vagy városi), és a szálláshely típusa, mint kovariánsok alapján mutattak változékonyságot.

- Kirilenko és szerzőtársai (2021) három ország turisztikai attrakcióinak vendégvéleményei alapján végeztek el témamodellezést. A kínai Terrakotta Hadsereg Múzeum vendégvéleményeinek elemzése (majd ennek validálása további két ország egy-egy attrakciójának vendégvéleményei alapján) arra a megállapításra jutott, hogy a negatív vélemények azonosítása nem felügyelt szövegbányászati algoritmusok segítségével problémás. Szinte az összes negatív értékelés nagyon kicsi valószínűséggel tartozott a témamodellezés egy adott témájához, kivéve néhány gyakori témát. Ezzel szemben a pozitív vendégvélemények elemzése tekintetében az LDA algoritmus jó teljesítményt mutatott.

Vizsgálatunk közvetlen előzménye Smith és szerzőtársai (2020) által elvégzett vizsgálat a Gellért-fürdővel kapcsolatos vendégelégedettségről a Tripadvisoron olvasható vendég-bejegyzések alapján. A szerzők a vendégvéleményekben előforduló témák azonosítására a KH Coder szövegbányászati alkalmazást használták, amely a szavak együttes előfordulását vizsgálja úgy, hogy a gyakran használt szavakból képez hálózatot. A hálózat csúcspontjainak mérete a korpuszt alkotó szavak gyakoriságától függ, a csúcspontokat összekötő élek, illetve ezek vastagsága azt jelzi, hogy az egyes szavak mely más szavakkal, milyen gyakran fordulnak elő együtt egy szövegegységben (egy vendégvéleményben). A felvázolt hálózat alapján a szerzők összesen 11 témát azonosítottak: 1. keretezés, azaz vendég személyes története a látogatásról, 2. épület, 3. bejárat és átöltözés, 4. úszósapka-használat, 5. törölköző, 6. alapszolgáltatások, a medencék és a szauna, 7. kiegészítő szolgáltatások, elsősorban a masszázskezelések, 8. higiénia és tisztaság, 9. személyzet, 10. látogatók, valamint 11. további általánosabb tapasztalatok. A vizsgálatot kiegészítették azzal, hogy az 1-5-ös skálán adott vendégértékelések egyes skálatételeihez tartozó értékeléseket is megvizsgálták, és arra jutottak, hogy legelégedettebb, 4-5-ös értékelést adott látogatók véleményei körében a pihenés, és a gyönyörű épület jelenik meg elsősorban, a 3-as értékelések esetében az épületbe belépést követően megfigyelhető elégedettség csökkenés jelenik meg, míg a legrosszabb minősítésű értékelések körében a törölköző kölcsönzés bérleti díja, a higiénia, és a személyzettel kapcsolatos negatív megjegyzések voltak a meghatározók (Smith *et al.* 2020).

A VIZSGÁLAT CÉLJA ÉS MÓDSZEREI
AIMS AND METHODS OF THE STUDY

Vizsgálatunk során három kérdésre kerestük a választ:

1. A látens Dirichlet allokációra épülő számítógépes témamodellezés alapján azonosíthatók-e koherens, és értelemezhető témák a Gellért-fürdő vendégeinek Tripadvisoron írt alapján, valamint ezek mennyire esnek egybe Smith és szerzőtársai (2020) által lefolytatott szövegbányászati vizsgálat során azonosított témákkal?
2. Kíváncsiak voltunk arra, hogy hogy egyes témák reprezentációja idővel hogyan változik, különös tekintettel arra, hogy a Tripadvisoron 2004-ig visszamenőleg olvashatók vendégek által értékelések a Gellért fürdő szolgáltatásairól.
3. Végül, de nem utolsó sorban kíváncsiak voltunk arra is, hogy hogyan alakul a témák prevalenciája, a hozzászólók által adott számszerű értékelések (1-5-ös skála) különböző szintjénél, illetve mutatkozik-e különbség a negatív és a pozitív értékelések esetén?

A Gellért fürdő vendégvéleményeit a Tripadvisor oldalairól töltöttük le, egy kifejezetten erre a célra készült kisalkalmazás segítségével. Összesen 10.692 nyilvánosan elérhető vendégvéleményt találtunk, amelyek 2004 és 2020 között születtek. A 2010 előtti időszakból igen kevés, mindössze néhány tucat véleményt tárolt a rendszer. Tömegesen, évente többszázas, majd többezres nagyságrendben 2012 után jelentek meg recenziók a Tripadvisoron. A legtöbb bejegyzés 2016-ban született, közel 2500 db, ezt követően azonban évről-évre csökkent a szöveges vendégvélemények száma, 2019-ben már csak a „csúcsév" kevesebb, mint fele volt a recenziók száma.1

2020-ra pedig szinte elfogytak a vélemények, ami a koronavírus járvány hatásának köszönhető, ugyanis ebben az évben a Gellért fürdő a lezárások miatt sokáig kényszerűen nem tudott vendégeket fogadni.

A közel 11 ezer bejegyzés 26 nyelven született, majdnem fele angolul, közel ötödük olaszul, tizedük franciául, öt százalékuk oroszul, négy százalékuk spanyolul. Minden más nyelv (köztük svéd, német, holland, portugál, dán, norvég, kínai, arab, magyar stb.) részaránya a három százalékot sem érte el. Ugyanakkor a bejegyzések nyelvek

szerinti megoszlása nem jó indikátora a látogatók származási ország szerinti megoszlásának, ugyanis a vendégek gyakran nem az anyanyelvükön, hanem angolul írtak véleményt.

Az elemezhetőség érdekében az összes vendégvéleményt angolra fordítottuk a Google Fordító segítségével. Noha a Google nem ad pontos fordítási eredményt, ám tekintettel arra, hogy az LDA „szózsák" modell, nem a kifejezések és mondatok nyelvtani koherenciája elsődleges, így a fordítás potenciális problémái önmagában nem lehetetlenítik el az elemzést. Az áttekintett kutatási előzményekben más kutatók is hasonlóan jártak el, egyebek mellett a Gellért fürdő vendégvéleményeinek korábbi vizsgálata során is számítógéppel fordították angolra a más nyelven íródott vendégvéleményeket (Kirilenko *et al.* 2021, Smith *et al.* 2020).

A nyelvi előfeldolgozáshoz és a témamodellezéshez az R (R Core Team 2021) STM csomagját használtuk (Roberts *et al.* 2019), az elemzést teljes egészében R-el folytattuk le. A szöveges adatok előfeldolgozása során a korpuszból eltávolítottuk az írásjeleket, számokat, valamint a gyakran előforduló, de a vizsgálat szempontjából jelentést nem hordozó szavakat, azaz a „stopszavakat" (pl. névelőket, segédigéket, határozószavalat, névmásokat stb.). Mivel ugyanannak a szónak többféle alakja is előfordulhat, ami torzíthatja az elemzést, így a szavakat lemmatizáltuk, azaz a ragozott és a képzett alakok szótöveit hagytuk meg.

Ezt követően az elemzés pontosságának javítása érdekében a ritkán előforduló szavakat is kizártuk, csak azokat tartottuk meg, amelyek legalább három dokumentumban előfordultak. Hasonlóképpen, a sok dokumentumban előforduló igen gyakori szavakat is érdemes elhagyni (jelen kutatásban ilyen szavak pl. a „Budapest" és a „Gellért" szavak), így úgy döntöttünk, hogy azokat a nagyon gyakori szavakat, amelyek a dokumentumok legalább tíz százalékában, konkrétan ezer véleményben egyaránt előfordulnak, kihagyjuk az elemzésből. Az előkészítési lépéseket követően 10.689 dokumentumot és 4.059 egyedi kifejezést (lemmatizált szót) tartalmazó szótárunk maradt a témamodellezés elvégzésére.

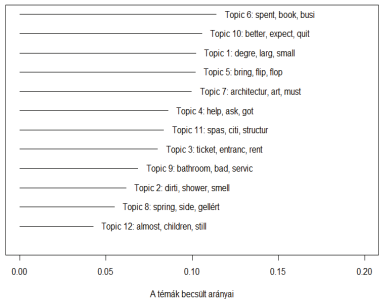
Az STM R csomag lehetőségeit kihasználva két dokumentumszintű metaadatot vontunk be az elemzésbe: a bejegyzés dátumát, mint folytonos kovariánst, valamint a vendégek által adott számszerű értékelést az 1-5-ös skálán, mint diszkrét kovariánst.

A látens Dirichlet allokációt alkalmazó témamodellezés esetén a témák számát (K) előzetesen kell meghatározni. Ehhez két mutatót vizsgáltunk meg: a szemantikus koherenciát, amely megméri az adott témához nagyobb valószínűséggel tartozó szavak szemantikai hasonlóságát, és a held-out likelihood-ot, amely azt méri, hogy mennyire általánosítható a modell. A mutatók magasabb értéke azt jelzi, hogy a témák azonosíthatósága és így a modell prediktív ereje javul. Együttesen a két mutató alapján a legmagasabb értékek 9, 10, és 12 téma esetében adódtak, így a kutatási előzményeket is figyelembe véve úgy döntöttünk, hogy 12 témával folytatjuk le a modellezést.

EREDMÉNYEK
RESULTS

Az STM algoritmust lefuttatva megkaptuk, hogy melyek az egyes témákhoz tartozó legvalószínűbb szavak, valamint mekkora a témák becsült részaránya az egyes hozzászólásokban és a korpusz egészében (3. ábra). A látogatói véleményekben a legnagyobb becsült gyakorisággal a 6., 10., 1., 5., és 7. téma jelenik meg, részarányuk meghaladja a 10%-ot, míg a további témák 4-8%-os valószínűséggel fordulnak elő.

3. ábra: A témák becsült arányai és a legnagyobb valószínűséggel a témákhoz tartozó szavak
Figure 3. Estimated proportions of the topics and words most likely to be associated with the topic



Forrás: Saját szerkesztés az STM R programcsomag segítségével

A témák azonosítása a legvalószínűbb szavak (*Highest Prob.*) alapján kutatói feladat. Az azonosítást támogatja a *FREX* mutató, amely az általános gyakoriságuk és az adott témában kizárólagosságuk alapján súlyozza a szavakat. A *Lift* mutató a téma szavait úgy súlyozza, hogy osztja őket a más témákban előforduló gyakoriságukkal, ezzel nagyobb súlyt adva a más témákban ritkábban előforduló szavaknak. A *Score* szintén gyakoriságuk alapján súlyozza a témák szavait. Ezen felül az STM a témák azonosítását azzal is támogatja, hogy az egyes témák esetében megvizsgálhatók az egyes önálló dokumentumok (Tripadvisor hozzászólások), amelyekben az adott téma a legmagasabb valószínűsíthető részaránnyal van jelen. A szavakat és az azonosított témákat az 1. táblázat tartalmazza.

1. táblázat: A témákat reprezentáló lemmatizált szavak (a számítógépes eljárás outputja) és a kutató által azonosított témák

Table 1. The lemmatised words representing the topics (output of the algorithm) and the topics identified by the researcher)

Témák	A témákhoz tartozó legvalószínűbb szavak különféle mutatók szerint	Azonosított téma
1.	*Highest Prob:* degre, larg, small, open, winter, close, tub *FREX:* degre, winter, larg, street, possibl, headset, headphon *Lift:* regardless, snobbish, waterfal, pompous, divis, inescap, econom *Score:* degre, regardless, larg, winter, tank, tub, street	Medencék
2.	*Highest Prob:* dirti, shower, smell, disappoint, bad, shoe, floor *FREX:* dirti, dirt, disgust, filthi, smell, smelli, barefoot *Lift:* recomend, indec, gastronomi, rusti, unmotiv, plaster, greasi *Score:* dirti, recomend, smell, disgust, shoe, poor, filthi	Higiénia hiányosságai
3.	*Highest Prob:* ticket, entranc, rent, buy, pay, huf, deposit *FREX:* card, bracelet, ticket, cash, huf, forint, deposit *Lift:* abus, hilton, seller, sensor, logo, rfid, credit *Score:* ticket, entranc, bracelet, huf, deposit, abus, card	Belépők, árak
4.	*Highest Prob:* help, ask, got, didnt, back, told, minut *FREX:* told, member, ask, ladi, said, desk, couldnt *Lift:* trekk, lagoon, apologis, injuri, spous, shrug, enquir *Score:* didnt, trekk, ask, told, got, desk, help	Személyzet és vendégkommunikáció
5.	*Highest Prob:* bring, flip, flop, main, walk, way, coupl *FREX:* flop, flip, plung, coupl, option, easi, navig *Lift:* attendee, poll, instanc, pol, gellart, prudish, reccomend *Score:* flip, flop, bring, dont, plung, coupl	Elrendezés, belső útvonalak, tájékozódás
6.	*Highest Prob:* spent, book, busi, morn, stay, earli, afternoon *FREX:* food, spent, busi, earli, morn, book, cafe *Lift:* attende, cafebar, fish, fisherman, caferestaur, -star, ish *Score:* busi, book, spent, attende, food, morn, earli	Kellemes, pihentető időtöltés
7.	*Highest Prob:* architectur, art, must, wonder, atmospher, decor, mosaic *FREX:* art, nouveau, style, deco, uniqu, fantast, atmospher *Lift:* regener, detour, fairi, secret, exud, stimul, rowdi *Score:* regener, art, nouveau, style, mosaic, architectur, atmospher	Gyönyörű épület, atmoszféra és dekorációk
8.	Highest Prob: spring, side, gellért, bridg, buda, hill, danub FREX: bridg, buda, danub, hill, bell, spring, pest Lift: athmospher, societi, szabadsag, champagn, hid, recharg, king Score: spring, athmospher, bridg, danub, buda, side, hill	Helyszín jellemzői
9.	Highest Prob: bathroom, bad, servic, poor, tourist, lack, understand FREX: bathroom, indic, renov, attent, piti, lack, averag Lift: divin, ongo, perspect, nickel, compass, decrepit, dime Score: bathroom, divin, poor, bad, hygien, lack, servic	Gyenge szolgáltatás, nincs megfelelően felkészülve a látogatókra
10.	Highest Prob: better, expect, quit, expens, szechenyi, less, say FREX: szechenyi, compar, better, mayb, expect, ruda, széchenyi Lift: rope, strand, lesser, favour, grander, palatinus, stroller Score: szechenyi, rope, better, expect, quit, expens, disappoint	Budapest további fürdői
11.	Highest Prob: spas, citi, structur, locat, offer, histor, famous FREX: histor, histori, ancient, structur, citi, properti, famous Lift: breadth, reachabl, anderson, wes, bormio, priceless Score: structur, citi, vodicka, spas, histor, tank, ancient	Műemlékfürdő
12.	Highest Prob: almost, children, still, read, shower, famili, review FREX: children, almost, late, famili, negat, happen, jump Lift: merit, stove, hub, trace, smoke, squeamish, unwelcom Score: merit, children, almost, famili, late, read, negat	Gyerekek, családok és más látogatók

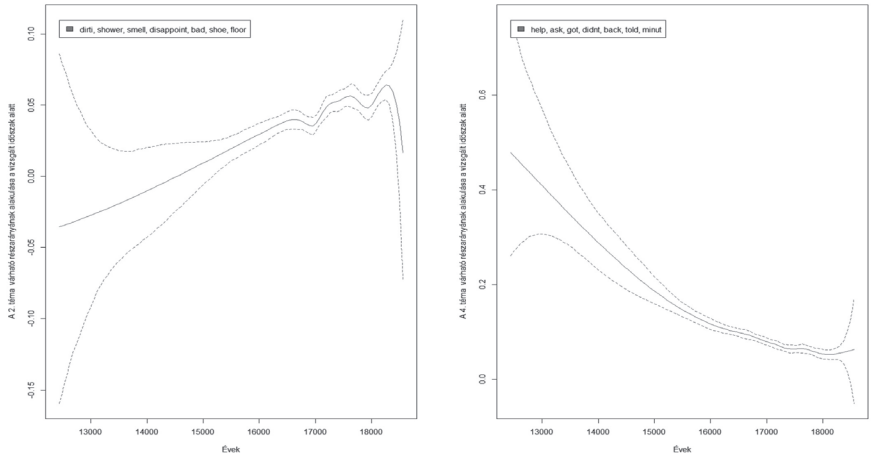
Forrás: Saját szerkesztés az STM R programcsomag segítségével

A tizenkét azonosított téma részben megfeleltethető Smith és szerzőtársai (2020) által azonosított témáknak, lásd pl. az épület, az alapszolgáltatások, a higiénia, a személyzet, a bejárat és átöltözés (a fürdő elrendezése), vagy a többi látogató, de akár a keretezés témáját. Budapest további fürdőit az STM azonosította témaként, a KH Coder nem, ahogy az árak is hangsúlyosan, önálló témaként jelennek meg jelen modellben. Összességében a témák kétharmada azonos a két vizsgálatban, ami jól jelzi mindkét modell alkalmasságát. Ez azonban nem jelenti azt, hogy a kiemelt szavak, és azok összefüggései ugyanazok lennének, a korábbi kutatásban azonosított csomópontok és kapcsolatok nem jelennek meg az LDA algoritmus eredményeiben. Ez részben az eltérő módszertannak is köszönhető, az egyező témák pedig annak, hogy egy fürdő vendégrecenzióiban „nyilvánvaló" témák megjelenése várható, így pl. a higiénia vagy a személyzet vendégekkel történő kommunikációja, a beléptetés, az árazás stb.

A témaprevalenciákat a kovariánsok alapján is megvizsgáltuk. A strukturált témamodellben a folytonos metaadat a bejegyzés dátuma volt, segítségével megvizsgálható, hogy hogyan változik egyes témák reprezentációja az idő múlásával. Az elemzés során a létesítmény menedzsmentje szempontjából is érdekes témák várható előfordulásának változását vizsgáltuk meg, így a 2. témát, a higiéniával kapcsolatos panaszokat, valamint a 4. témát, a vendégekkel történő kommunikációt.

A higiéniával kapcsolatos panaszok előfordulása 2005 és 2019 között nőtt, majd ezután csökkent, ami vélhetően azzal függ össze, hogy a koronavírusválság időszakában a létesítmény látogatottsága visszaesett. Bár részaránya a recenziókban nem volt magas, mindössze 5-10%, de a téma fokozatosan vált egyre relevánsabbá a vizsgált időszakban (4. ábra, 1. panel).

4. ábra: A 2. (higiénia) és 4. téma (vendégkommunikáció) valószínűsíthető arányának alakulása a hozzászólásokban 2005 és 2020 között +/- 5% konfidenciaintervallumok mellett
Figure 4. Expected prevalence of topic 2 (hygiene) and topic 4 (guest communication) in the guest reviews between 2005 and 2020 with +/- 5% confidence intervals)



Megjegyzés: Az STM a dátum formátumú adatokat nem tudta kezelni. A dátumok természetes számokká alakítását követően a kezdőnap 1970.01.01. A panelek vízszintes tengelyén az azóta eltelt napok száma látható, így 13000 = ***2005.****08.05; 14000 =* ***2008.****05.01; 15000 =* ***2011.****01.26; 16000 =* ***2013.****10.22; 17000 =* ***2016.****07.18; 18000 =* ***2019.****04.14*
Forrás: Saját szerkesztés az STM R programcsomag segítségével

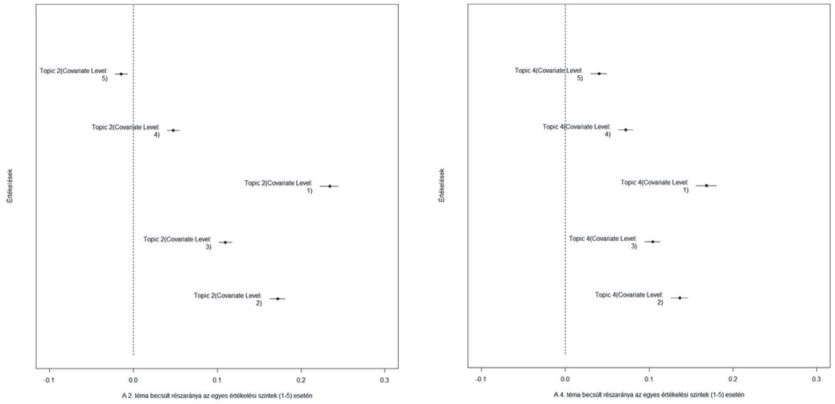
Látványosabb a vendégkommunikáció témakörének reprezentációja, amely szintén negatív kontextusban került elő több olyan recenzióban, amelyben ennek a témának a reprezentációja a legnagyobb volt. A vizsgált időszak elején – amikor a hozzászólások száma alacsony volt – közel 50%-ban determinálta a recenziókat, míg 2019-2020 időszakában már „csak" 20 százalék alatti a téma részaránya a hozzászólásokban, ami arra utal, hogy javult a vendégkommunikáció minősége helyzet a fürdőben (4. ábra, 2. panel).

A 2. és a 4. téma reprezentációját az 1-5-ös skála, mint értékelési szintek esetén is megvizsgáltuk. Az STM módszertanban ez az ún. pontbecslés, ami egy adott kovariáns meghatározott értéke mellett mutatja be a téma valószínűsíthető részarányát. Nem meglepetés, hogy a 2. téma, a higiéniával kapcsolat panaszok reprezentációja magasabb az alacsonyabb vendégértékelések körében. Az 1-es értékelésű vendégvélemények körében a higiénia prevalenciája majdnem 25%-os arányú, míg a magas, 4-5-ös értékelésű vendégvélemények esetében gyakorlatilag nulla a reprezentációja (5. ábra, 1. panel).

Hasonló, de lényegesen kisebb mértékű variabilitás a vendégkommunikáció (4. téma) esetében is megfigyelhető, itt az 1-es értékelésű recenziók esetében a téma várható részaránya közel 20%, míg a legmagasabb, 5-ös értékelések esetében mindössze 5% körül alakul (5. ábra, 2. panel).

Ezek az eredmények szintén összevethetők a korábbi kutatással, Smith *et al.* (2020) hasonló eredményekre jutottak.

5. ábra: A 2. (higiénia) és 4. (vendégkommunikáció) témák reprezentációja a hozzászólásokban a számszerű vendégértékelések (1-5-ös skála) egyes szintjein
Figure 5. Expected prevalence of topic 2 (hygiene) and 4 (guest communication) in the guest reviews at each level of the numerical guest ratings (1-5 scale)



Forrás: Saját szerkesztés az STM R programcsomag segítségével

A KUTATÁS KORLÁTAI
LIMITATIONS OF THE RESEARCH

Technikai, módszertani és terjedelmi okokból nem volt arra mód, hogy bővebb terjedelemben vizsgáljuk a témák és a lehetséges kovarianciák kapcsolatát, így a továbbfejlesztés egyik releváns iránya ez. A vizsgálat további lehetséges kiterjesztése Budapest összes fürdőjének és akár jelentősebb attrakcióinak együttes vizsgálata az internetes vendégvélemények alapján, rámutatva a vendégek elvárásaira, a szolgáltatási szűk keresztmeszetekre és a fejlesztés lehetőségeire.

KÖVETKEZTETÉSEK ÉS JAVASLATOK
CONCLUSIONS AND PROPOSALS

Eredményeink azt mutatják, hogy az LDA által gépi tanulási algoritmus segítségével generált témák szemantikailag is koherensek, ténylegesen a vizsgált szolgáltatás fontosabb dimenzióira, a vendégélményt meghatározó tényezőkre mutatnak rá. Emellett a vizsgálat eredményei jelentős részben átfedésben vannak a szavak együttes előfordulásának gyakorisága alapján képzett témákkal, sőt a számszerű vendégértékelések esetében is hasonló eredményekre jutott a két vizsgálat. Ez azonban nem jelenti azt, hogy a különböző szövegbányászati algoritmusok eredményei teljesen összecsengenek, más és más szavak alapján körvonalazódtak a témák az együttes gyakorisági vizsgálatban és a strukturált témamodellben. Eredményeink alapján – ugyan messze nem teljeskörű – de a már a létesítmény menedzsmentje szempontjából is megfontolható javaslatot tudunk megfogalmazni. Eszerint figyelni kell a létesítmény állapotára, tisztaságára, a higiénia helyzetére, ugyanis a téma reprezentációja az legutolsó időszakot leszámítva folyamatosan nőtt a vendégértékelésekben.

ÖSSZEFOGLALÁS
SUMMARY

Tanulmányunkban arra tettünk kísérletet, hogy a rejtett Dirichlet eloszlást alkalmazó strukturált témamodell (stm) segítségével határozzuk meg a Gellért fürdő Tripadvisoron olvasható vendégvéleményeinek témáit, és ezeket összevessük egy korábbi kutatásban, más módszerrel felvázolt témastruktúrával. Tizenkét releváns témát tudtunk azonosítani, amely megfeleltethető egy korábbi kutatás eredményeként generált témákkal. Elemeztük azt is, hogy a témák reprezentációja hogyan változott a vizsgált időszakban, 2004-2020 között, valamint megvizsgáltuk, hogy a szöveges recenziókhoz tartozó számszerű értékelések esetében mely témák megjelenése volt valószínűbb.

A témaprevalenciák időbeli alakulásának változása a megosztóbb tartalmú témák (higiénia, vendégkommunikáció) esetében volt megfigyelhető. A számszerű értékelések és a szöveges recenziók kapcsolata arra enged következtetni, hogy a higiénia az egyik legfontosabb csalódást keltő tényező a vendégek körében, az alacsony értékelések esetén a téma reprezentációja jelentősen magasabb volt a vendégértékelésekben és a téma részaránya folyamatosan nőtt a vizsgált időhorizonton. A vendégkommunikáció esetében is megfigyelhető, hogy az alacsonyabb vendégértékelések esetében nagyobb a téma részaránya, de az idő előre haladtával a téma egyre kisebb arányban jelent meg a szöveges értékelések körében.

[1] Hasonló tendenciát tapasztaltunk korábbi kutatásunk során a Sziget Fesztivál Facebookon megjelenő vendégvéleményei esetében (Hinek 2021).

HIVATKOZÁSOK
REFERENCES

Balogh K. (2015). A *látens Dirichlet allokáció társadalomtudományi alkalmazása* [ELTE Társadalomtudományi Kar]. https://tas.precognox.com/labs/kuruc-info-visualization/A_latens_Dirichlet_allokacio_tarsadalomtudomanyi_alkalmazasa_Balogh_Kitti.pdf

Blei, D. M. (2012), Probabilistic topic models. *Communications of the ACM,* 55(4), 77–84. DOI: 10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003), Latent Dirichlet Allocation. *Journal of Machine Learning Research,* 3(4–5), 993–1022. DOI: 10.1162/JMLR.2003.3.4-5.993

Calheiros, A. C., Moro, S., & Rita, P. (2017), Sentiment Classification of Consumer-Generated Online Reviews Using Topic Modeling. *Journal of Hospitality Marketing & Management,* 26(7), 675–693. DOI: 10.1080/19368623.2017.1310075

Gerrish, S., & Blei, D. M. (2012), How They Vote: Issue-Adjusted Models of Legislative Behavior. *Advances in Neural Information Processing Systems,* 25, 2753–2761. https://proceedings.neurips.cc/paper/2012/hash/193002e668758ea9762904da1a22337c-Abstract.html

Hinek M. (2021), Fesztiválátogatók véleményeinek számítógéppel támogatott tematikus modellezése – egy kísérlet eredményei Computer-aided topic modelling based on festival-goers' opinions – results of an experiment. *Turizmus Bulletin,* 21(1), 4–12. DOI: 10.14267/TURBULL.2021v21n1.1

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

Kirilenko, A. P., Stepchenkova, S. O., & Dai, X. (2021), Automated topic modeling of tourist reviews: Does the Anna Karenina principle apply? *Tourism Management,* 83, 104241. DOI: 10.1016/j.tourman.2020.104241

Korfiatis, N., Stamolampros, P., Kourouthanassis, P., & Sagiadinos, V. (2019), Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews. *Expert Systems with Applications,* 116, 472–486. DOI: 10.1016/j.eswa.2018.09.037

Park, K., & Ha, S. H. (2017), Customer Service Evaluation based on Online Text Analytics: Sentiment Analysis and Structural Topic Modeling. *The Journal of Information Systems,* 26(4), 327–353. DOI: 10.5859/KAIS.2017.26.4.327

Paul, M. J., & Dredze, M. (2014), A Model for Mining Public Health Topics from Twitter. *PLoS ONE,* 9(8), e103408. DOI: 10.1371/journal.pone.0103408

R Core Team. (2021), R: A language and environment for statistical computing (4.1.1) [Computer software]. https://www.R-project.org/

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016), A model of text for experimentation in the social sciences. *Journal of the American Statistical Association,* 111(515), 988–1003. DOI: 10.1080/01621459.2016.1141684

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019), stm: An R Package for Structural Topic Models. *Journal of Statistical Software,* 91, 1–40. DOI: 10.18637/jss.v091.i02

Roberts, M. E., Tingley, D., Stewart, B. M., & Airoldi, E. M. (2013), The Structural Topic Model and Applied Social Science. NIPS 2013 Workshop on Topic Models: Computation, *Application, and Evaluation.* DOI: 10.1080/01621459.2016.1141684

Smith, M. K., Jancsik, A., & Puczkó, L. (2020), Customer satisfaction in post-socialist Spas: A case study of Budapest, City of Spas. *International Journal of Spa and Wellness,* 3(2–3), 165–186. DOI: 10.1080/24721735.2020.1866330

Sutherland, I., Sim, Y., Lee, S. K., Byun, J., & Kiatkawsin, K. (2020), Topic Modeling of Online Accommodation Reviews via Latent Dirichlet Allocation. *Sustainability,* 12(5), 1821. DOI: 10.3390/su12051821

Weinshall, D., Levi, G., & Hanukaev, D. (2013), LDA Topic Model with Soft Assignment of Descriptors to Words. P*roceedings of the 30th International Conference on Machine Learning,* 711–719. https://proceedings.mlr.press/v28/weinshall13.html

Hinek Mátyás, PhD, főiskolai tanár
mhinek@metropolitan.hu

Budapesti Metropolitan Egyetem

TOPIC MODELLING OF SPA VISITOR REVIEWS USING THE EXAMPLE OF GELLÉRT SPA AND SWIMMING POOL

THE AIMS OF THE PAPER

The study presents the results of a computer based topical modelling of guest reviews written by visitors of the Gellért Spa and Swimming Pool between 2004 and 2021. From a tourism marketing point of view, the analysis of guest reviews is of particular importance, especially for attractions with a high turnover of visitors. The Gellért is an iconic monumental bath of Budapest and Hungary, a health tourism attraction that is an unmissable experience for many visitors to Budapest. This is reflected in the more than 10,000 guest reviews written in almost 30 languages on Tripadvisor over the last decade and a half, a number so huge that it can only be comprehensively understood by machine.

METHODOLOGY

All reviews written on Tripadvisor between 2005 and 2021 were downloaded using a dedicated app. Reviews written in languages other than English were translated into English using Google Translate. The resulting corpus was analysed using structured topic modelling with latent Dirichlet allocation, a rapidly developing method in text mining, to identify the topics that typically occur in multiple guest reviews. We did this using the statistical software R and the structured topic modelling application STM running in R environment.

MOST IMPORTANT RESULTS

The modelling identified 12 typical themes across the guest reviews. These were also compared with an earlier analysis of the same corpus based on word frequency analysis, which showed that similar themes could be identified using both methodologies. We also separately analysed the representation of opinions on service features that were largely negatively evaluated by guests over the time horizon studied. According to this, the proportion of themes related to hygiene and cleanliness increased, while the proportion of themes related to guest communication, also largely negatively rated, decreased in the written guest reviews of the Gellért Spa.

Keywords: natural language processing, topic modelling, latent Dirichlet distribution, Gellért Spa and Swimming Pool