Bianka PARMOVÁ, Mária VOJTKOVÁ
Katedra štatistiky Fakulty hospodárskej informatiky Ekonomickej univerzity v Bratislave

SEGMENTÁCIA ČITATEĽOV S VYUŽITÍM TEXT MININGU

SEGMENTATION OF READERS USING TEXT MINING

ABSTRAKT
Analýza správania čitateľov na internetovej stránke poskytuje tvorcom webu i tvorcom webového obsahu cenné informácie, ktorých využitie môže zvýšiť zisk prevádzkovateľovi stránky. Odhalenie najpopulárnejších tém preferovaných čitateľmi umožňuje získať lepší prehľad o tom, o ktorý typ obsahu je najmenší záujem a, naopak, ktorá téma je pre čitateľov najzaujímavejšia. Cieľom tohto príspevku je segmentácia online čitateľov na základe prečítaných článkov podľa tém s využitím hĺbkovej analýzy textu. Metódy hĺbkovej analýzy textu využívame na extrahovanie tém článkov spravodajského webu. Prostredníctvom výsledkov tejto analýzy v spojení s údajmi o návštevnosti stránky vytvoríme profily čitateľov na základe tém, ktoré preferujú.

ABSTRACT
A behavioral analysis of readers at the website, provides valuable information for the creators of the site and content, the utilization of which may help to increase revenue of the website operator. Revealing the most popular topics preferred by readers enables the web creators to gain deeper insights into which topics are the least interesting for users and vice versa, which topics attract readers the most. The aim of this paper is to create segments of online readers using text mining techniques, based on the articles read. We are using text mining techniques to extract topics from the articles of the news website. We can create user profiles based on the topic preferences of readers through the analysis results in connection with web traffic data.

KĽÚČOVÉ SLOVÁ
hĺbková analýza textu, extrahovanie tém, model latentnej Dirichletovej alokácie, segmentácia čitateľov

KEY WORDS
text mining, topic modeling, model of Latent Dirichlet allocation, segmentation of readers

1. ÚVOD
V súčasnom trende rozvoja informačných technológií a internetu sa podnikanie v mnohých oblastiach presúva do online prostredia. Výnimkou nie sú ani spravodajské médiá, ktoré ponúkajú svoj obsah na webových stránkach. Výhodou online prostredia je možnosť získania obrovského množstva údajov o zákazníkoch, ktorými sú v prípade online médií samotní návštevníci (čitatelia) stránky. Analýzou dát o návštevnosti stránky v spojení s analýzou textových údajov priamo zo stránky médií môžeme dospieť k cenným informáciám o návštevníkoch stránky, ktoré na základe tradičných nástrojov webovej analytiky nemožno získať.

Hlavným cieľom tohto článku je vytvorenie segmentov čitateľov spravodajského webu s rovnakými preferenciami tém článkov. Tento cieľ pozostáva z dvoch základných parciálnych cieľov. Prvým z nich je charakteristika jednotlivých článkov prostredníctvom tém využitím metód hĺbkovej analýzy textu. Po naplnení prvého základného parciálneho cieľa získame charakteristiku analyzovaných článkov prostredníctvom tém, ktorých sa týkajú. Tieto údaje budú vstupom do druhej časti analýzy a podmienkou naplnenia druhého parciálneho cieľa, ktorým je vytvorenie segmentov čitateľov s rovnakými preferenciami pomocou zhlukovej analýzy.

Výsledok hlavného cieľa práce spočíva v poskytnutí nového, netradičného pohľadu na štruktúru čitateľov spravodajského webu vzhľadom na preferované témy. Kombináciou techniky hĺbkovej analýzy textu a zhlukovej analýzy poskytuje tento článok nový pohľad na segmentáciu v oblasti webovej analytiky, ktorú možno využiť v mnohých ďalších oblastiach okrem aplikovanej oblasti spravodajských médií.

Samotná analýza je založená na dvoch zdrojoch dát. Prvým zdrojom je html kód analyzovanej stránky www.sme.sk, z ktorého sme pomocou kódu v jazyku *R* extrahovali URL adresy článkov a ich kľúčové slová, ktoré vytvárajú autori článkov pri publikovaní. Druhým zdrojom dát sú logy servera analyzovanej stránky, kde sa zaznamenávajú všetky interakcie návštevníkov so stránkou. Tieto dáta sme mali k dispozícii od firmy Piano Media, s. r. o., ktorá prevádzkuje na stránke www.sme.sk služby spoplatnenia online obsahu[1]. Prostredníctvom dát o návštevníkoch tohto spravodajského webu budeme analyzovať ich správanie z pohľadu obsahu prečítaných článkov za analyzovaný mesiac, ktorým bol júl 2015.

2. HĹBKOVÁ ANALÝZA TEXTU A JEJ VÝZNAM V SÚČASNOSTI

Definícia hĺbkovej analýzy textu nie je jednoznačná, pretože sa opiera o poznatky z mnohých ďalších oblastí, ako napríklad vyhľadávanie informácií, hĺbková analýza údajov alebo objavovanie znalostí v dokumentoch. V tejto časti článku by sme chceli objasniť vývoj tohto pojmu.

S pojmom procesu hĺbkovej analýzy textu (Text Mining), často nazývaného aj procesom objavovania znalostí v textových dokumentoch (Knowledge Discovery in Texts – KDT), sa v literatúre stretávame prvýkrát v roku 1999 [4], i keď s výskumom tejto problematiky sa možno stretnúť aj skôr. Autorka tu definuje hĺbkovú analýzu textu pomocou už v tom čase známych disciplín, ako napr. vyhľadávanie informácií (Information Retrieval). Kľúčovou vlastnosťou je hľadanie rôznych typov vzorov v textových súboroch, čo považovala za analogické klasickému procesu hĺbkovej analýzy údajov [7].

Publikácia [6] z roku 2003 rozdeľuje hĺbkovú analýzu textu do dvoch kategórií – klasické a inteligentné dolovanie. „Pod klasickým dolovaním v textoch sa chápe napr. kategorizácia textov, zhlukovanie textov, extrakcia lexikálnych a syntaktických znakov, hľadanie asociácií medzi termami, extrakcia liniek medzi entitami v rámci textu a medzi rôznymi textami, ale napr. aj sumarizácia textov. Inteligentným dolovaním v textoch sa potom nazýva interakcia výskumníka a počítačového nástroja, ako aj použitie metód umelej inteligencie s cieľom vytvárať znalosti o okolitom svete na základe odvodených lingvistických znakov a ďalších typov

[1] *Menovaná firma ako vlastník dát súhlasí s publikovaním výsledkov v predloženej podobe.*

vzorov. Napr. odráža novoidentifikovaná téma v prúde dokumentov realitu? Aké stratégie skúmania implikujú nájdené prepojenia, sú naozaj relevantné? Aké obchodné rozhodnutia možno na ich základe učiniť?“ [7, s. 10].

Publikácia [3] z roku 2007, ktorá sa venuje hĺbkovej analýze textu veľmi podrobne, ju definuje ako „znalostne intenzívny proces, v ktorom používateľ priebežne interaguje s kolekciou dokumentov za pomoci analytických nástrojov“. Autori kladú dôraz na analýzu prepojení medzi informáciami v kolekcii dokumentov, pričom kategorizáciu, zhlukovanie a extrakciu informácií chápu ako súčasť predspracovania textových údajov [7, s. 11].

Náš názor na definíciu hĺbkovej analýzy textu sa vo veľkej miere zhoduje s názorom kolektívu autorov z Technickej univerzity v Košiciach, ktorý vydal publikáciu z oblasti hĺbkovej analýzy textu v slovenskom jazyku Dolovanie znalostí z textov [7]. Autori sa prikláňajú k definícii hĺbkovej analýzy textu podľa základnej definície hĺbkovej analýzy údajov, ktorú prezentujú ako „interaktívny a iteratívny proces získavania platných, pre danú aplikáciu užitočných a doposiaľ neznámych znalostí“. Za dôležitú súčasť tohto procesu považujú taktiež interakciu používateľa s analytickým systémom. S týmto názorom sa stotožňujeme, pretože považujeme zohľadňovanie subjektívnych názorov a vedomostí analytika z analyzovanej oblasti za veľmi dôležitú súčasť analýzy textových údajov.

Po definícii hĺbkovej analýzy textu sa zameriame bližšie na jej základné techniky a ich aplikácie v praxi. Medzi základné techniky hĺbkovej analýzy textu môžeme zaradiť výskumnú analýzu (Exploratory Analysis) a kategorizáciu (Categorization) [5, s. 2].

Výskumná analýza zahŕňa techniky, ako extrahovanie tém (Topic Extraction) či zhlukovú analýzu (Clustering). Hlavnou myšlienkou extrahovania tém je vytvorenie skupín dokumentov, ktoré sa týkajú rovnakých tém. Cieľom zhlukovej analýzy v oblasti hĺbkovej analýzy textu je priradiť jednotlivé dokumenty zo súboru k jednému zhluku, pričom dokumenty v zhlukoch sú si navzájom podobné a jednotlivé zhluky dokumentov sú navzájom odlišné. Rozdiel medzi zhlukovaním a extrahovaním tém zo súboru dokumentov spočíva v tom, že v prípade extrahovania tém sa môže jeden dokument zaradiť súčasne do viacerých skupín v závislosti od toho, koľko tém je v ňom obsiahnutých [5, s. 111].

Kategorizácia súvisí s oblasťou manažmentu obsahu (Content Management) zameranou na organizáciu veľkého množstva dokumentov získaných z rôznych zdrojov na základe ich obsahu. Hlavným rozdielom medzi kategorizáciou obsahu a extrahovaním tém je, že v prípade kategorizácie ide o učenie s učiteľom, teda je potrebné určiť kategórie obsahu, kam majú byť jednotlivé dokumenty zaradené. V prípade extrahovania tém sa tieto kategórie vytvárajú automaticky na základe štatistických metód (napr. latentná sémantická analýza) [5, s. 159].

Technika kategorizácie textu sa využíva v mnohých oblastiach podnikania. Niektoré online médiá využívajú túto techniku na automatické zaraďovanie nových článkov do sekcií na stránke (napr. šport, politika, veda, ekonomika atď.) alebo tiež na vytváranie personalizovaných odporúčaní na obsah súvisiaci s témami v okruhu záujmov čitateľa.

3. EXTRAHOVANIE TÉM SPRAVODAJSKÝCH ČLÁNKOV

Zdrojom dát našej analýzy bola spomínaná internetová stránka spravodajského portálu www.sme.sk, z ktorej sme prostredníctvom html kódu získali kľúčové slová článkov dostupných na stránke v júli 2015. Vstupné dáta obsahujúce tieto údaje pozostávali z 8 203 unikátnych URL adries jednotlivých článkov. Každý z článkov mal vo svojej URL adrese zahrnutý unikátny identifikátor v tvare /c/“sedem číslic“/, pričom niektoré identifikátory článkov sa v dátovom súbore vyskytovali viac než jedenkrát, z čoho vyplýva, že niektoré články boli zahrnuté viackrát s rôznym tvarom URL adresy, čo je pre našu analýzu nežiaduce. Preto sme najprv odstránili duplicitné pozorovania (články).

Po tejto modifikácii tvorili naše dáta 4 043 unikátnych článkov, ktorých obsah reprezentovali kľúčové slová priradené ku každému článku. V ďalšej časti sme sa zamerali na analýzu kľúčových slov. Vstupom do analýzy textu boli teda samotné kľúčové slová, pričom naša databáza obsahovala 7 408 unikátnych kľúčových slov.

Pred samotnou analýzou bolo však potrebné textové údaje predspracovať. Vo všeobecnosti sa v procese predspracovania textových údajov dodržiavajú nasledujúce kroky:

1. Konverzia na čistý text – znamená úpravu dát vzhľadom na to, že kľúčové slová z formálneho hľadiska nie sú jednotne upravené, niektoré slová sa začínajú veľkými písmenami, iné malými napriek tomu, že ich význam je totožný; jednotlivé slová sú oddelené čiarkami, ale s rôznym počtom medzier za nimi atď. Konverzia na čistý text bola aj v našom prípade úvodným krokom analýzy kľúčových slov.

2. Tokenizácia a segmentácia – rozdelenie textu na elementárne textové jednotky. V našom prípade išlo o rozdelenie zoznamu kľúčových slov pri jednotlivých článkoch pomocou jednotného separátora – medzery.

3. Lematizácia, morfologická analýza – jednotlivé slová sa v textových dokumentoch môžu vyskytovať v rôznych morfologických tvaroch, teda v rôznych pádoch, osobách, číslach atď. Preto je nevyhnutné transformovať ich do základného tvaru, tzv. lemu. V našom prípade sme na vstupné údaje aplikovali všetky úpravy okrem lematizácie, pretože sme pracovali už priamo s kľúčovými slovami článkov, ktoré sa uvádzali v prevažne zjednotenom morfologickom tvare.

4. Eliminácia neplnovýznamových slov – znamená vylúčenie takých slov, pri ktorých sa predpokladá malý prínos ku charakteristike obsahu dokumentov. V našom prípade sme spomedzi kľúčových slov vylučovali číslice, znaky a chybné výrazy pozostávajúce z jedného písmena.

5. Váhovanie termov – označuje sa ako úprava frekvencie slov každého dokumentu v celom korpuse. V našom prípade hral tento krok predspracovania údajov významnú úlohu – až 63 % všetkých kľúčových slov (4 637 z celkového počtu 7 408) bolo priradených k článku iba jedenkrát, 15 % len v dvoch prípadoch. Hranicu na elimináciu nízko frekventovaných slov sme stanovili na úroveň 5 priradení v rámci všetkých článkov.

Všetky doteraz opísané kroky predspracovania textových dát boli nevyhnutné na transformáciu textových dokumentov do formy vektorovej reprezentácie.

Po úpravách sme mali k dispozícii vstupné dáta pozostávajúce zo 4 043 článkov opísaných 922 kľúčovými slovami. Aplikácia techniky hĺbkovej analýzy textu (modelu latentnej Dirichletovej alokácie – LDA)) nám umožnila nahradiť kľúčové slová charakterizujúce obsah článkov témami generovanými spomínaným modelom.

Latentná Dirichletova alokácia je súčasťou oblasti pravdepodobnostného modelovania (*probabilistic modeling*). Generatívne pravdepodobnostné modelovanie je postavené na myšlienke, že analyzované dáta považujeme za výsledok generatívneho pravdepodobnostného procesu, ktorý obsahuje skryté (latentné) premenné (*hidden variables*). Tento generatívny proces je definovaný združeným rozdelením pravdepodobnosti pozorovaných a skrytých premenných. Prostredníctvom tohto združeného rozdelenia pravdepodobnosti vieme vypočítať podmienenú pravdepodobnosť skrytých premenných za predpokladu pozorovaných premenných. V prípade LDA sú pozorovanými premennými slová dokumentov a skrytými premennými rozumieme neznámu štruktúru tém v kolekcii dokumentov, ktorú chceme odhaliť. Výpočtovým problémom odvodenia skrytej štruktúry tém je problém výpočtu aposteriórneho rozdelenia pravdepodobnosti, teda podmienenej pravdepodobnosti skrytých tém za predpokladu pozorovaných premenných [1, s. 77 – 84]. Práve z dôvodu využitia štatistickej inferencie založenej na výpočte podmienenej pravdepodobnosti neznámych premenných za predpokladu známych pozorovaní (slov) sa tento model označuje za bayesiánsky [2, s. 71 – 93].

Tabuľka č. 1: Výsledok procesu extrahovania tém

Téma	Poradie	1	2	3	4	5	Názov témy
1	Kľúčové slovo	eko	zvuk	vzi	video	foto	Ekonomika
	Významnosť	0,1099	0,0674	0,0475	0,0275	0,0258	
2	Kľúčové slovo	maďarsko	migranti	čr	migrácia	počasie	Migranti
	Významnosť	0,0752	0,0624	0,0554	0,0267	0,0248	
3	Kľúčové slovo	irán	dohoda	Usa	vzi	horúčavy	Zahraničné správy
	Významnosť	0,0590	0,0501	0,0457	0,0354	0,0354	
4	Kľúčové slovo	polícia	vzi	bax	nehoda	bbx	Policajné správy
	Významnosť	0,1024	0,0300	0,0300	0,0300	0,0248	
5	Kľúčové slovo	hokej	nhl	khl	usa	slovan	Hokej
	Významnosť	0,2299	0,0729	0,0540	0,0440	0,0364	
6	Kľúčové slovo	is	sýria	turecko	usa	útok	Blízky východ
	Významnosť	0,0665	0,0525	0,0402	0,0359	0,0350	
7	Kľúčové slovo	futbal	el	prestup	anglicko	usa	Futbal
	Významnosť	0,2553	0,0245	0,0238	0,0224	0,0175	
8	Kľúčové slovo	zákon	nrsr	novela	návrh	prezident	Politika SR
	Významnosť	0,0472	0,0460	0,0331	0,0224	0,0213	
9	Kľúčové slovo	re	ba	bax	doprava	tt	Dopravné správy
	Významnosť	0,1485	0,0748	0,0408	0,0283	0,0261	
10	Kľúčové slovo	cyklistika	tourdefrance	svet	motorizmus	tdf	Cyklistika
	Významnosť	0,1252	0,0429	0,0340	0,0322	0,0322	

Téma	Poradie	1	2	3	4	5	Názov témy
11	Kľúčové slovo	školstvo	vláda	školy	súdy	bbx	Školstvo
	Významnosť	0,0443	0,0246	0,0172	0,0148	0,0123	
12	Kľúčové slovo	tenis	wimbledon	dvojhra	výsledok	výsledky	Tenis
	Významnosť	0,1271	0,0535	0,0524	0,0468	0,0401	
13	Kľúčové slovo	rusko	ukrajina	usa	francúzsko	eko	Rusko, Ukrajina
	Významnosť	0,0867	0,0766	0,0615	0,0242	0,0232	
14	Kľúčové slovo	usa	obete	útok	británia	čína	Krimi zahraničie
	Významnosť	0,0457	0,0431	0,0326	0,0300	0,0222	
15	Kľúčové slovo	eko	grécko	eú	nemecko	eurozóna	EÚ
	Významnosť	0,1351	0,1210	0,0507	0,0244	0,0192	

Poznámka: *eko* – ekonomika, *vzi* – prevzaté články, *čr* – Česká republika, *bax* – okres Bratislava, *bbx* – okres Banská Bystrica, *nhl* – Americká hokejová liga, *khl* – Kontinentálna hokejová liga, *is* – Islamský štát, *el* – Európska futbalová liga, *nrsr* – Národná rada Slovenskej republiky, *re* – regionálne spravodajstvo, *ba* – Bratislava, *tt* – Trnava, *tdf* – Tour de France, *eú* – Európska únia.
Zdroj: vlastné spracovanie v jazyku R

Výstupom modelu LDA je okrem iného matica početnosti priradení jednotlivých slov k vytvoreným témam, na základe ktorej dokážeme získať obsah matice *phi* s hodnotami pravdepodobnosti, s akou jednotlivé slová súvisia s vytvorenými témami. Použitím pomocnej funkcie v jazyku R získame pre každú tému päť najvýznamnejších kľúčových slov a ich pravdepodobnosť, s akou súvisia s danou témou. Súčet pravdepodobností všetkých slov sa pre každú vytvorenú tému rovná jednej. Tento výstup upravený do tabuľky 1 nám umožňuje interpretovať a pomenovať jednotlivé témy. V prvom stĺpci tabuľky 1 je uvedené číslo témy, každá z tém je charakterizovaná kľúčovým slovom (v riadku *Kľúčové slovo*) spolu s pravdepodobnosťou súvislosti tohto slova s danou témou (riadok *Významnosť*). Na základe najvýznamnejších slov subjektívne určíme názvy tém, ktoré budeme používať v ďalších častiach článku. V poznámke pod tabuľkou uvádzame podrobnejší opis kľúčových slov, ktoré majú formu skratiek.

V prvej téme je najvýznamnejšie slovo *eko* indikujúce správy z ekonomiky. Ďalšie kľúčové slová sa týkajú prevzatých článkov (*vzi*) či videí a fotiek. Ide o viac-menej všeobecnú tému, nazveme ju *ekonomika*.

Druhá téma sa týka vyslovene problematiky migrácie, medzi najvýznamnejšie kľúčové slová patria slová ako *maďarsko, migranti* či *migrácia*. Objavuje sa tu aj *počasie*, no tému napriek tomu nazveme *migranti*.

Tretiu tému tvoria zahraničné správy a najvýznamnejšie kľúčové slová sú *irán, dohoda* a *usa*. Aj v tomto prípade sa medzi prvými piatimi najvýznamnejšími kľúčovými slovami vyskytuje termín týkajúci sa počasia (*horúčavy*), čo je pravdepodobne spôsobené obdobím zberu dát, ktorým bol mesiac júl, keď sa o počasie zaujíma viacero čitateľov.

Medzi pätnástimi témami sa vyskytujú štyri témy týkajúce sa vyslovene športových disciplín: *hokej, futbal, cyklistika* a *tenis*. V prípade týchto tém sa medzi najvýznamnejšími kľúčovými slovami nachádzajú výlučne slová charakteristické pre dané športy. Napríklad téma *tenis* je charakterizovaná kľúčovými slovami, ako *tenis,*

wimbledon, dvojhra, výsledok, výsledky. V prípade témy *futbal* ide o slová *futbal, el* (európska liga), *prestup, anglicko, usa*.

Ďalšími zaujímavými témami s kľúčovými slovami vzťahujúcimi sa vyslovene na danú tému sú napríklad *politika SR* s kľúčovými slovami, ako *zákon, nrsr, novela, návrh, prezident*, alebo téma *EÚ* charakterizovaná najvýznamnejšími termínmi, ako *eko*, *grécko, eú, nemecko, eurozóna*.

Doposiaľ sme analyzovali vzniknuté témy len prostredníctvom ich najvýznamnejších kľúčových slov. Na získanie lepších interpretačných výsledkov použijeme vizualizáciu výsledkov extrahovania tém, ktorá nám umožní porovnať jednotlivé témy a tiež zistiť, ktorá z nich má najväčšiu prevahu.

Na vizualizáciu potrebujeme zostrojiť dve matice. Prvou je *theta*, matica opisujúca relevanciu jednotlivých článkov k vytvoreným témam (tabuľka č. 2). Obsahom matice sú koeficienty *theta*, ktoré sme získali vydelením počtu priradení slov dokumentov k jednotlivým témam v rámci iteračného procesu (celkovým počtom priradení slov dokumentu ku všetkým témam). Téma s najvyššou hodnotou tejto pravdepodobnosti vyjadrená koeficientom *theta* bude daný článok charakterizovať najväčšou mierou. Počet iterácií hovorí o počte opakovaní procesu priraďovania jednotlivých kľúčových slov k témam a jednotlivých tém k článkom, pričom v našom prípade sme nastavili jeho úroveň na 1000 opakovaní.

Tabuľka č. 2: Ukážka matice theta určujúcej závislosti medzi článkami a témami

Kľúčové slová	sr nr sieť poriadok exekučný novela beblavý	usa futbal turné chelsea výhra barcelona sumár	srbsko kosovo albánsko rama návšteva neohlásená	eko rusko meny cudzie nákup pozastavený	sr elektro odborníci priemysel elektro-technický nedostatok zep stretnutie kiska	usa new york černoch policajt uškrtenie dohoda
Ekonomika	0,0024	0,0016	0,0024	0,0047	0,4698	0,0024
Migranti	0,0024	0,0016	0,4843	0,0047	0,0047	0,0024
Zahraničné správy	0,0024	0,0016	0,0024	0,0047	0,0047	0,4843
Policajné správy	0,0024	0,0016	0,0024	0,0047	0,0047	0,4843
Hokej	0,0024	0,0016	0,0024	0,0047	0,0047	0,0024
Blízky východ	0,0024	0,0016	0,0024	0,0047	0,0047	0,0024
Futbal	0,0024	0,9772	0,0024	0,0047	0,0047	0,0024
Politika SR	0,9663	0,0016	0,0024	0,0047	0,4698	0,0024
Dopravné správy	0,0024	0,0016	0,0024	0,0047	0,0047	0,0024
Cyklistika	0,0024	0,0016	0,0024	0,0047	0,0047	0,0024
Školstvo	0,0024	0,0016	0,0024	0,0047	0,0047	0,0024
Tenis	0,0024	0,0016	0,0024	0,0047	0,0047	0,0024
Rusko, Ukrajina	0,0024	0,0016	0,4843	0,4698	0,0047	0,0024

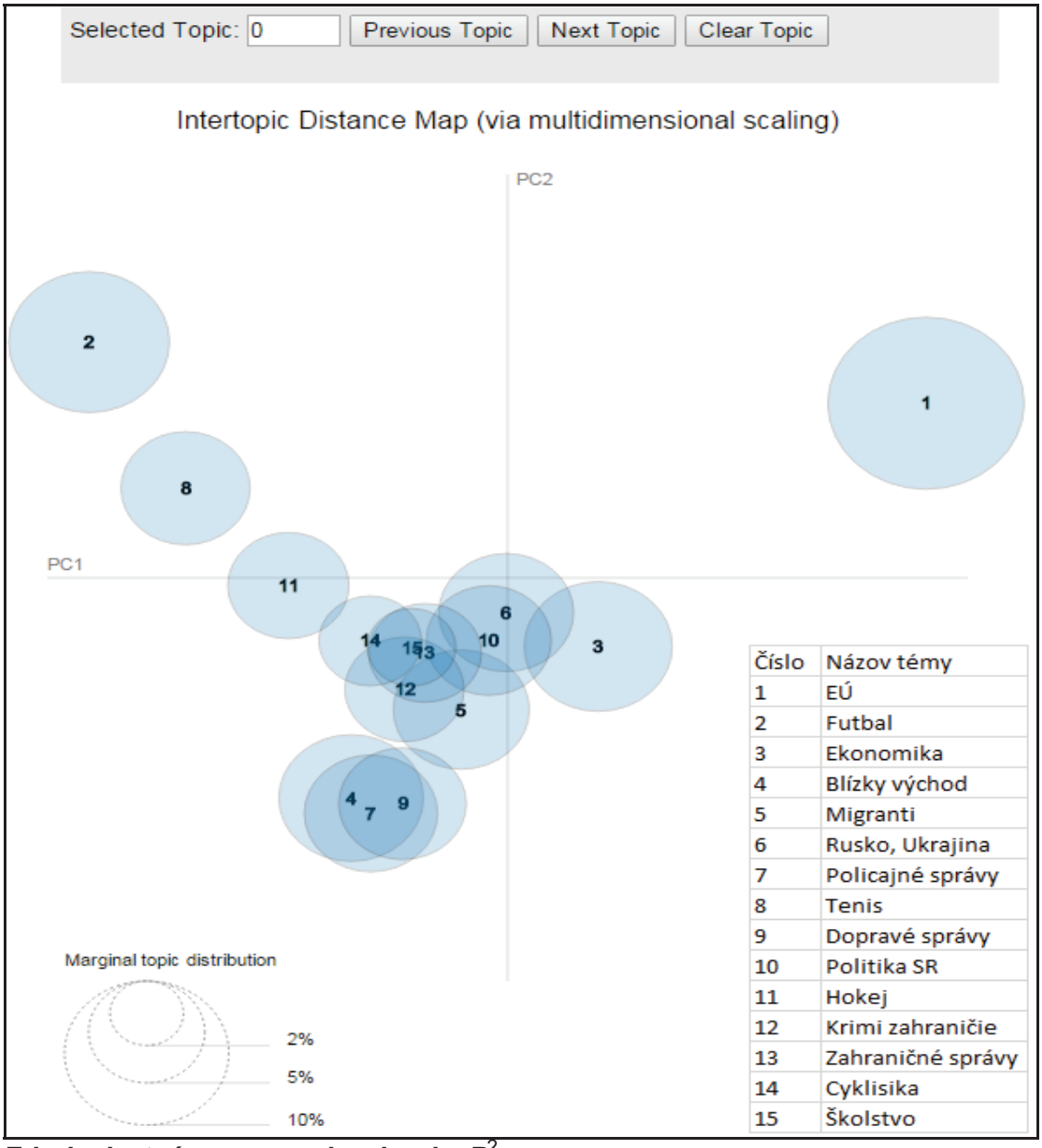
Kľúčové slová	sr nr sieť poriadok exekučný novela beblavý	usa futbal turné chelsea výhra barcelona sumár	srbsko kosovo albánsko rama návšteva neohlásená	eko rusko meny cudzie nákup pozastavený	sr elektro odborníci priemysel elektro-technický nedostatok zep stretnutie kiska	usa new york černoch policajt uškrtenie dohoda
Krimi zahraničie	0,0024	0,0016	0,0024	0,0047	0,0047	0,0024
EÚ	0,0024	0,0016	0,0024	0,4698	0,0047	0,0024

Zdroj: vlastné spracovanie v jazyku R

Prvý článok s kľúčovými slovami *SR, NR, Sieť, poriadok* atď. sa s pravdepodobnosťou 96,63 % týka témy *politika SR*. S ešte vyššou pravdepodobnosťou je v poradí druhý článok priradený k téme *futbal*. Pravdepodobnosť súvislosti tretieho článku s kľúčovými slovami ako *Srbsko, Kosovo, Albánsko* je rozdelená medzi dve témy – *migranti* a *Rusko, Ukrajina*. Príkladom takéhoto rozdelenia pravdepodobnosti je aj nasledujúci článok, ktorý sa týka témy *ekonomika* a *politika SR*. Posledný článok zobrazený v našej tabuľke je príkladom článku z oblasti zahraničných policajných správ s kľúčovými slovami, ako *USA, New York, černoch, policajt*.

Takto upravená matica obsahujúca hodnoty pravdepodobnosti, s akou jednotlivé články súvisia s vytvorenými témami, bude vstupom do ďalšej časti našej analýzy. Charakteristiky článkov prostredníctvom pätnástich premenných využijeme na segmentáciu čitateľov týchto článkov.

Graf č. 1: Interaktívna vizualizácia tém



Zdroj: vlastné spracovanie v jazyku R[2]

Druhou maticou potrebnou na vizualizáciu tém je matica *phi* charakterizujúca vzťah tém a kľúčových slov, na základe ktorej sme v predchádzajúcej časti určili názvy tém. Hodnoty koeficientov *phi* sme určili ako podiel počtu priradení slov k jednotlivým témam a celkového počtu priradení slov ku všetkým témam spolu.

Vizualizácia vytvorených tém je možná prostredníctvom interaktívnej mapy poskytujúcej všeobecný pohľad na model ako celok. Na základe nej vieme analyzovať dominanciu jednotlivých tém v rámci modelu či vzájomnú podobnosť tém.

Jednotlivé témy znázorňujú kružnice s proporcionálnym obsahom k pomeru[3], ktorý

[2] *Číslovanie tém nie je zhodné s označením vo výstupoch modelu, názvy tém uvádzame v legende.*
[3] *LDAvis Vignette, dostupné na internete, dátum posledného prístupu k zdroju: november 2016: <https://cran.r-project.org/web/packages/LDAvis/vignettes/details.pdf>*

vyjadruje odhadovaný počet slov vybraných na základe danej témy v generatívnom procese. Na základe toho platí, že čím väčšia je celková frekvencia výskytu slov charakteristických pre danú tému, tým väčší je obsah kružnice, ktorou sa označuje daná téma.

V našom prípade (graf č. 1) je najviac dominantná téma *EÚ*, nasledovaná témami *futbal* a *ekonomika*. Zaujímavým spôsobom sa od ostatných tém odlišujú športové témy. Najvýznamnejšou spomedzi športových tém je *futbal* ležiaci najviac vychýlene od ostatných.

Najväčší zhluk tvorí skupina tém, v ktorej sa nachádzajú *politika SR*, *migranti* a *ekonomika*. Najmenej významnou v tejto skupine je *školstvo*. Nižšie umiestnená je skupina, ktorú tvoria témy *Blízky východ*, *dopravné správy* a *policajné správy*.

4. SEGMENTÁCIA ČITATEĽOV SPRAVODAJSKÝCH ČLÁNKOV

Druhou časťou analýzy bola segmentácia čitateľov stránky www.sme.sk s využitím výsledkov získaných v predchádzajúcej časti práce. Analyzovali sme 4 043 článkov, ktoré sme opísali prostredníctvom pätnástich premenných – tém, ktoré sme získali na základe analýzy ich kľúčových slov modelom LDA.

Údaje o článkoch a ich pravdepodobnosti súvisu s vytvorenými témami sme spojili s údajmi o správaní čitateľov na stránke, a tak sme vytvorili profil čitateľov podľa charakteru prečítaného obsahu na webe.

Na získanie informácií o správaní čitateľov na stránke sme spracovali náš druhý zdroj dát – databázu logov servera stránky. Táto databáza obsahuje informácie o každom pozretí stránky v priebehu jedného mesiaca ako čas pozretia, URL adresa pozretej stránky, unikátny identifikátor návštevníka, zariadenie a prehliadač, ktorý použil, či URL adresa stránky, z ktorej návštevník prišiel. Pre našu analýzu sme vybrali z tejto databázy potrebné premenné a upravili ich.

Z dôvodu nepresnej identifikácie návštevníkov používajúcich mobilné zariadenia sme sa zamerali len na analýzu správania čitateľov používajúcich na prezeranie stránky osobné počítače. Podiel pozretí stránky z mobilných zariadení sa pohyboval na úrovni 37,5 % z celkových 64 miliónov pozretí v júli 2015. Vzhľadom na cieľ, ktorým bola segmentácia čitateľov na základe preferovaného obsahu prečítaných článkov, sme do analýzy zahrnuli iba čitateľov, ktorí v priebehu analyzovaného mesiaca prečítali aspoň 5 článkov. Do kategórie tzv. lojálnych čitateľov patrilo 244 457 čitateľov, čo predstavovalo približne 12 % všetkých návštevníkov stránky v danom mesiaci.

V nasledujúcom kroku sme spojili dva zdroje údajov – tabuľku s údajmi o čitateľoch a prečítaných článkoch s tabuľkou vytvorenou procedúrou LDA charakterizujúcou články pomocou pätnástich tém, a to na základe unikátneho identifikátora článkov. Výstupom týchto krokov bola teda matica s 244 457 riadkami (počet čitateľov) a 16 stĺpcami, obsahujúca unikátny identifikátor čitateľov spolu s pätnástimi koeficientmi charakterizujúcimi relevanciu prečítaných článkov k vytvoreným témam. Súčet pravdepodobností sa pre každé pozorovanie (čitateľa) rovná jednej. Ukážku matice v transponovanej podobe zobrazuje tabuľka č. 3.

V riadkoch ukážky matice vstupujúcej do zhlukovej analýzy je 15 tém, v stĺpcoch zobrazujeme 6 vybraných čitateľov. Čitateľ 1 sa najviac zaujíma o témy *EÚ* a *cyklistika*, no čitateľ 2 sa zaujíma o tému *EÚ* dvakrát viac. U tohto čitateľa výrazne prevládajú ďalšie dve témy (*politika SR* a *zahraničné správy*), o zvyšné sa zaujíma menej v porovnaní s inými čitateľmi. Preferencie čitateľa 3 súvisia najmä s problematikou zahraničnej politiky, v jeho prípade prevažujú témy *migranti* a *Rusko a Ukrajina*. Články prečítané čitateľom 4 sa týkali všetkých tém s približne rovnakou pravdepodobnosťou. Naopak, v prípade čitateľa 5 vidíme, že v jeho zozname článkov prevládali športové články, ktoré súviseli najmä s témami *cyklistika, hokej* či *futbal.*

Tabuľka č. 3: Ukážka vstupných dát zhlukovej analýzy

Téma	Čitateľ 1	Čitateľ 2	Čitateľ 3	Čitateľ 4	Čitateľ 5
Ekonomika	0,0558	0,0308	0,0878	0,0487	0,0296
Migranti	0,0683	0,0308	0,2504	0,0993	0,0144
Zahraničné správy	0,0548	0,1695	0,0878	0,0466	0,0144
Policajné správy	0,0509	0,0308	0,0084	0,0726	0,1136
Hokej	0,0573	0,0308	0,0084	0,0698	0,2108
Blízky východ	0,0358	0,0308	0,0878	0,0633	0,0307
Futbal	0,0770	0,0308	0,0084	0,0693	0,1659
Politika SR	0,0360	0,1550	0,0084	0,0429	0,0307
Dopravné správy	0,0138	0,0308	0,0084	0,0436	0,0296
Cyklistika	0,1573	0,0308	0,0888	0,1232	0,1975
Školstvo	0,0393	0,0308	0,0084	0,0358	0,0612
Tenis	0,0645	0,0308	0,0084	0,0328	0,0270
Rusko, Ukrajina	0,0633	0,0308	0,1697	0,0876	0,0459
Krimi zahraničie	0,0599	0,0308	0,0878	0,0480	0,0144
EÚ	0,1659	0,3062	0,0809	0,1164	0,0144

Zdroj: vlastné spracovanie v jazyku R

4.1. Metóda *k*-priemerov

S cieľom získať segmenty čitateľov sme sa rozhodli použiť nehierarchickú metódu *k*-priemerov (*k*-means) [9, s. 140]. Jedným z dôvodov bola vyššia efektívnosť spracovania pomerne veľkého počtu 244 457 pozorovaní. Ako mieru podobnosti jednotlivých čitateľov sme použili euklidovskú vzdialenosť, pričom sme overili nezávislosť vstupných premenných ako prvú podmienku na použitie uvedenej miery vzdialenosti. Keďže články boli reprezentované ako reálne vektory v normovanom tvare, nebolo potrebné uvažovať o Mahalanobisovej vzdialenosti [bližšie 8, s. 51 – 80].

Ďalším predpokladom použitia metódy *k*-priemerov je vopred určený počet zhlukov, ktoré chceme dostať. Heuristicky by sme tento počet určili na úrovni štyroch až ôsmich zhlukov, no naše predpoklady sme najskôr overili pomocou grafickej analýzy, ktorá hodnotila výsledky procedúry *k*-priemerov pre počet zhlukov od 2 do 15. Krivka znázorňujúca vnútrozhlukovú variabilitu výraznejšie poklesla pri počte 5 zhlukov, ktorý sme v konečnom dôsledku zvolili za optimálny.

Výsledkom procesu *k*-priemerov je päť homogénnych zhlukov podobnej veľkosti (tabuľka č. 4). Výnimkou je prvý segment, do ktorého patrí 40 % zo všetkých 244 457 čitateľov. Ide o segment čitateľov, ktorí čítajú články zo všetkých tém približne rovnakým podielom.

Tabuľka č. 4: Početnosť zhlukov čitateľov

Segment	Absolútna početnosť	Relatívna početnosť (%)
1	98 276	40,4
2	32 917	13,5
3	36 458	15,0
4	41 835	17,2
5	33 792	13,9

Zdroj: vlastné spracovanie v jazyku R

Tabuľka č. 5: Charakteristika piatich segmentov čitateľov

Téma	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5
Ekonomika	0,0604	0,0679	0,0335	0,0505	0,0354
Migranti	0,1124	0,0645	0,0367	0,0570	0,0504
Zahraničné správy	0,0547	0,0463	0,0251	0,0361	0,0321
Policajné správy	0,0612	0,2481	0,0378	0,0516	0,0523
Hokej	0,0528	0,0418	0,1638	0,0408	0,0660
Blízky východ	0,0789	0,0390	0,0246	0,0392	0,0288
Futbal	0,0456	0,0357	0,3030	0,0347	0,0679
Politika SR	0,0568	0,0516	0,0273	0,0394	0,0290
Dopravné správy	0,0510	0,0879	0,0296	0,0341	0,0376
Cyklistika	0,0802	0,0748	0,1050	0,0710	0,3808
Školstvo	0,0310	0,0332	0,0157	0,0213	0,0179
Tenis	0,0329	0,0260	0,0742	0,0236	0,0473
Rusko, Ukrajina	0,1086	0,0632	0,0364	0,0616	0,0452
Krimi zahraničie	0,0575	0,0422	0,0253	0,0351	0,0273
EÚ	0,1159	0,0780	0,0621	0,4037	0,0821

Zdroj: vlastné spracovanie v jazyku R

Tabuľka č. 5 charakterizuje jednotlivé zhluky prostredníctvom všetkých 15 premenných, teda tém, kde sa pre každú tému uvádza hodnota koeficienta *theta* vypovedajúca o pravdepodobnosti súvislosti medzi danou témou a článkami prečítanými čitateľmi v jednotlivých segmentoch. Súčet hodnôt koeficientov *theta* pre každý segment sa rovná jednej. Čím je hodnota koeficienta vyššia, tým obľúbenejšia je téma v danom segmente. Najpopulárnejšie témy sú zobrazené zelenou farbou a, naopak, najmenej populárne témy v rámci segmentov znázorňujeme červenou farbou.

Najobľúbenejšou zo všetkých segmentov je u čitateľov téma *EÚ*. Až v štyroch segmentoch sa objavila medzi tromi najpopulárnejšími témami. Už sme spomenuli, že téma *EÚ* dominuje všetkým témam, pretože je charakterizovaná kľúčovými slovami s častým výskytom. Kľúčové slová ako *eko, Grécko, EÚ, Nemecko* či *eurozóna* sa v korpuse dát z júla 2015 objavovali často zrejme najmä preto, že sa v tomto období viedli rokovania o poskytnutí pomoci Európskej únie Grécku.

Z podobného dôvodu sa medzi najpopulárnejšími témami segmentov vyskytuje téma *cyklistika*. O tejto téme sa v období zberu dát veľa písalo, pretože v júli 2015 sa konali preteky Tour de France, ktoré boli na Slovensku mimoriadne sledované.

Medzi najmenej populárne témy segmentov patrí *školstvo* definované kľúčovými slovami ako *školstvo, vláda, školy* či *súdy.* Tieto slová sa v sledovanom období vyskytovali v korpuse s malou frekvenciou. O školstve sa publikovalo v júli 2015 menej článkov v porovnaní s ostatnými témami.

Na základe výsledkov zhrnutých v tabuľke č. 5 môžeme tvrdiť, že obdobie zberu dát ovplyvňuje výsledky extrahovania tém a tým aj charakteristiku segmentov čitateľov. Napríklad segment 5, pre ktorý je charakteristická najmä téma *cyklistika*, by sme mohli považovať za dočasný; v analýze dát z iného obdobia by sa pravdepodobne neobjavil. Tento fakt však neplatí pre zvyšné segmenty, ktoré môžeme považovať za stabilné segmenty čitateľov webu s nasledujúcimi charakteristikami:

Prvý segment je najpočetnejší spomedzi všetkých piatich, tvorí ho 40 % analyzovaných čitateľov, ktorí čítajú správy zo všetkých tém s približne vyrovnaným podielom. Najviac prevažujú témy *EÚ, migranti* a problematika *Ruska a Ukrajiny.* Priemerná hodnota koeficienta *theta* pre tieto témy sa pohybuje na úrovni 0,1, čo je v porovnaní s najdominantnejšími témami iných segmentov nízka hodnota.

Témami s najväčšou prevahou sú v prípade *druhého segmentu* témy z oblasti policajných a dopravných správ, kde medzi najdominantnejšie kľúčové slová patrili slová, ako *polícia, nehoda, vzi* (indikuje prevzaté články)*, tt, ba*, resp. *bax* (indikujúce správy z hlavného mesta). Medzi najmenej populárne témy tohto segmentu patria okrem témy školstva a športových tém najmä témy týkajúce sa zahraničia (*krimi zahraničie, Blízky východ, zahraničné správy*). Pomocou týchto indikátorov by sme mohli túto skupinu čitateľov označiť ako čitateľov aktualít z lokálnej oblasti.

Tretí segment tvoria čitatelia športových správ. Výrazne dominantnou témou je *futbal*, kde priemerná hodnota pravdepodobnosti relevancie danej témy priradená k článkom prečítaným čitateľmi tohto segmentu dosiahla hodnotu 0,30. Druhou najpopulárnejšou témou tohto segmentu je *hokej* takisto s relatívne vysokou hodnotou koeficientu *theta* 0,16, nasledovaný v tomto období populárnou *cyklistikou*. Téma *tenis* dosahuje v tomto segmente hodnotu 0,07, čo je najvyššia hodnota koeficienta *theta* spomedzi všetkých segmentov pre túto tému.

Vo štvrtom segmente výrazne dominuje téma *EÚ*, ktorej priemerná hodnota koeficienta *theta* pre tento segment dosahuje hodnotu až 0,4. Druhé poradie v obľúbenosti tém v tejto skupine čitateľov obsadzuje *cyklistika*, treťou najobľúbenejšou témou je *Rusko a Ukrajina*, no hodnoty koeficientov *theta* pre tieto témy sú výrazne nižšie, dosahujú úroveň 0,07, resp. 0,06. Najmenej obľúbenými témami tohto segmentu čitateľov sú *tenis* a *školstvo*, ktoré patria medzi najmenej populárne témy aj v ostatných segmentoch. Keďže vo štvrtom segmente extrémne dominuje téma *EÚ*, ktorú charakterizujú kľúčové slová, ako *eko, Grécko, Nemecko, EÚ, eurozóna, financie, banky,* usudzujeme, že ide o segment čitateľov zaujímajúcich sa najmä o správy z oblasti ekonomiky.

Ako sme už spomenuli, *piaty segment*, ktorý tvorí približne 13,9 % všetkých analyzovaných čitateľov, považujeme za dočasný, keďže jeho vznik ovplyvnilo obdobie zberu dát. V tomto zhluku dominuje v danom období populárna téma *cyklistika*, priemerná hodnota koeficienta *theta* dosahuje pre túto tému úroveň 0,38.

5. ZÁVER

Výsledkami segmentácie spravodajského denníka je päť homogénnych zhlukov čitateľov, ktoré sa navzájom odlišujú preferenciami obsahu článkov. Identifikovali sme segment fanúšikov športových správ, kde dominovali všetky štyri športové témy – *futbal, hokej, tenis* a *cyklistika*. Taktiež sme zistili, že segment s najväčším počtom čitateľov sa zaujíma o široké spektrum tém. Na základe toho môžeme predpokladať, že čitatelia tohto segmentu čítajú väčšie množstvo článkov z rôznych oblastí a patria tak medzi najlojálnejšiu časť publika. Medzi piatimi segmentmi čitateľov sa objavil jeden dočasný, v ktorom výrazne prevažovala téma cyklistiky, ktorá bola populárna najmä v období zberu dát.

V súčasnosti sa stáva trendom v oblasti spravodajských médií spoplatňovanie obsahu (Paid Content), teda prijímanie platieb od čitateľov za sprístupnenie obsahu vybraných (prípadne všetkých) článkov na stránke. S cieľom zvyšovania zisku prostredníctvom spoplatnenia obsahu sa prikladá dôraz na ďalší cieľ – zvyšovanie spokojnosti a lojality zákazníka. Zaplatenie za sprístupnenie obsahu je viac pravdepodobné u lojálneho čitateľa ako u náhodného návštevníka stránky.

Práve otázka týkajúca sa obsahu spravodajských článkov bola pre nás v tomto príspevku najpodstatnejšia. Analyzovať obsah článkov je možné napríklad na základe sekcie, do ktorej sú priradené, na základe kľúčových slov priradených ku každému článku alebo na základe samotného textu článkov. Výsledkom analýzy tohto druhu môžu byť napríklad segmenty čitateľov, ktorí sú zoskupovaní podľa obsahu článkov, ktoré za určité obdobie prečítali. Výsledky segmentácie čitateľov podľa prečítaného obsahu poskytujú pre tvorcov webu cenné informácie o svojom publiku, ktoré na základe tradičných nástrojov webovej analytiky nemožno získať.

Odhalenie najpopulárnejších tém preferovaných čitateľmi umožňuje tvorcom webu získať lepší prehľad o tom, o ktorý typ obsahu je najmenší záujem a, naopak, ktorá téma je pre čitateľov najzaujímavejšia. Ak sa napríklad ukáže, že medzi najpopulárnejšie témy patrí taká, na ktorú sa tvorcovia nezameriavajú takou mierou ako na ostatné, tak je ideálne pozmeniť štruktúru obsahu článkov a zvýšiť počet článkov tvorených na dopytovanú tému a zvýšiť tým lojalitu návštevníkov.

Môže tiež nastať prípad, keď tvorcovia obsahu pomocou takejto analýzy zistia, že články zamerané na témy, z ktorých publikujú väčší počet článkov v porovnaní s ostatnými témami, oslovujú iba malú časť publika. V tomto prípade ponuka prevyšuje dopyt a je pre tvorcov optimálne znížiť počet publikovaných článkov na danú tému.

Informácie z analýz takéhoto druhu sa dajú využiť napríklad aj v oblasti marketingu, kde možno zostavovať emaily odberu noviniek na základe obľúbených tém čitateľa a zvyšovať tak frekvenciu návštev a lojalitu odberateľov noviniek. V prípade spravodajských webov spoplatňujúcich obsah je možné na základe segmentácie čitateľov vytvárať cielené marketingové kampane so zľavami na časť

obsahu podľa záujmov čitateľov alebo vytvárať špeciálne balíčky predplatného zamerané napr. len na čitateľov ekonomiky, regionálnych správ alebo športu.

Uvedený článok poskytuje komplexný pohľad na štruktúru čitateľov spravodajského webu vzhľadom na preferované témy. Spojením výsledkov analýzy textových dát charakterizujúcich obsah článkov a údajov o návštevnosti stránky sme opísali charakter čitateľov z nového, netradičného pohľadu. Doplňujúcou úlohou tohto článku bola ukážka aplikácie hĺbkovej analýzy textu v oblasti webovej analytiky ako možnosť zlepšenia tradičných analýz o návštevnosti stránky.

LITERATÚRA

[1] BLEI, D.: Probabilistic Topic Models. In: Communications of the ACM, 2012, Vol. 55, No. 4, p. 77-84.
[2] BLEI, D. – LAFFERTY, J. D.: Topic models. In: Text Mining: Classification, Clustering and Applications, 2009, p. 71-93.
[3] FELDMAN, R. – SANGER, J.: The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, 2007. ISBN 978-0-521-83657-9.
[4] HEARST, M. A.: Untangling text data mining. In: Proceedings of ACL ’99: the 37th annual meeting of the Association for Computational Linguistic. University of Maryland, 1999, p. 3-10.
[5] CHAKRABORTY, G. – MURALI, P. – SATISH, G.: Text Mining and Analysis: Practical Methods, Examples and Case Studies Using SAS. Cary, NC:SAS Institute Inc., 2013. ISBN 978-1-61290-551-8.
[6] KROEZE, J. H. – MATTHEE, M. C. – BOTHMA, T. J. D.: Differentiating Data- and Text-Mining Terminology. In: Proceeding of SAICSIT, 2003, p. 93-101. ISBN:1-58113-774-5.
[7] PARALIČ, J. a kol.: Dolovanie znalostí z textov. Košice: Equillibria, 2010. ISBN 978-80-89284-62-7.
[8] ŘEZÁNKOVÁ, H. – HÚSEK, D. – SNÁŠEL, V.: Shluková analýza dat. Příbram: Professional Publishing, 2007. ISBN 978-80-86946-26-9.
[9] STANKOVIČOVÁ, I. – VOJTKOVÁ, M.: Viacrozmerné štatistické metódy s aplikáciami. Bratislava: Iura Edition, 2007. ISBN 978-80-8078-152-1.

RESUME

The outcome of the analysis described in this paper, is to identify the segments of online news readers based on topic preferences. The traditional methods of web traffic are completed with text analyses, in order to describe the readers from a new perspective.
Based on the results of cluster analysis, we concluded that the period of data collection affects the results of topic extraction and thus the readers´ segment description. One of the five segments were composed of readers preferring the topic of cycling, which was specific for the analyzed period when the Tour de France was held. The other four segments of readers are considered stable, and expected to appear in a similar form also in other month’s analysis. The most important segment was the most numerous one, consisting of approximately forty percent of the analyzed readers. Readers in this segment, prefer reading articles on all topics with approximately the same proportions. We also identified a segment consisting of sports fans, where particularly the sport topics were dominant.

PROFESIJNÝ ŽIVOTOPIS

Ing. Bianka Parmová *v máji 2016 ukončila druhý stupeň štúdia na Fakulte hospodárskej informatiky Ekonomickej univerzity v Bratislave v študijnom odbore štatistické metódy v ekonómii. V súčasnosti pracuje ako dátová analytička (Data Scientist) vo firme Piano Media, s. r. o. Článok vychádza z výsledkov jej diplomovej práce, ktorú vypracovala pod vedením doc. Ing. Márie Vojtkovej, PhD.*

Doc. Ing. Mária Vojtková, PhD., *pôsobí vo funkcii docentky na Katedre štatistiky Fakulty hospodárskej informatiky Ekonomickej univerzity v Bratislave. Vo svojej vedeckovýskumnej a pedagogickej činnosti sa venuje viacrozmerným štatistickým metódam, ktoré sú zamerané na aplikáciu viackriteriálneho hodnotenia v rôznych oblastiach sociálno-ekonomického života, hľadanie skrytých vzťahov pomocou metód zníženia dimenzie, segmentáciu, čiže zhlukovanie podobných objektov charakterizovaných určitými vlastnosťami, a určenie diskriminačnej funkcie ako spôsobu rozlíšenia medzi vytvorenými skupinami a klasifikáciu nových objektov. Je spoluautorkou vedeckých monografií, niekoľkých učebníc, skrípt a mnohých vedeckých článkov publikovaných doma i v zahraničí.*

KONTAKT

bianka.parmova@gmail.com
maria.vojtkova@euba.sk