APLIKÁCIE TEXT MININGOVÝCH METÓD

Jana Rusinková

Abstrakt

Článok analyzuje možnosti aplikácie text miningových metód pre textové dokumenty. Bližšie informuje čitateľa o možnostiach práce s neštrukturovaných textoch a poukazuje na ďalšie príležitosti v oblasti objavovania znalostí v textoch. Článok sa bližšie venuje aplikácií text miningových metód v praxi a poukazuje na ich možnosti využitia.

Kľúčové slová: text mining, metódy text miningu, analýza sentimentu, extrakcia informácií, sumarizácia textu, identifikácia autora, prisudzovanie autorstva, automatická identifikácia jazyka

Abstract

This article examines application of text-mining methods for text documents. Furthermore this paper informs the reader about text mining technologies for unstructured text and proposes methods that improve knowledge discovery in texts. The article pays more attention to the application of text-mining methods in practice and highlights their potential usage.

Keywords: text mining, text mining methods, sentiment analysis, information extraction, text summarization, author identification, automatic language detection

Úvod

S príchodom počítačov sa čoraz častejšie začali výskumníci pokúšať o to, aby mohli byť počítače schopné rozumieť prirodzenému jazyku. Komunikácia medzi počítačom a človekom by sa tak stala omnoho prirodzenejšia, jednoduchšia a efektívnejšia. Spolu s masívnou elektronizáciou dokumentov vo všetkých sférach ľudských činností dochádza neustále k masívnemu nárastu množstva dát uložených v elektronickej podobe. Približne 80% dát uložených v najrôznejších databázach má podobu textu, je teda v podobe neštruktúrovaných dát. Pre používateľov je preto nevyhnutné, aby vo všetkých neštruktúrovaných textových

informáciách uložených nestratili prehľad, priebežne ich systematizovali a samozrejme tiež vyťažili prospech z včasného získania ďalších informácií analýzou existujúcich dát.

Text mining všeobecne spadá pod súbor data miningových metód, kde vznikol ako ďalšie odvetvie data miningu, ktoré pokrýva požiadavky po spracovaní textov za súbežného vyhľadania zanlostí v nich obsiahnutých. Dôvodom separácie text miningu od data miningu je predovšetkým skutočnosť, že data mining má všeobecnejší záber, vyhľadáva a spracováva informácie aj v číslach, nominálnych a ordinárních premenných, naopak text mining sa špecializuje výhradne na prácu s neštruktúrovaným textom.[1]

Formálnejšie by sa text mining dal definovať ako metóda netriviálnej automatickej extrakcie skrytých, implicitných, vopred neznámych a potenciálne užitočných a dôležitých informácií z veľkého množstva neštruktúrovaných alebo čiastočne štruktúrovaných textových dát pomocou kombinácie nástrojov strojového učenia, pokročilých štatistických analýz, rôznych algoritmov, postojov a trendov. Výstupom sú zmysluplné informácie. Metódy text miningu pracujú s veľkým objemom dát a vyhľadávajú informácie v texte, ktorý je napísaný v prirodzenom jazyku. Z tohto dôvodu tak bojujú so šumom a inými problémami.

Analýza sentimentu

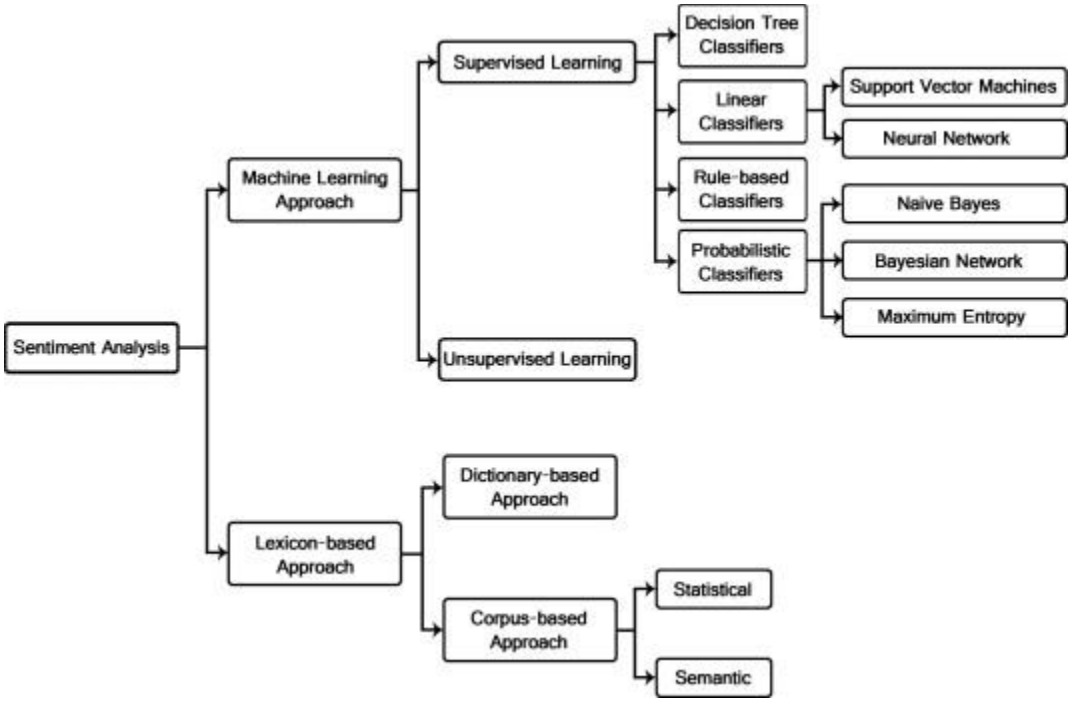
Analýza sentimentu (sentiment analysis) využíva metódy textminingu na členenie dokumentov podľa emočného obsahu do troch skupín:

- pozitívny (positive)
- negatívny (negative)
- neutrálny (neutral).

Softvér pri triedení textov pracuje s citovo zafarbenými slovami, pričom tieto slová porovnáva s ostatnými všeobecne použitými slovami v dokumente. Analýza textu prebieha na základe hodnotenia sentimentu, pričom sa prihliada na frekvenciiu použitia expresívnych slov. Ak je ich frekvencia v texte nadpriemerná, na základe tohto hodnotenia je text definovaný ako pozitívny

[1] ULDRICH, M. Text mining aneb Kladivo na nestrukturovaná data. In: *IT Systems* [online]. 2011, **12** [cit. 2015-03-12]. ISSN 1802-615X. Dostupné na: http://www.systemonline.cz/clanky/text-mining-kladivo-nanestrukturovana-data.htm

alebo negatívny. Ak mu výskyt týchto expresných slov v texte nízku mieru, text je kategorizovaný ako neutrálny. Analýza sentimentu je zložitu textminingovou úlohou, ktorá využíva techniky strojového učenia. Možnosti, ako pristupovať k analýze sentimentu si uvádzame na nasledujúcom Obrázku 1.



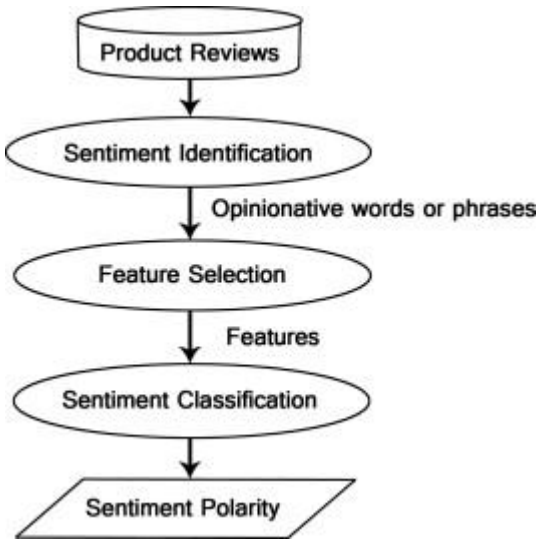
Obr. 1 Techniky analýzy sentimentu[2]

Aplikačné oblasti analýzy sentimentu:

- monitoring sociálnych sietí
- identifikácia relevantných a dôležitých miest diskusií
- identifikácia tzv. opinion makers alebo opinion leaders (diskutujúcich s veľkým vplyvom)
- sledovanie hodnotení produktov, služieb, značiek (dôležité pre samotné firmy, prípadne pre konkurenciu)

[2] MEDHAT, et al. Sentiment analysis algorithms and applications: A survey. In: *Ain Shams Engineering Journal* [online]. 2014, **5**, 1093–1113. [cit. 2015-03-20]. ISSN 2090-4479.
Dostupné na: http://ac.els-cdn.com/S2090447914000550/1-s2.0-S2090447914000550-main.pdf?_tid=b322cee8-d61f-11e4-8789- 00000aacb35e&acdnat=1427639416_e6dbe68cef0a7c15058d2f2a1c649e61

- predikcia vývoja trhu - predajnosť a zmeny cien na základe analýzy on-line hodnotení a recenzií
- ovplyvňovanie politických rozhodnutí - identifikovanie sentimentu v diskusii o konkrétnej oblasti (terorizmus, rómska otázka, ismal v Európe a pod.)



Obr. 2 Diagram analýzy sentimentu v produktových recenziách[3]

Analýza sentimentu v prostredí online reputačného manažmentu

Spoločnosť Visibility[4] v roku 2013 vypracovala štúdiu zaoberajúcu sa analýzou sentimentu výsledkov organického vyhľadávania. Výsledky z roku 2013 potom porovnávala s rovnakým výskumom, ktorý vypracovali v roku 2010. V rámci štúdie autori definovali vybrané segmenty jednotlivých predstaviteľov z rôznych polí podnikateľského pôsobenia. Sentiment prvých 10 výsledkov vo vyhľadávaní na Google.sk bol kontrolovaný pomocou portálu www.hidemyass.sk, ktorý slúži ako proxy server. Takýto postup zabezpečoval objektívnosť zobrazených výsledkov a zabraňoval zobrazovaniu personalizovaných informácii, ktoré sú ovplyvňované predovšetkým

[3] MEDHAT, et al. Sentiment analysis algorithms and applications: A survey. In: *Ain Shams Engineering Journal* [online]. 2014, **5**, 1093–1113. [cit. 2015-03-20]. ISSN 2090-4479.
Dostupné na: http://ac.els-cdn.com/S2090447914000550/1-s2.0-S2090447914000550-main.pdf?_tid=b322cee8-d61f-11e4-8789- 00000aacb35e&acdnat=1427639416_e6dbe68cef0a7c15058d2f2a1c649e61

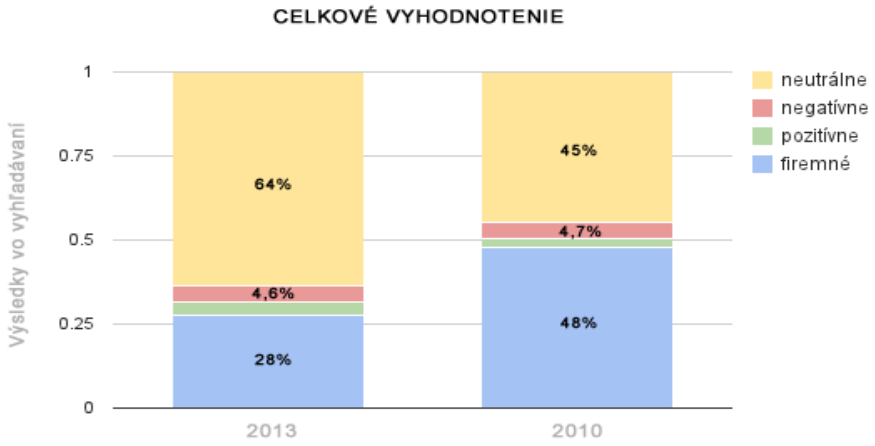
[4] SASKO, J. *Analýza sentimentu výsledkov slovenských firiem* [online]. 2014 [cit. 2015-03-25]. Dostupné na: http://www.reputation.sk/wp-content/uploads/2014/02/ORMreport.pdf

polohou vyhľadávajúceho, jeho históriou či cookies. V analýze sa zohľadňovali len organické výsledky vyhľadávania.

Tab. 1: Hodnota sentimentu prvých 10 výsledkov vo vyhľadávaní[4]

pozícia výsledku	1	2	3	4	5	6	7	8	9	10
pozitívny sentiment	20	19	18	17	16	15	14	13	12	11
web spoločnosti	10	9	8	7	6	5	4	3	2	1
neutrálny sentiment	2	2	2	2	2	2	2	2	2	2
negatívny sentiment	-20	-19	-18	-17	-16	-15	-14	-13	-12	-11

V tabuľke 1 môžeme vidieť hodnoty pridelené sentimentu prvých 10 výsledkov vo vyhľadávaní. Po spočítaní jednotlivých bodov sentimentu všetkých desiatich výsledkov na Google.sk pre danú firmu dostaneme výslednú sumu. Tá predstavuje finálny faktor pre hodnotenie úspechu resp. neúspechu spoločnosti v konkrétnom segmente. Analýza sentimentu výsledkov vo vyhľadávaní pre rok 2013, v porovnaní s rokom 2010, neskôr jasne ukázala, že mnohé slovenské firmy stále zanedbávajú svoju online reputáciu, čo môže mať za následok stratu nielen potenciálnych, ale aj aktuálnych zákazníkov. V súčasnosti stavajú zákazníci svoje nákupné rozhodnutia na tom, čo si o spoločnosti, výrobku či jednotlivcovi zistia z diskusií či recenzií na internete. Aj z tohto dôvodu je veľmi dôležité sledovať správanie zákazníkov, k čomu môže pozitívne prispieť aplikovanie analýzy sentimentu do praxe.



Obr. 3: Analýza sentimentu výsledkov vo vyhľadávaní[5]

[5] SASKO, J. *Analýza sentimentu výsledkov slovenských firiem* [online]. 2014 [cit. 2015-03-25]. Dostupné na: http://www.reputation.sk/wp-content/uploads/2014/02/ORMreport.pdf

Extrakcia informácií

Extrakcia informácií (information extraction) je identifikácia a následná klasifikácia špecifickej informácie nájdenej v neštruktúrovaných zdrojoch (ako napr. v textových zdrojoch v prirodzenom jazyku) do sémantických kategórií tak, aby sa daná informácia stala jednoduchšie spracovateľnou.[6]

Sedláček[7] definuje cieľ extrakcie informácií ako prevod neštruktúrovaného textu do štruktúrovaného formátu (napríklad databáza) podľa jeho obsahu. Jedným z hlavných účelov tejto činnosti je dodanie informácií ďalším metódam v procese spracovania. Zvolený prístup záleží na štruktúre textu. Tá môže nadobúdať tieto formy:

- neštruktúrovaný text (free text) - text s celými gramatickými vetami, ktoré umožňujú spracovanie pomocou metód spracovania prirodzeného jazyka. Vzory sú získavané na základe syntaktickej a sémantickej analýzy;
- štruktúrovaný text - text s vopred definovanou štruktúrou, v ktorom sú vzory oddelené určitým pevne daným spôsobom;
- semi-štruktúrovaný text - text, ktorý nie je gramaticky štruktúrovaný, je často písaný heslovitým spôsobom.

Podľa Neta[8] je extrakcia informácií zjednodušene iba vytvorenie podmnožiny viet pôvodného textu. Extrakcia informácií je metóda, ktorá poskytuje veľmi dobré výsledky. Text je hlboko analyzovaný, kedy na základe sémantickej reprezentácie je parafrázovaný jeho obsah. Ide o čisto strojovú úpravu, čo môže spôsobovať problémy, najmä v prípade slovenského jazyka. Výsledné generovanie súvislého textu nie je na takej úrovni, aby dosahovalo zrozumiteľnosť výstupu ako po použití textovej sumarizácie.

---

[6] MOENS, Marie-Francine. *Information Extraction: Algorithms and Prospects in a Retrieval Context. (The Information Retrieval Series).* London: Springer, 2006. ISBN 978-1-402-04987-3.

[7] SEDLÁČEK, P. *Text mining a jeho možnosti (aplikace)* [online]. 2003 [cit. 2015-02-27]. Dostupné na: http://www.fi.muni.cz/usr/jkucera/pv109/2003p/xsedlac5.htm

[8] NETO, J. L., et al. *Automatic Text Summarization using a Machine Learning Approach* [online]. 2002 [cit.2015-03-25]. Dostupné na: http://www.cs.ukc.ac.uk/people/staff/aaf/pub_papers.dir/SBIA-2002- Joel.pdf

Tento problém sa však týka najmä automatickej tvorby abtraktov. Pri čítaní textu tak môže dochádzať k náročnejšiemu pochopeniu a jednotlivé vety môžu pôsobiť, ako by boli vytrhnuté z kontextu. Avšak extrahovaný text by mal byť po prečítaní pochopiteľný.

Medzi základné podúlohy extrakcie informácií patrí[9]:

- rozpoznávanie pomenovaných entít (nájdenie a identifikácia zaujímavých (pomenovaných) objektov v texte, napr. mená osôb, názvy miest, lokalít),
- extrakcia relácií (nájdenie vzťahov medzi pomenovanými objektmi),
- extrakcia a dolovanie názorov,
- extrakcia kľúčových slov a termínov,
- automatická tvorba abstraktov.

Textová sumarizácia

Hlavnou úlohou automatickej sumarizácie textu je zredukovanie veľkosti dokumetu s cieľom zachovania informačej hodnoty[8]. Textová sumarizácia je ďalším spôsobom automatického zhrňovania textu. Ide o metódu, ktorá poskytuje horšie výsledky ako extrakcia informácií. Princípom textovej sumarizácie je vytvoriť zhrnutie textu automaticky a v krátkom čase. Ak by bolo nutné čítať celý dokument a manuálne vytvoriť jeho zhrnutie, bolo by to časovo i finančne veľmi náročné. Najmä ak ide o obrovský korpus textov. Aj v tomto prípade je možné využiť nástroje textmiringu a text sumarizovať automaticky. Vytváranie súhrnu (sumarizácie) textu definuje Ježek[10] ako:

- vytvorenie stručnej a presnej reprezentácie obsahu dokumentu,
- vybratie najdôležitejších informácií zo zdrojového textu, ktoré ho rekapitulujú pre účely a úlohy používateľa.

[9] MOENS, Marie-Francine. *Information Extraction: Algorithms and Prospects in a Retrieval Context. (The Information Retrieval Series).* London: Springer, 2006. ISBN 978-1-402-04987-3.
[10] JEŽEK, K. a J. STEINBERGER. *Sumarizace textů* [online]. 2010 [cit.2015-03-25]. Dostupné na: http://textmining.zcu.cz/publications/SumarizDATAKON.pdf

Spôsob hodnotenia môže vychádzať z porovnávania výsledného súhrnu s pôvodným textom, s ručne vytvoreným súhrnom alebo so súhrnom vytvoreným iným sumarizačným systémom. Hodnotiace metódy môžu byť rozdelené nasledovne[11]:

- **Priame metódy** - sú založené priamo na analýze súhrnu a jeho porovnanie s originálom čo do miery tematickej obsažnosti, kontextu, čitateľnosti, gramatiky a pod. Porovnávať výsledok je možné aj s ručne vytvoreným abstraktom (od autora originálu alebo od profesionálneho abstraktora),
- **Nepriame metódy** - sú založené na miere použiteľnosti súhrnu pre zadaný účel. Tým môže byť napr. klasifikačné úloha, filtrovanie, vyhľadávanie alebo odpovedanie na otázky. Kvalita súhrnu potom môže byť určená kvantitatívnymi ukazovateľmi ako sú napríklad presnosť a úplnosť výberu podľa súhrnu v porovnaní s výberom podľa originálu alebo "ideálneho", ručne konštruovaného súhrnu.Je potrebné poznamenať, že žiadny súhrn dokumentu nie je možné považovať za ideálny.

Prisudzovanie autorstva

Spory a polemika v oblasti prisudzovania autorstva textu začali vzbudzovať záujem už v antike[12]. Až do konca 19. storočia sa pristupovalo k zisťovaniu autorstva na základe postojov a svedectva samotného autora. Počas posledných rokov sa výskum v oblasti určovania autorstva opiera o poznatky z oblasti strojového učenia, vyhľadávania informácií a spracovania prirodzeného jazyka[13].

Od konca 90. rokov neustále narastá množstvo elektronických textov (blogov, článkov, elektronických kníh, zdigitalizovaných materiálov), ktoré umožňuje moderným technológiám automatizovať proces zisťovania autorstva. Techniky určovania autorstva majú v praxi široké využitie. Je možné overovať pravosť svedectiev, závetí či iných úradných dokumentov. Firmy a zákazníci napríklad využívajú metódy na overenie, či v príspevkoch, vo fórach alebo v recenziách diskutuje neustále tá istá osoba, iba pod iným pseudonymom.

---

[11] JEŽEK, K. a J. STEINBERGER. *Sumarizace textů* [online]. 2010 [cit.2015-03-25]. Dostupné na: http://textmining.zcu.cz/publications/SumarizDATAKON.pdf

[12] MURRAY, G. *The Rise of the Greek Epic*. 4th ed. London: Oxford University Press, 1967.

[13] RYGL, J. *Autorství a stylometrie* [online]. 2015 [cit. 2015-01-07]. Dostupné na: https://nlp.fi.muni.cz/cs/AutorstviAStylometrie

V prípade rozdeľovania a zhlukovaniu textu podľa autorstva sa zisťuje, či bol text napísaný jedným autorom. Ak sa na texte podieľalo viac autorov, texty sa rozdelia do častí podľa autorov. V dnešnej dobe tiež často počujeme o plagiátorstve. Mnoho autorov však svoje diela uverejňuje anonymne alebo pod pseudonymom. V tomto prípade je tak možné využiť omnoho jednoduchšie metódy než pri probléme identifikácie autorstva.

Štýl autora

Osobitný štýl je typický pre každého autora. Najznámejší autori klasických anglických románov a divadelných hier sú známe práve kvôli osobitnému štýlu, ktorý robí ich diela okamžite rozpoznateľné. Od zložitosti Henryho Jamesa, humoru Dickensa, priamočiarosti Jane Austenovej až k neformálnosti Marca Twaina a jednoduchosti Jacka Londona. Čitateľ, ktorý dobre pozná týchto románopiscov, vie rozpoznať spoznať ich štýl písania. Niektorí autori sú čítanejší nie kvôli obsahu ich textov a tomu, čo chcú dielom povedať, ale práve pre výnimočný štýl písania.[14]

Štýl však nie je jednoduché definovať. Preto je úlohou prisudzovania autorstva identifikovať spoľahlivé ukazovatele, ktoré môžeme považovať za prvok autorovho štýlu. Techniky určovania autorstva sa v literatúre využívajú nielen na overenie pravosti autora či plagiátorstva, ale tiež na prisúdenie diela od anonymného autora.

Podľa najnovších výskumov je možné autorstvo spoľahlivo určit, ak sa spoja lingvistické metódy s metódami strojového učenia. V prípade lingvistiky ide o stylometriu, na základe ktorej je možné identifikovať typické znaky a vlastnosti konkrétneho autora. Tie sa prejavujú vo všetkých jeho dielach a zároveň sú unikátne, odlišujú sa od znakov iných autorov. Pri určovaní autorstva pomocou analýzy prirodzeného jazyka je potrebné si definovať jednotlivé metriky, ktoré sa budú porovnávať[15].

[14] ZHAO, Y. a J. ZOBEL. *Authorship Attribution in Classic Literature* [online]. 2006 [cit. 2015-03-07]. Dostupné na: http://goanna.cs.rmit.edu.au/˜jz/fulltext/acsc07yz.pdf

[15] SEDLÁČIKOVÁ, B. *Historie matematické lingvistiky* [online]. Brno: Akademické nakladatelství CERM v Brně. 2012 [cit. 2015-01-07]. ISBN 978-80-7204- 815-1. Dostupné na: http://dml.cz/handle/10338.dmlcz/402330

Typické rysy autora

Metriky, ktoré potrebujeme merať, nazývame tiež rysy autora. Skúmajú sa charakteristiky, typické pre konkrétneho autora. Je nutné poznamenať, že jednotlivé charakteristiky sú rozdielne aplikovateľné pre každý jazyk. Hodnotené metriky uvádzame podľa Rygla[16]:

- rým a rytmika
- dĺžka slov
- dĺžka viet
- interpunkcia
- spájanie slov a používanie nespisovných slovných tvarov
- bohatosť slovnej zásoby
- frekvencia znakov
- pôvod slov
- chyby
- frekvencia slov
- pozícia slov vo vete
- syntax

Ak jednotlivé rysy autora spojíme, po spracovaní textov získame zoznam čísel, ktoré autora (dokument) charakterizujú. Tomuto zoznamu sa často hovorí odtlačok autora. S týmto odtlačkom potom pracujú nástroje strojového učenia. Podľa typy úlohy automatické štatistické metódy vytvorí klasifikátor. Klasifikátor môže riešiť rôzne úlohy. Typické je porovnanie, či bol dokument napísaný autorom (prípad porovnaní napadnutej závetu s korešpondenciou zosnulého), priradenie autorstva anonymnému dokumentu a zhlukovaniu podľa autorstva (spoj na diskusiu tých autorov, ktorí píšu rovnako). Klasifikátor potom vracia riešenie s pravdepodobnosťou správnosti odpovedí.[17]

Nepodarilo sa nám nájsť žiadne relevantné štúdie venujúce sa prisudzovaniu autorstva v slovenskom jazyku. Existujúca literatúra sa venuje iba oblasti autorského práva. V prípade

[16] RYGL, J. *Autorství a stylometrie* [online]. 2015 [cit. 2015-01-07]. Dostupné na: https://nlp.fi.muni.cz/cs/AutorstviAStylometrie
[17] RYGL, J. *Autorství a stylometrie* [online]. 2015 [cit. 2015-01-07]. Dostupné na: https://nlp.fi.muni.cz/cs/AutorstviAStylometrie

slovanských jazykov (konkrétne češtine) sa tejto oblasti venujú na Masarykovej univerzite v Brne v Centre spracovania prirodzeného jazyka. Ich vyvinuté metódy sa v praxi využívajú napríklad na Ministerstva vnútra v rámci Bezpečnostného výzkumu ČR.[16]

Automatická identifikácia jazyka

Úlohou tejto metódy je špecifikácia jazyka, v ktorom je dokument napísaný. Výsledok je možné dosiahnuť vybudovaním tabuliek so špecifikovanými frekvenciami dvojíc či trojíc písmen, typickými pre daný jazyk. Problematické sú pre tento typ spracovania krátke súbory, kde dochádza k veľkej odchýlke týchto frekvencií, ktoré tvoria podklad pre korektné určenie jazyka. Inou, skôr mechanickou metódou, môže byť implementácia slovníkov a následné porovnanie dokumentu na základe rozličných kritérií, ako napríklad slovníkové frázy, gramatika alebo diakritika.[18]

Záver

V článku sme odprezentovali najznámejšie a najvyužívanejšie aplikačné metódy text miningu pre neštruktúrované textové dáta. Poukázali sme na aktuálne príležitostí a možnosti využitia týchto metód v praxi a poskytli náhľad do dostupnej literatúry zaoberajúcej sa toto problematikou vo svete.

Zoznam bibliografických odkazov

JEŽEK, K. a J. STEINBERGER. *Sumarizace textů* [online]. 2010 [cit.2015-03-25]. Dostupné na: http://textmining.zcu.cz/publications/SumarizDATAKON.pdf

MEDHAT, et al. 2014. Sentiment analysis algorithms and applications: A survey. In: *Ain Shams Engineering Journal* [online]. 2014, **5**, 1093–1113. [cit. 2015-03-20]. ISSN 2090-4479. Dostupné na: http://ac.els-cdn.com/S2090447914000550/1-s2.0-S2090447914000550-main.pdf?_tid=b322cee8-d61f-11e4-8789-00000aacb35e&acdnat=1427639416_e6dbe68cef0a7c15058d2f2a1c649e61

[18] ŘEHŮŘEK, R. a M. KOLKUS. Language Identification on the Web: Extending the Dictionary Method. In: *Computational Linguistics and Intelligent Text Processing, 10th International Conference*. Mexico City: Springer-Verlag, 2009, s. 357-368. ISBN 978-3-642-00381-3.

MOENS, Marie-Francine. *Information Extraction: Algorithms and Prospects in a Retrieval Context. (The Information Retrieval Series).* London: Springer, 2006. ISBN 978-1-402-04987-3.
MURRAY, G. *The Rise of the Greek Epic*. 4th ed. London: Oxford University Press, 1967.
NETO, J. L., et al. *Automatic Text Summarization using a Machine Learning Approach* [online]. 2002 [cit.2015-03-25]. Dostupné na: http://www.cs.ukc.ac.uk/people/staff/aaf/pub_papers.dir/SBIA-2002- Joel.pdf
RYGL, J. *Autorství a stylometrie* [online]. 2015 [cit. 2015-01-07]. Dostupné na: https://nlp.fi.muni.cz/cs/AutorstviAStylometrie
ŘEHŮŘEK, R. a M. KOLKUS. Language Identification on the Web: Extending the Dictionary Method. In*: Computational Linguistics and Intelligent Text Processing, 10th International Conference*. Mexico City: Springer-Verlag, 2009, s. 357-368. ISBN 978-3-642-00381-3.
SASKO, J. *Analýza sentimentu výsledkov slovenských firiem* [online]. 2014 [cit. 2015-03-25]. Dostupné na: http://www.reputation.sk/wp-content/uploads/2014/02/ORMreport.pdf
SEDLÁČEK, P. *Text mining a jeho možnosti (aplikace)* [online]. 2003 [cit. 2015-02-27]. Dostupné na: http://www.fi.muni.cz/usr/jkucera/pv109/2003p/xsedlac5.htm
SEDLÁČIKOVÁ, B. *Historie matematické lingvistiky* [online]. Brno: Akademické nakladatelství CERM v Brně. 2012 [cit. 2015-01-07]. ISBN 978-80-7204- 815-1. Dostupné na: http://dml.cz/handle/10338.dmlcz/402330
ULDRICH, M. Text mining aneb Kladivo na nestrukturovaná data. In: *IT Systems* [online]. 2011, **12** [cit. 2015-03-12]. ISSN 1802-615X. Dostupné na: http://www.systemonline.cz/clanky/text-mining-kladivo-nanestrukturovana-data.htm
ZHAO, Y. a J. ZOBEL. *Authorship Attribution in Classic Literature* [online]. 2006 [cit. 2015-03-07]. Dostupné na: http://goanna.cs.rmit.edu.au/˜jz/fulltext/acsc07yz.pdf

Kontakt

Mgr. Jana Rusinková
Žilinská univerzita v Žiline
Fakulta humanitných vied
Katedra mediamatiky a kultúrneho dedičstva
email: jana.rusinková@mediamatika.sk