Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

FIIT-5208-46082

Bc. Peter Gašpar

PREPÁJANIE MULTIMEDIÁLNYCH METADÁT S VYUŽITÍM MIKROBLOGOVEJ SIETE

Diplomový projekt

Študijný program: Informačné systémy
Študijný odbor: 9.2.6 Informačné systémy
Miesto vypracovania: Ústav informatiky, informačných systémov a softvérového inžinierstva
Vedúci práce: Ing. Jakub Šimko PhD.

Máj 2016

Anotácia

Slovenská technická univerzita v Bratislave
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ
Študijný program: Informačné systémy

Autor: Bc. Peter Gašpar
Diplomový projekt: Prepájanie multimediálnych metadát s využitím mikroblogovej siete
Vedúci diplomového projektu: Ing. Jakub Šimko PhD.
Máj 2016

Metadáta sa stali obľúbeným prostriedkom ako charakterizovať informácie okolo nás, a to najmä v doméne multimédií a televízie. Mnohé existujúce prístupy sa opierajú o statický obsah, ktorý môžeme nájsť v rôznych znalostných bázach a encyklopédiách. Väčší potenciál však v sebe skrývajú sociálne siete. Sociálne siete sa vyznačujú väčšou dynamikou obsahu a najmä vzájomnou interakciou používateľov. Viaceré televízne spoločnosti ich používajú na propagáciu svojich televíznych relácií, pretože sú tiež prostriedkom na komunikáciu so svojimi divákmi. Často však priame prepojenie sociálneho obsahu na televízny program neexistuje. V našej práci prichádzame s metódou, ktorá využíva pomenované entity, hashtagy a externé prepojenia ako referenčné črty extrahované z príspevkov na sociálnej sieti. Tieto črty zároveň generujeme z televízneho programu pre vytvorenie mapovania medzi príspevkami a televíznymi reláciami. Navrhnutú metódu sme implementovali a overili na vybranej vzorke dát zo sociálnej siete. Získanými výsledkami sme potvrdili nami vyslovené hypotézy a zároveň sme preukázali prínos navrhnutých vylepšení v danej doméne.

Anotation

Slovak University of Technology in Bratislava
FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGY
Degree course: Information Systems

Author: Bc. Peter Gašpar
Diploma project: Linking Multimedia Metadata by Using Microblogging Network
Supervisor: Ing. Jakub Šimko PhD.
May 2016

Metadata has become a popular way to characterize any information, especially in the domain of multimedia and television. There are many approaches, which are based on the static content from the knowledge bases or encyclopaedias. A bigger potential is hidden in the Social Networking Services (SNS). SNSs are characterized by the highly dynamic content and users' interactions. Many television companies use SNSs to propagate their programmes, because they are also one of the most straightforward ways to get in touch with their audience. However, a straight connection between the social content and a television is often missing. In our work we are introducing a new approach that uses named entities, hashtags and external links from the SNS's posts. These features are also generated from the TV schedule to create mapping between the posts and the TV programmes. We have implemented and evaluated our approach by using annotated dataset of posts from the SNS. Presented results confirm our hypothesis and also reveal our contribution to the state of the art.

Poďakovanie

Chcel by som sa poďakovať svojmu školiteľovi Jakubovi Šimkovi za cenné rady, pripomienky, povzbudenie a odbornú pomoc, ktorú mi ochotne poskytol.

Čestné prehlásenie

Vyhlasujem, že som diplomovú prácu vypracoval samostatne s použitím uvedenej odbornej literatúry.

Bratislava 13.5.2016 .....

Bc. Peter Gašpar

Obsah

1 Úvod **1**

1.1 Televízia ako zaujímavá výskumná doména 1

1.2 Ciele práce 3

1.3 Slovník použitých pojmov a skratiek 3

2 Získanie a spracovanie informácií **5**

2.1 Televízny program ako východisko 5

2.2 Rozširujúce metadáta 5

2.3 Sociálne siete 6

2.4 Analýza obsahu zo sociálnych sietí 7

2.4.1 Stránky slovenských televízií 7

3 Reprezentácia a predspracovanie textu **11**

3.1 Text a jeho štruktúra 11

3.1.1 Vektor váh termov 12

3.1.2 N-gramy 13

3.1.3 Grafy N-gramov 13

3.1.4 Podobnosť textov 14

4 Prepájanie obsahu **15**

4.1 Klasifikácia sociálneho obsahu 16

4.1.1 Existujúce prístupy 17

4.2 Ko-učenie 19

4.3 Zhrnutie 21

5 Úvod k extrakcii metadát **23**

5.1 Detekcia tém 23

5.2 Detekcia sentimentu 25

5.3 Meranie sledovanosti 28

6 Prepájanie sociálneho a televízneho obsahu **31**

6.1 Úvod k mapovaniu 32

6.2 Vytvorenie televíznych metadát 33

6.2.1 Kandidátne pomenované entity 35

6.2.2 Kandidátne hashtagy 35

6.2.3 Prefixový index 36

6.3 Predspracovanie príspevku 37

6.3.1 Extrakcia pomenovaných entít 38

6.3.2 Extrakcia hashtagov 39
6.3.3 Spracovanie prepojení 39
6.4 Realizácia mapovania 41
6.5 Určenie výsledného mapovania 45
6.6 Rozšírenie štandardného mapovania 49
6.6.1 Vytvorenie dynamických televíznych metadát 50
6.6.2 Použitie dynamických televíznych metadát 50
6.7 Zhrnutie 52

7 Implementácia mapovania **53**
7.1 Architektúra riešenia 53
7.2 Moduly nižšej úrovne 54
7.3 Dátový model 55

8 Overenie mapovania **59**
8.1 Použité dátové množiny 59
8.2 Priebeh experimentov 61
8.3 Výsledky realizovaných experimentov 63
8.3.1 Prahová hodnota 63
8.3.2 Črty príspevku 65
8.3.3 Prefixový index 65
8.3.4 Ko-učenie 66
8.3.5 Úplne nastavenie 67
8.4 Porovnanie s existujúcimi prístupmi 69
8.5 Zhrnutie 70

9 Záver **71**

Literatúra **73**

A Technická dokumentácia **i**
A.1 Štruktúra dát príspevkov sociálnej siete i
A.1.1 Post i
A.1.2 Comment ii
A.2 Štruktúra dát televízneho programu iii
A.2.1 Channel iii
A.2.2 Programme iii
A.3 Detaily implementácie iv
A.3.1 Realizácia mapovania v
A.3.2 Anotovanie dátovej množiny vii

B Inštalácia a návod na použitie **ix**
B.1 Vytvorenie bázy televíznych metadát ix
B.2 Inštalácia ix
B.3 Realizácia mapovania x

C Príloha k overeniu riešenia **xiii**
C.1 Črty príspevku xiii
C.2 Časová náročnosť xiii

D Obsah elektronického média **xv**

E Návrh vedeckého článku na konferenciu ENIC 2016

Zoznam obrázkov

1.1 Aktivita vykonávaná na druhej obrazovke 2

2.1 Aktivita používateľov na Facebook stránke TV JOJ (2.2. - 15.2.2015) 8

6.1 Všeobecný pohľad na proces mapovania 33
6.2 Pohľad na nízkourovňové a vysokoúrovňové predspracovanie 34
6.3 Príklad invertovaného indexu pomenovaných entít 35
6.4 Príklad invertovaného indexu hashtagov 36
6.5 Príklad extrakcie pomenovaných entít z príspevku. Uvedená ukážka ilustruje hľadanie slov začínajúcich na veľké písmená a následné porovnanie s prefixovým indexom. 39
6.6 Ukážka rozdelenia URL na doménu a subdoménu 40
6.7 Pohľad na celý proces predspracovania 41
6.8 Ukážka výpočtu n-gramovej podobnosti s prefixom. 44
6.9 Ukážka mapovania pomocou hashtagov 46
6.10 Príklad mapovania pomocou hashtagov 46
6.11 Ukážka vytvorenia TF-IDF modelov 51
6.12 Celkový pohľad na proces mapovania 52

7.1 Diagram komponentov implementovaného riešenia 56
7.2 Diagram tried implementovaného riešenia 57

8.1 Graf F1-skóre pre rôzne prahové hodnoty 64

A.1 Ukážka zdrojového kódu mapovania vi
A.2 Ukážka mapovača vii

1. Úvod

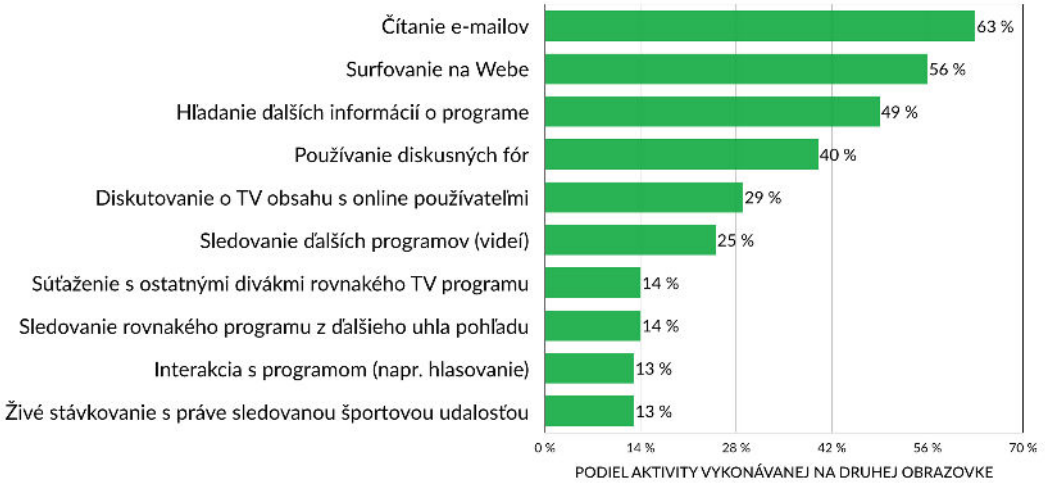
S rozmachom Webu sa informácie stali dôležitou súčasťou života ľudí. Či už sa zaujímame o obrázok, video, dokument alebo príspevok na sociálnej sieti, vždy sa snažíme objaviť niečo nové a zaujímavé. Každý objekt, ktorý sa vyskytuje okolo nás (napr. obrázok, video, a pod.) charakterizujeme prostredníctvom jeho špecifických vlastností - teda určitých overených informácií. Výskumníci z celého sveta hľadajú spôsoby, akými možno získať tieto informácie tak, aby sa proces spoznávania zrýchlil, uľahčil a v neposlednom rade aj skvalitnil.

O každom objekte môžeme povedať niekoľko základných vlastností. V našej práci sa primárne zaujímame o **multimediálne objekty** a ich charakteristické vlastnosti budeme spoločne nazývať **metadáta**. Práve oblasť multimédií predstavuje výhodu v množstve dostupných metadát, pričom mnohé z nich máme k dispozícii už z priameho skúmania určitého objektu. V prípade obrázka ide napríklad o výšku, šírku, miesto odfotografovania a mnohé iné. Tieto základné informácie môžu byť síce v mnohých prípadoch užitočné, no nielen pre zvedavca, ale aj pre ďalšie skúmanie, spravidla nepostačujúce.

Vhodným príkladom ilustrujúcim potrebu rozširujúcich metadát môže byť napríklad televízny program. Základné informácie o jednotlivých televíznych reláciách pozostávajú najmä z názvu, anotácie, obsadenia a tiež času začiatku a konca vysielania. Dôležitou informáciou pre diváka však môže byť napríklad aj hodnotenie danej relácie, čo má význam nielen u filmov ale aj u opakovane vysielaných programov (napr. seriálov). Popri hodnotení však zohráva významnú úlohu aj ďalší dostupný obsah: články, fotografie, udalosti, videá, a pod. Doménu televízie sme si zvolili ako hlavnú doménu pre výskum v oblasti získavania multimediálnych metadát.

1.1 Televízia ako zaujímavá výskumná doména

Klasické televízne vysielanie sa v posledných rokoch stretáva s náročnou konkurenciou v podobe rôznych videoslužieb na sledovanie, nahrávanie a zdieľanie (online) obsahu. Významnú rolu v tomto smere hrá najmä popularita a rozšírenie Internetu a rôznych multimediálnych zariadení - notebookov, mobilných telefónov a tabletov. V poslednej dobe sa však táto - na prvý pohľad - konkurencia skloňuje skôr ako pomocný nástroj pre televízne vysielanie. V tejto súvislosti sa stretávame s pojmom **Druhá obrazovka** (*angl.* Second screen). Pod druhou obrazovkou si môžeme predstaviť práve už spomínané multimediálne zariadenia, ktoré používajú diváci súbežne so sledovaním televízneho vysielania.



Obr. 1.1: Aktivita vykonávaná na druhej obrazovke

Na zariadeniach pritom vykonávajú rôznu činnosť, ktorá vôbec nemusí, ale často práve súvisí s televíznym vysielaním. Ako ukázal prieskum z roku 2013 [14] až 75% divákov využíva mobilné zariadenia počas sledovania televízie. Najčastejšou aktivitou sa s 63% percentami stalo čítanie e-mailov, na druhom mieste sa ocitlo používanie mobilných aplikácií a surfovanie na Internete s 56%, pričom tretiu priečku s 49% obsadilo používanie aplikácií/Internetu na hľadanie ďalších informácií o sledovanom programe. Celkovú prehľadovú štatistiku zobrazuje graf na obrázku č. 1.1.

Prieskum taktiež odhalil dôležitosť používateľsky-generovaného obsahu, a to najmä rôznych edukatívnych videí z videoslužby YouTube, kde návody a recenzie sledovalo (všeobecne, nielen cez druhú obrazovku) až 82% účastníkov.

Moderné technológie umožnili a podnietili tiež rozmach **nelineárneho vysielania**. To sa od lineárneho líši v tom, že divák má možnosť zasahovať do sledovaného programu - pozastavením, posunom v programe, prípadne výberom času, v ktorom si daný program spustí. Nelineárne vysielanie sa vyskytuje najčastejšie ako súčasť doplnkových služieb sprostredkovateľov televízneho vysielania, a to najmä ako súčasť *videa na požiadanie* (*angl.* video-on-demand, VOD) alebo televízneho archívu.

Význam prepojenia televízie a Internetu diskutujú autori v článku [4]. Podľa súčasných prieskumov až 34% amerických domácností vlastní inteligentný televízor, 20% vlastní zariadenie na prenos videa (napr. Chromecast[1] alebo Apple TV[2]) a až 62% hracie konzoly. Navyše 60% domácností, ktoré vlastnia televíziu, sledujú vysielanie aj cez Internet.

Z uvedených skutočností môžeme vyvodiť nasledujúce závery:

[1] https://www.google.com/chromecast
[2] http://www.apple.com/sk/tv/

- diváci sa počas sledovania televízie snažia získať ďalšie informácie o sledovanej televíznej relácií,
- diváci často diskutujú sledovaný TV obsah alebo sa iným spôsobom zapájajú do aktivít súvisiacich s televíznou reláciou.

Zaujímavý priestor pre aktivitu divákoch pritom môžeme nájsť v prostredí sociálnych sietí, ktorých popularita v súčasnosti ešte stále rastie. Vzniká nám však zároveň problém ako prepojiť dva rôzne svety, a to svet televízie a svet sociálnych sietí.

1.2 Ciele práce

Metadáta z domény televízie, ktoré divák získava prostredníctvom systémov EPG alebo tlačených či internetových programových plánov, často nie sú postačujúce. Cieľom našej práce je preskúmať použitie nového prvku - sociálnej siete - na sprístupnenie metadát, ktoré môžu obohatiť existujúce statické prístupy. V nasledujúcich kapitolách sa pozrieme na túto problematiku z pohľadu viacerých existujúcich prístupov. Významnú časť práce v tejto fáze venujeme hľadaniu spôsobu, ako prepojiť televízny a sociálny obsah, čím sa nám sprístupnia možnosti ako zo sociálneho obsahu extrahovať nové metadáta.

1.3 Slovník použitých pojmov a skratiek

V tejto podkapitole uvádzame pojmy a skratky používané v našej práci.

EPG - elektronický programový sprievodca (*angl.* electronic programme guide)

hashtag - krátka slovná značka, ktorej primárnym cieľom je vytvoriť určitú základnú kategorizáciu textu, v ktorom sa nachádza. Hashtag začína symbolom „#“ a obsahuje výhradne alfanumerické znaky (zvyčajne bez interpunkcie a diakritiky).

príspevok - textový alebo multimediálny prejav na sociálnej sieti (spoločný pojem pre „status“, „tweet“ alebo akýkoľvek iný prejav, ktorý používateľ zverejní priamo na svojom používateľskom profile).

sociálny obsah - používateľský obsah prístupný na sociálnej sieti. Medzi sociálny obsah patria najmä príspevky.

televízny program - zoznam časovo zoradených položiek, ktoré má v pláne vysielať televízia.

televízna relácia - prvok z televízneho programu. V našej práci zahŕňame do tohto pojmu: filmy, seriály, relácie - akýkoľvek úsek z vysielania televízie (okrem

reklamných spotov).

tweet - príspevok na sociálnej sieti Twitter[3].

[3] www.twitter.com

2. Získanie a spracovanie informácií

Procesu dolovania metadát v doméne televízneho vysielania predchádza získanie pôvodného obsahu, z ktorého budeme dolovať. Či už ide o textové alebo multimediálne dáta, dôležitým krokom je výber kvalitného a použiteľného zdroja informácií. V tejto kapitole sa bližšie pozrieme na procesy spojené so získavaním a predspracovávaním informácií, ktoré plnia kľúčovú úlohu v ďalšom postupe.

2.1 Televízny program ako východisko

V oblasti televízneho vysielania sa stretávame s obsahom rôzneho typu - dokumentmi, seriálmi, filmami či reláciami, ale i rôzneho žánru, od komédie cez romantiku až po horor. Ucelený pohľad na dostupný obsah nachádzame v zoznamoch televízneho vysielania nazývaných tiež **televízny program**.

Pod televíznym programom TV stanice rozumieme chronologicky zoradený zoznam televíznych relácií pre konkrétnu stanicu. Tento zoznam pritom spravidla obsahuje okrem základných informácií (názov TV relácie, začiatok a koniec vysielania) aj rozširujúce: popis a zoznamy hercov, producentov či režisérov. Rozširujúce informácie môžeme v tomto prípade považovať za statické metadáta. Spravidla ich totiž dodávajú priamo vysielatelia a zachytávajú obsah konkrétnej relácie. Špeciálnym prípadom sú seriály, pri ktorých sa často stretávame s opisom jednotlivých ich častí. V dobe rozmachu Internetu však už tieto informácie nie sú postačujúce a potrebujeme preto prekročiť hranice televízneho programu tak, ako ho poznáme.

2.2 Rozširujúce metadáta

Problematiku metadát sme načrtli v úvodnej kapitole práce. Mutlimediálny obsah z domény televízie má bohaté zastúpenie v rôznych bázach metadát. V súčasnosti možno za jednu z najznámejších považovať databázu DBpedia[1], ktorá obsahuje prepájané dáta pre viac ako 87 000 filmových titulov a viac ako 29 000 televíznych relácií. Za jednu z hlavných nevýhod pokladáme absenciu slovenskej lokalizácie pre túto oblasť. Taktiež, hoci sa snaží používateľská komunita (dav) databázu aktívne udržiavať, s využívaním manuálnych anotácií sú spojené isté riziká. Jednou z nevýhod tohto prístupu je možná subjektívnosť autorov anotácií. Ďalšia nevýhoda vyplýva priamo zo statickosti spracovávaného obsahu. Statický obsah nedokáže pru-

[1] http://wiki.dbpedia.org/

žne reagovať na dynamiku, ktorá je charakteristická pre súčasný svet a prostredie Webu. Ak sme závislí na statickom obsahu, metadáta sa môžu stať po čase neplatnými. Riešenie tohto problému predstavuje spojenie statického (overeného) obsahu s dynamickým.

Tento prístup nachádzame aj v mnohých komerčných webových portáloch, ktoré prezentujú metadáta vizuálne prívetivou formou pre používateľa. Tie sú navyše obohatené o dynamické prvky, akými sú hodnotenie a používateľské či odborné komentáre. Medzi takéto portály patrí napríklad IMDB.com[2] alebo v slovenskom prostredí známy CSFD.cz[3].

2.3 Sociálne siete

Fenoménom 21. storočia sa na Webe stalo publikovanie používateľského obsahu. Blogy umožnili zdieľať názory a myšlienky ľudí v podobe textov a multimédií a podnietili ich tak k väčšej tvorivosti. Neskôr sa stali populárne aj mikroblogy, ktoré sa zameriavali na kratšie texty. Z mikroblogov sa postupne vyvinuli sociálne siete (*angl.* social networking services), ktoré sa v oblasti sémantického Webu stali jeden z najskúmanejších fenoménov. Ich výhoda spočívala v tom, že sa na nich najviac prejavovala dynamika v podobe aktivity používateľov. Zároveň, ako už z názvu vyplýva, kládli dôraz na spájanie používateľov (vytvorenie akejsi siete). Používatelia sa tak vďaka sociálnym sieťam stali primárnymi tvorcami a širiteľmi zaujímavého obsahu.

Jadro sociálnej siete spravidla tvorí práve obsah, ktorý tiež nazývame **príspevok**. Keď hovoríme o príspevkoch, najčastejšie máme na mysli textové vyjadrenia používateľov (nazývané tiež statusy), fotografie alebo odkazy na iné webové stránky, ktoré používatelia na sociálnej sieti zdieľajú. Iní používatelia sociálnej siete navyše na takéto príspevky reagujú, a to najčastejšie vo forme komentárov, čím sa dynamika ešte viac zvýrazňuje. Problémom, s ktorým sa musia mnohí výskumníci zaoberať, je kategorizácia takýchto príspevkov.

V súčasnosti sa čoraz častejšie stretávame s používaním tzv. **hashtagov** (*angl.* hashtag). Hashtagom nazývame krátku slovnú značku, ktorej primárnym cieľom je vytvoriť určitú kategorizáciu príspevku. Každý hashtag sa začína symbolom „#“ a spravidla obsahuje iba alfanumerické znaky (bez medzier). Túto techniku pritom používajú samotní tvorcovia príspevkov - ich motivácia spočíva v tom, že iní používatelia, ktorí sociálnu sieť prehľadávajú, nájdu ich príspevok rýchlejšie a jednoduchšie.

Ďalšou nemenej dôležitou charakteristikou sociálnych sietí sú entity, ktoré sa

[2] http://www.imdb.com/
[3] http://www.csfd.cz/

často v príspevkoch vyskytujú. Používatelia pritom uvádzajú mená ľudí, názvy miest alebo obľúbené knižky priamo v texte alebo využívajú nepriame prepojenie pomocou zdieľaných článkov, videí alebo iných odkazov na externé stránky. Tieto prepojenia pritom v sebe nesú nemalú informačnú hodnotu, nakoľko príspevku na sociálnej sieti môžu dodať kontext.

Podľa [13] patria celosvetovo medzi najpopulárnejšie (mikroblogové) sociálne siete Facebook[4] a Twitter[5]. V prípade Slovenska je situácia mierne odlišná, kde sa podľa [2] spomedzi svetových sociálnych sietí umiestnili Facebook na 2. mieste a Twitter na 39. mieste. Obe služby umožňujú svojim používateľom zdieľať svoj každodenný život. Primárnym prostriedkom na zdieľanie sú v prípade **Twittera** krátke, najviac 140 znakové správy, nazývané tiež tweety (*angl.* tweets). Limit v počte znakov však nemusí vždy predstavovať obmedzenie. Tento problém sa na Twitteri odstraňuje používaním hashtagov, ktorými sa tiež dosahuje prirodzená kategorizácia obsahu. Hoci **Facebook** neobmedzuje používateľa[6] až tak striktne, hashtagy na ňom nie sú až natoľko populárne. Obe služby pritom podporujú vytváranie špeciálnych fanúšikovských profilov pre osobnosti, udalosti, produkty, televízie (televízne relácie) a iné. Keďže ide o používateľský obsah, nie je vždy postačujúce spoliehať sa iba na hashtagy alebo kategorizáciu medzi jednotlivé profily. Na tento účel je nevyhnutné pracovať priamo s prirodzeným jazykom.

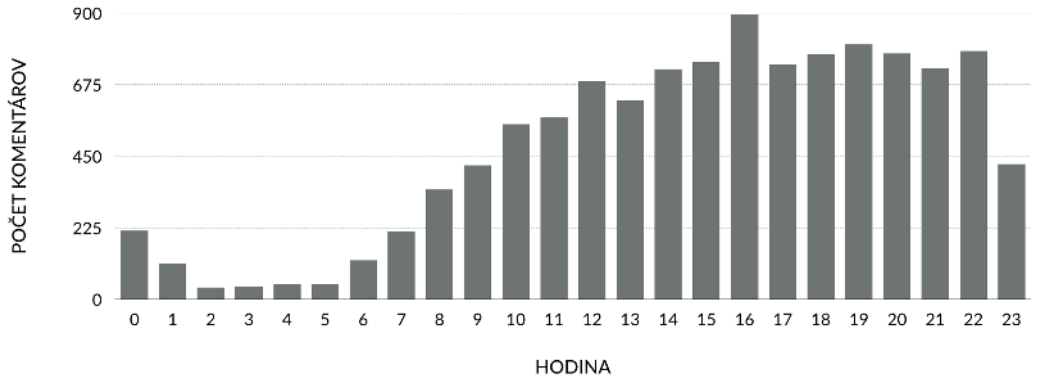
2.4 Analýza obsahu zo sociálnych sietí

V našich hypotézach sme považovali sociálne siete za primárny zdroj dynamických informácií pre získavanie metadát. Tieto hypotézy sme však museli podložiť preskúmaním denne publikované obsahu. V tejto podkapitole uvádzame zistenia, ku ktorým sme dospeli automatickým a manuálnym analyzovaním fanúšikovských stránok na sociálnej sieti Facebook.

2.4.1 Stránky slovenských televízií

Pri skúmaní slovenských televízií sme sa zamerali na dve najsledovanejšie[7] celoplošné komerčné televízne stanice - TV Markíza[8] a TV JOJ[9]. Analyzovali sme hlavné fanúšikovské stránky oboch televízií, ale taktiež aj vybrané podstránky ich najsledo-

[4] http://www.facebook.com
[5] http://www.twitter.com
[6] Maximálny počet znakov na príspevok je viac ako 63 000. Zdroj: http://blog.hubspot.com/marketing/character-limit-social-media-blog-posts (platné k 9.5.2016)
[7] Ku dňu 6. 12. 2015 podľa údajov spoločnosti PMT, s.r.o. (http://www.pmt.sk).
[8] http://www.markiza.sk
[9] http://www.joj.sk



Obr. 2.1: Aktivita používateľov na Facebook stránke TV JOJ (2.2. - 15.2.2015)

vanejších vlastných seriálov, a to v priebehu 2 týždňov (2.2. - 15.2.2015).

Vo všeobecnosti vyvíjali obe televízie najvyššiu aktivitu najmä v priebehu dňa a nové príspevky pridávali spravidla v 30-minútových intervaloch. Medzi najčastejšie príspevky televízneho obsahu patrili:

- zákulisné články k vysielaným programom,
- fotografie a videá zo zákulisia,
- upútavky na program,
- rôzne súťaže,
- fotografie a články, ktorých predmetom boli známe osobnosti, herci a herečky.

Pozitívom, s ktorým sme sa stretávali, bolo živé publikovanie fotografií z priamo vysielaného televízneho filmu, seriálu alebo živého prenosu. Divák, ktorý tento program nesledoval (prípadne nemal možnosť sledovať), si mohol vytvoriť predstavu o deji, zapojiť sa do diskusie alebo ho daný program mohol upútať. Častým (a negatívnym) javom bolo publikovanie bulvárnych článkov a tiež rôznych obrázkov, ktoré s televíznym prostredím nemali súvis. Snahou televízie bolo v tomto prípade vyvolať u diváka pozornosť a získať tak cenné označenia „Páči sa mi to“, komentáre a zdieľania. Príspevky tohto charakteru pre nás predstavujú istú formu šumu.

Okrem príspevkov, ktoré odosielali samotné televízie, sme preskúmali aj komentáre používateľov. Ako vidíme na grafe (obrázok č. 2.1), najväčšia aktivita sa odohrávala medzi 12 - 22 hodinou, kedy bol absolútny počet komentárov zároveň najvyšší.

Medzi najčastejšie činnosti, ktoré vyplynuli z obsahu komentárov, patrili:

- komentovanie aktuálneho deja programu - niekedy však bez ohľadu na obsah

(televíziou) zdieľaného príspevku,

- vyjadrovanie názorov na publikované fotografie a videá - nie zriedkavým javom bola negatívna kritika, v prípade, že išlo o fotografiu osobnosti,
- diskusia k postavám filmu alebo seriálu.

Mnohé komentáre obsahovali názory používateľov a tiež v sebe zahŕňali sentiment - či už vyplývajúci z textového obsahu alebo použitých emotikonov. Častým javom, ktorý sa vyskytoval najmä pri pravidelne vysielaných programoch, boli sťažnosti divákov na (negatívne) zmeny v čase ich vysielania.

Popri slovenských televíziách sme analyzovali aj stránky svetových televízií, a to konkrétne zameraných na dokumentárny žáner: Discovery Channel, History, Animal Planet a Travel Channel. Keďže lokálne verzie stránok na Facebooku (v čase nášho výskumu) buď neexistovali alebo vykazovali nízku aktivitu, zvolili sme si oficiálne stránky, ktorých obsah bol v angličtine.

Vzhľadom na charakter televíznych staníc, aj obsah príspevkov na sociálnej sieti jednotlivých staníc bol veľmi podobný. Stránky pravidelne publikovali fotografie a články, ktoré odkazovali na niektorú z relácií, a to či už priamo (napr. odkazom na webovú stránku relácie alebo formou hashtagu) alebo nepriamo (napr. v prípade Animal Planet obrázky rôznych zvierat bez uvedenia konkrétnej relácie). V porovnaní so slovenskými televíziami bol obsah na vyššej kvalitatívnej úrovni, avšak komentáre používateľov už neobsahovali žiaden sentiment[10] alebo odkazy na pomenované entity. Ďalší, avšak menej závažný problém predstavuje potreba mapovania anglického obsahu (najmä relácií s anglickým názvom) na slovenský televízny program.

Na príspevky zo sociálnej siete sme sa pozreli aj z pohľadu **štruktúry**. Identifikovali sme pritom tri základné črty, ktoré sa v nich vyskytovali a ktoré sme už diskutovali v predchádzajúcej kapitole: **pomenované entity, hashtagy a externé prepojenia**. Väčšina príspevkov vždy obsahovala aspoň dve z čŕt, pričom najčastejšie sme sa stretávali s pomenovanými entitami a externými prepojeniami.

Nemožno pritom úplne jednoznačne prehlásiť, ktorá z čŕt prevládala, keďže to záviselo od toho, s ktorou televíziou sme pracovali. Oficiálna Facebook stránka americkej televízie CBS mala iba v 13% príspevkov použité hashtagy, ale externé odkazy sme identifikovali až v 90% príspevkov. Naopak, v prípade slovenskej televízie Markíza sme sa stretli s použitím hashtagu v takmer každom príspevku. Takéto správanie iba dokazuje, že tvorcovia obsahu využívajú viaceré prvky, ktoré často kombinujú. Preto pri návrhu riešení, ktoré sa opierajú o príspevky zo sociálnej siete, musíme uvažovať súčasne viaceré možnosti.

[10] S výnimkou komentárov k obrázkom zvierat.

3. Reprezentácia a predspracovanie textu

Obsah získaný zo sociálnej siete môže mať rôzny charakter: absolútne číslo (napr. počty obľúbení, komentárov, zdieľaní), inštanciu entity (napr. používateľa, ako autora príspevku) alebo text (napr. text príspevku, komentára). Definovať správnu reprezentáciu textu pritom predstavuje často zložitú, no dôležitú úlohu. Text príspevkov na sociálnych sieťach v sebe totiž skrýva potenciál, ktorý sa budeme snažiť v nasledujúcich úvahách odhaliť.

3.1 Text a jeho štruktúra

Súvislý text, a to bez ohľadu na jeho dĺžku, môžeme spravidla rozdeliť na samostatné slová - reťazce znakov. Pre mnohé úlohy využívajúce text ako primárny zdroj informácií je však takáto reprezentácia nepostačujúca v prípade, že ide o text v prirodzenom jazyku. Prirodzený jazyk je totiž charakteristický svojou rôznorodosťou, a to najmä z dvoch hľadísk:

- **morfológie** (tvaroslovia) - vyplývajúce z využívania rôznych slovných druhov (napr. počítanie, počítač, počítač, počítaný, ...),
- **sémantiky** (významu slov) - vyplývajúce z častého výskytu javov, akými sú polysémia, homonymia, synonymia, a pod.

Aby sme vedeli text vhodne reprezentovať, je nevyhnutné ho rozdeliť na slová a tie následne normalizovať. Tento proces sa tiež nazýva **tokenizácia** a jeho výsledkom sú reťazce znakov - **tokeny** - oddelené od predchádzajúceho a nasledujúceho reťazca medzerou (definované v [15]). Využívajú sa pritom súčasne viaceré prístupy. Medzi najpoužívanejšie zaraďujeme:

- **odstránenie tzv. stop slov** - teda odstránenie slov, ktoré sa vyskytujú v texte s vysokou frekvenciou a nemajú vplyv na jeho sémantický význam (napr. spojky, predložky),
- **lematizácia** (*angl.* lemmatization) - hľadanie slovníkového tvaru slova (napr. kritickejší - kritický),
- **stemming** - hľadanie koreňa slov (napr. ženatý - žen).

V prípade anglického textu existujú mnohé algoritmy, ktoré tieto úlohy dokážu vyriešiť. Hoci je v slovenčine vzhľadom na komplexnosť jazyka táto úloha náročnejšia, existujú práce, ktoré sa jej podrobne venovali a celý proces môžu uľahčiť. Príkladom

je práca [15] Garabíka, a kol. alebo Lematizátor slovenského jazyka[1].

Pri spracovávaní textového obsahu je nevyhnutné zvoliť vhodnú reprezentáciu. Dôležitosť tohto kroku tkvie najmä v tom, aký by mohla mať nesprávna reprezentácia negatívny dopad na ďalšie spracovávanie - či už z hľadiska pamäťovej náročnosti alebo výpočtovej zložitosti aplikovanej metódy. Taktiež musíme mať vedomosť o tom, akú úlohu budeme pri práci s textom riešiť. Medzi najčastejšie úlohy v oblasti práce s prirodzeným jazykom patrí hľadanie podobnosti medzi dokumentmi. Či už ide o webový vyhľadávač alebo nástroj, ktorý sa snaží klasifikovať dokumenty na témy, je navyše potrebné pri voľbe reprezentácie myslieť aj na spracovávaný obsah. V našom prípade budeme pracovať s obsahom zo sociálnej siete. V nasledujúcich podkapitolách sa pozrieme na dva najčastejšie prístupy využívané v tejto oblasti. Vychádzali sme pritom z práce [1].

3.1.1 Vektor váh termov

Mnohé reprezentácie dokumentov vychádzajú z predstavy dokumentu ako vektora reálnych čísiel (najčastejšie váh). Klasickým prístupom je model vektoru termov (*angl.* term vector model). Jeho najčastejšie využitie nachádzame najmä v oblasti objavovania znalostí (*angl.* knowledge discovery) v prípade, že hľadáme dokumenty relevantné pre nejaký dopyt. Tento dopyt pritom spravidla pozostáva najmä z určitých kľúčových slov. Slová (alebo skupiny slov) pochádzajúce z dokumentu a dopytu nazývame spoločne **termy**. Skupiny termov následne reprezentujeme pomocou **vektora váh**. Pri určovaní hodnoty každej váhy termu v dokumente alebo dopyte môžeme vychádzať z dvoch modelov:

- **binárny model** - váha nadobúda hodnotu 1, ak sa term v dokumente vyskytuje a váhu 0, ak sa v dokumente nevyskytuje,
- **pravdepodobnostný model TF-IDF** - váha termu je vypočítaná pomocou metódy TF-IDF.

Pri **metóde TF-IDF** (s myšlienkou prichádzajú Salton a McGill v [29]) sa pri výpočte relevancie termu vo vektore berie do úvahy výskyt termu v rámci samotného dokumentu a tiež naprieč viacerými dokumentami. Rozlišujeme pritom:

- **frekvenciu termu** (TF, *angl.* term frequency), ktorá vyjadruje počet výskytov termu v dokumente. Vychádza z myšlienky, že často opakujúci sa term môže byť pre dokument viac relevantný ako term s mizivým výskytom.
- **inverznú dokumentovú frekvenciu** (IDF, *angl.* inverse document frequency), ktorá penalizuje termy, ktoré sa často vyskytujú vo viacerých doku-

[1] http://www.forma.sk/tech-jm-lem.aspx

mentoch. IDF preto môže pomôcť pri minimalizovaní tzv. „stop“ slov.

Výstupom metódy je matica, ktorá pre každý term t (z množiny termov) stanovuje jeho váhu TF-IDF v danom dokumente d (z množiny dokumentov). Pre určenie miery podobnosti termov v dopyte a termov v dokumente používame **koeficient podobnosti** (*angl.* coeficient of similarity). Koeficientom podobnosti teda určíme, do akej miery sa dopyt vzťahuje na skúmaný dokument.

3.1.2 N-gramy

Ďalší často využívaný model reprezentácie dokumentov predstavujú **n-gramy**. N-gramy sú časti slov pevne stanovenej dĺžky **n**, ktoré vzniknú z pôvodného dokumentu postupným rozdeľovaním slov. Medzery medzi slovami nahrádzame spravidla špeciálnym symbolom, najčastejšie podčiarkovníkom („_“). Medzi najpoužívanejšie n-gramy zaraďujeme unigramy ($n = 1$), bigramy ($n = 2$) a trigramy ($n = 3$). Napríklad z textu „Dobrý deň“ získame nasledujúce trigramy (teda n-gramy dĺžky 3):

dob, obr, brý, rý_, ý_d, _de, deň

Alternatívnym prístupom je tiež rozdelenie dokumentu na **n-gramy slov**. V tomto prípade dĺžka n nebude predstavovať celkovú dĺžku n-gramu, ale **počet slov** n-gramu. Postupujeme teda podobne ako v predchádzajúcom príklade, avšak teraz nepracujeme s postupnosťou znakov, ale s postupnosťou slov. Napríklad z textu „Dobrý deň, ako sa máte“ získame nasledujúce trigramy (za predpokladu vylúčenia interpunkčných znamienok):

dobrý deň ako, deň ako sa, ako sa máte

V porovnaní s technikou váhovania pomocou TF-IDF, hlavná výhoda n-gramov spočíva najmä vo väčšej tolerancii preklepov a chýb. Podobne ako termy, aj n-gramy môžeme reprezentovať vektorom váh.

Autori v [1] vhodne poznamenali, že ak nezohľadníme aj poradie n-gramov v texte, môže dôjsť k strate pôvodného významu. Príkladom je slovo „wiki“, ktoré pri n-gramoch veľkosti 2 môžeme ľahko zameniť za slovo „kiwi“. Riešenie tohto problému môžeme nájsť v grafoch n-gramov.

3.1.3 Grafy N-gramov

Obohatením štandardného modelu n-gramov sú grafy n-gramov prezentované v [17]. Graf n-gramov pritom vyjadrujeme nasledovne:

- uzly predstavujú samotné n-gramy,
- (ohodnotené) hrany spájajú n-gramy, pričom sú proporcionálne ohodnotené

podľa ich vzájomnej vzdialenosti (vzhľadom na pôvodný text).

Pri konštrukcii grafu sa analyzuje vstupný text postupne pretváraný do n-gramov, ktoré sa prekrývajú. Zároveň sa ukladajú informácie o susedných uzloch, čím môžeme získať prehľad o tom, ktorá hrana a s akou vzdialenosťou (resp. mierou blízkosti) spája dvojicu n-gramov. Priamej aplikácii tohto modelu sa venujeme v podkapitole *Detekcia sentimentu*.

3.1.4 Podobnosť textov

Z pohľadu hľadania podobnosti medzi textami máme k dispozícii opäť viacero prístupov. Na úrovni dokumentov sa najčastejšie stretávame s modelom TF-IDF a modelom n-gramov. Ako sme už predtým uviedli, problémy v modeli TF-IDF spôsobuje najmä intolerancia preklepov alebo chýb v slovách. V doméne sociálnych sietí sa preto častejšie stretávame s používaním modelu n-gramov.

Pri hľadaní podobnosti medzi dvoma dokumentmi analyzujeme získané n-gramy. Grzegorz Kondrak v [22] zhrnul najpoužívanejšie riešenia využívané pre analýzu podobnosti reťazcov. Keďže v prípade sociálnych sietí sú príspevky používateľov spravidla relatívne krátke (často iba 1 rozvitá veta), môžeme dokument ďalej uvažovať ako reťazec znakov. Referenčnou a veľmi často používanou metódou je **Levenshteinova vzdialenosť**, ktorú definujeme ako minimálny počet elementárnych úprav potrebných na pretransformovanie z prvého reťazca na druhý reťazec. Obdobou je **Najdlhšia spoločná podsekvencia** (*angl.* Longest common subsequence), teda hľadanie najdlhšieho spoločného podreťazca oboch reťazcov. Medzi ďalšie prístupy patrí získanie počtu spoločných n-gramov medzi reťazcami, ale taktiež sa môžeme pozrieť aj na postupnosť n-gramov alebo ich vzájomnú vzdialenosť v dokumentoch.

Kondrak v [22] spojil viacero prístupov a navrhol algoritmus, ktorý umožňuje efektívne vypočítať podobnosť (alebo, naopak, vzdialenosť) medzi dvoma reťazcami znakov založený na modeli n-gramov. Tento algoritmus sa pritom snaží eliminovať viaceré nevýhody predchádzajúcich prístupov, ktorými boli najmä zanedbanie poradia n-gramov v texte a ich vzájomnej vzdialenosti. Metóda navyše zahŕňa aj normalizácie z pohľadu dĺžky reťazcov a kladie tiež vyšší dôraz na n-gramy, ktoré sa nachádzajú na začiatku reťazca (spravidla sa jedná o predpony slov).

4. Prepájanie obsahu

Významným krokom v procese obohacovanie metadát k multimediálnemu obsahu je správna klasifikácia informácií a príspevkov získaných zo sociálnych sietí. Pod klasifikáciou v doméne sociálnych sietí rozumieme priradenie príspevku sociálnej siete do kategórie. Touto kategóriou pritom môže byť na jednej strane téma, ale taktiež aj pomenovaná entita - osoba, miesto alebo udalosť, ktorá sa v príspevku priamo alebo nepriamo nachádza.

V doméne televízneho vysielania uvažujeme najmä dve základné typy pomenovaných entít:

- konkrétnu televíznu **reláciu** (napr. konkrétny názov seriálu, filmu, relácie, ...),
- **osoby** súvisiace alebo priamo vystupujúce v televíznej relácii (napr. herci, moderátori, režiséri, ...).

Úlohou **mapovania** je identifikovať a priradiť skúmané entity k textom a rozhodnúť o tom, do akej miery spolu navzájom súvisia. Pristupovať k tejto úlohe môžeme dvoma spôsobmi:

- hľadaním **podobnosti** medzi (známou) pomenovanou entitou a samotným dokumentom,
- **rozpoznaním** / identifikáciou (neznámej) pomenovanej entity v skúmanom dokumente a následným porovnaním s databázou pomenovaných entít. Tento proces tiež nazývame **rozpoznávanie pomenovaných entít** (*angl.* named-entity recognition).

Mnohé riešenia vychádzajú najmä z prvého prístupu, kedy známu pomenovanú entitu získavajú z existujúcich bázy znalostí. Spravidla však lepšie výsledky dosiahneme v kombinácii s druhým prístupom, kedy sa snažíme v textoch najskôr potenciálne entity identifikovať a následne ich namapovať na konkrétne inštancie v báze znalostí, čím im zároveň priradíme aj pomenovanie (teda typ entity, napr. herec).

Rozpoznávaniu pomenovaných entít v slovenských textoch sa práve s využitím tejto kombinácie venovali Kaššák a kol. v [21]. Vo svojej práci sa zameriavali na novinové články dostupné z rôznych zdrojov na Webe, pričom sa im podarilo rozpoznať viaceré kategórie pomenovaných entít s priaznivou úspešnosťou. Pomenované entity v textoch extrahovali prostredníctvom vybraných pravidiel. Extrahované entity následne porovnávali s bázou znalostí - Wikipediou[1].

Li a kol. sa v [24] na proces rozpoznávania pozreli cez dvojvrstvový prav-

[1] http://www.wikipedia.org

depodobnostný model, ktorý pozostával z dvoch postupných krokov: (1) nájdenie relevantných dokumentov a (2) extrakcia entít z nájdených relevantných dokumentov. Autori sa svojím návrhom snažili skvalitniť prácu internetových vyhľadávačov, ktoré spravidla ponúkajú ako odpoveď na zvedavcove dopyty dokumenty (webové stránky) a nie konkrétne odpovede (entity).

4.1 Klasifikácia sociálneho obsahu

Špeciálnym prípadom klasifikácie textov je klasifikácia príspevkov na sociálnych sieťach. Častým problémom existujúcich prístupov je najmä závislosť od jazyka, na ktorý sú zamerané. A to nielen z pohľadu špecifík jednotlivých národov a kultúr, ale taktiež aj špecifík prostredia, v ktorom sa jazyk používa. Pri skúmaní obsahu sociálnych sietí sa iba zriedkavo stretávame s používaním spisovnej reči. Pri identifikácii pomenovaných entít je preto nevyhnutné zamerať sa aj na nasledujúce problémy:

Mnohojazyčnosť. V prípade, že chceme, aby bol navrhovaný prístup prenositeľný bez ohľadu na spracovávaný jazyk, nestačí sa spoliehať na vysokú popularitu vybraného (napr. anglického) jazyka. Naopak, ak sa zameriame na konkrétny jazyk s minoritným zastúpením (napr. slovenský jazyk), nesmieme zabudnúť na možný výskyt slov aj z iných jazykov.

Slang a neologizmy. S rastúcim počtom používateľov sociálnych sietí (a online služieb vo všeobecnosti) vzrastá aj množstvo nových slov, ktoré v konverzáciách vznikajú. Či už ide o slangové výrazy, rôzne skratky alebo neologizmy, spoločne ich zaraďujeme do neformálnej (a spravidla aj nespisovnej) komunikácie, ktorá je pre komunikáciu na sociálnej sieti charakteristická. V prípade sociálnej siete sa používanie takýchto výrazov ešte väčšmi umocňuje, nakoľko je každá správa používateľa limitovaná na 140 znakov (Twitter). Použitie existujúcich slovníkov alebo anotovaných zoznamov slov sa preto v mnohých prípadoch (výrazne) sťažuje.

Šum. Princíp, na ktorom sú založené súčasné sociálne siete, je veľmi podobný online diskusným fóram (chat), kde sa príspevky odosielajú v reálnom čase. Okrem iného môžeme považovať aj toto za jeden z faktorov, ktoré vplývajú na gramatickú korektnosť textu.

Poznámka: *Uvedená problematika bola diskutovaná v práci [1], ktorej sa detailnejšie venujeme v kapitole Detekcia sentimentu.*

4.1.1 Existujúce prístupy

Rozpoznávaniu korektných pomenovaných entít sa venovali autori v [33]. Zamerali sa pritom na nejednoznačnosť mien, ktorá sa vyskytuje najmä v prípade názvov spoločností. Príkladom môže byť spoločnosť Orange (v preklade tiež: orange = pomaranč), Apple (v preklade tiež: apple = jablko) alebo Blackberry (v preklade tiež: blackberry = černica). Ich prístup spočíval v zostavovaní tzv. základných profilov (*angl.* basic profiles) spoločností, ktoré ich charakterizovali a boli preto východiskom pri rozpoznávaní. Autori zostavili základné profily z kľúčových slov získaných prevažne z webových zdrojov (napr. internetové stránky jednotlivých spoločností, Wikipedia, a pod.). Následne ich s využitím n-gramov mapovali na tweety. Výhoda uvedeného prístupu bola najmä možnosť dynamického obohatenia základných profilov o nové kľúčové slová z neustáleho prúdu nových tweetov. Navrhovaná metóda sa dá aplikovať nielen na tweety, ale taktiež aj na príspevky a komentáre z iných sociálnych sietí (napr. Facebook).

Cremonesi a kol. v [9] sa zamerali na tweety z domény televízie a snažili sa v nich nájsť entity. Využili pritom **klasifikátor jednej triedy** (*angl.* one-class classifier), ktorý sa trénuje na objektoch patriacich do vybranej triedy. Klasifikátor sa následne snaží spomedzi všetkých testovacích dát vyberať iba tie, ktoré patria (s určitou pravdepodobnosťou) do tejto vybranej triedy. Na každú triedu je pritom potrebný samostatný klasifikátor, čím sa eliminuje nutnosť preklasifikovať všetky existujúce klasifikátory v prípade, že do množiny dát pribudne nová trieda. Autori si v práci za trénovacie tweety zvolili tie, ktoré poukazovali na určitú entitu (položku z katalógu - TV reláciu). Ich vlastná implementácia klasifikátora rozdelila celý proces na tri kroky, ktoré sa vykonávali až kým aspoň jeden z nich neuspel (t.j. tweet nebol klasifikovaný):

1. **Presná zhoda** - priame vyhľadávanie názvu filmu v tweete.
2. **Voľná zhoda** - hľadanie (pod)reťazcov z názvu filmu v tweete. Tento krok je odolný aj voči niektorým preklepom.
3. **Použitie SVM**[2] - použitý v prípade, že sa žiadna časť z názvu filmu nenachádza v tweete. Autori implementovali klasifikátor založený na SVM. Samotný tweet reprezentovali ako vektor atribútov, ktoré vznikli najmä tokenizáciou a predspracovaním textu (tweetu). Za základné atribúty si zvolili unigramy (samotné termy), bigramy (dvojice susedných termov), hashtagy a názov (stupeň zhody medzi obsahom tweetu a názvom filmu). Normalizáciu atribútov vykonali prostredníctvom IDF (z modelu TF-IDF).

Autori experimentom ukázali, že postupná aplikácia jednotlivých krokov zlepšovala

[2]Support vector machine

celkové výsledky. Za hlavnú nevýhodu ich prístupu však považujeme priamu úmernosť počtu potrebných klasifikátorov od počtu skúmaných filmov, čo však autori priamo v práci obhajujú kvalitnou robustnosťou navrhovanej metódy pri zväčšujúcom sa počte spracovávaných filmov.

Gattani a kol. riešili v [16] problém extrakcie, prepojenia, klasifikácie a značkovania obsahu zo sociálnej sieti. Navrhli rozsiahly systém, v ktorom sa spája viacero prístupov prispievajúcich k riešeniu tejto problematiky. Jadrom ich metódy bola báza znalostí - Wikipedia, ktorá poskytovala zoznam kandidátnych pomenovaných entít. Prínosom bolo zahrnutie kontextu z diskusie na sociálnej sieti. Kontext umožnil klasifikovať aj tie príspevky, v ktorých sa samotné pomenované entity priamo nevyskytovali, ale napríklad nadväzovali na predchádzajúcu aktivitu používateľov.

Do procesu mapovania pomenovaných entít vstúpili používatelia aj v práci ([19]) Hua a kol., v ktorej autori vychádzali zo vzájomnej sociálnej interakcie - a to najmä vzťahu nasledovateľ - nasledovaný (*angl.* follower - followee). Prichádzajú s myšlienkou váhovanej dosiahnuteľnosti (*angl.* weighted reachability), ktorá zohľadňuje vzdialenosť medzi dvoma používateľmi, a to vrátane počtu uzlov na ceste medzi nimi. Obohatením je uvažovanie aktuálnej popularity entity (t.j. pohľad na záujem o entitu naprieč nedávnymi príspevkami). Autori sa vo svojej práci tiež zamerali na problematiku nejednoznačnosti pomenovaní.

Vytvorením modelu používateľa sa pokúsili rozpoznať pomenované entity Shen a kol. v [30]. Zamerali sa na identifikáciou tém (resp. skupiny tém), o ktoré sa používateľ - autor tweetu - vo svojom profile najviac zaujíma. Použitím tohto princípu navyše otvorili cestu k riešeniu problému nejednoznačnosti pomenovaných entít.

Mnohé ďalšie práce sa venovali tejto problematike. Li a kol. navrhli v [23] systém TwiNER určený pre identifikáciu pomenovaných entít v tzv. tematických príspevkoch na Twitteri (*angl.* Targeted Twitter Stream). Tematickými príspevkami nazývame prúd príspevkov, ktoré sú spravidla filtrované podľa vopred zvolených kľúčových slov (napr. príspevky o určitej spoločnosti, udalosti alebo osobe). Ritter a kol. sa snažili vo svojej práci [28] vylepšiť existujúce nástroje pre prácu s prirodzeným jazykom, ktoré dosahovali nedostatočné výsledky pri analýze príspevkov zo sociálnych sietí. Ich príspevok možno nájsť najmä v skvalitnení identifikácie slovných druhov (*angl.* Part-of-speech tagging), slovných spojení, oprave chýbajúcich kapitálok a v neposlednom rade aj v rozpoznávaní pomenovaných entít.

4.2 Ko-učenie

Častým problémom, s ktorým sa pri rozpoznávaní pomenovaných entít musíme vysporiadať, je ich absencia. Štandardné webové dokumenty, akými sú články alebo príspevky z internetových encyklopédií, zvyčajne obsahujú dostatok textu, prípadne pomenované entity dokážeme rozpoznať priamo z ich názvov. V prípade sociálnych sietí, kde je dĺžka príspevku limitovaná, sa ale nezriedka stretávame s prípadom, kedy používatelia vyjadrujú svoj postoj k entite, ktorá sa však nevyskytuje priamo v texte, avšak je k danému textu relevantná. Relevancia sa prejavuje napríklad prítomnosťou súvisiaceho obrázka alebo externého prepojenia.

Mnohé prístupy využívajú v procese rozpoznávania pomenovaných entít strojové učenie. Pod strojovým učením, ako už názov napovedá, môžeme rozumieť činnosť, pri ktorej sa algoritmus snaží naučiť (natrénovať) postup, ako dospieť k riešeniu určitého problému. Problém rozpoznávania pomenovaných entít môžeme tiež nazvať klasifikáciou, teda úlohou priradenia triedy/kategórie (pomenovaná entita) k dokumentu. Pri klasifikácií pritom môžeme rozlišovať dva základné prístupy k učeniu:

- **učenie s učiteľom** (*angl.* supervised learning), pri ktorom algoritmus vychádza z označených trénovacích dát (množina dát, ktorej položky majú určenú klasifikáciu na základe zlatého štandardu),
- **učenie bez učiteľa** (*angl.* unsupervised learning), pri ktorom nie je množina trénovacích dát označená. Algoritmus sa v tomto prípade snaží z trénovacích dát vyvodiť vzory, ktoré využije pre samotnú klasifikáciu.

V predchádzajúcich kapitolách sme sa stretávali najmä s metódou SVM. Cremonesi vo svojej práci [9] využil SVM pre rozpoznanie názvov TV relácií v prípade, že sa pomenovaná entita v príspevku nenachádzala. Postup diskutovaný v práci využíval učenie s učiteľom, ktoré vychádzalo z tweetov klasifikovaných do jednotlivých tried (trieda predstavovala pomenovanú entitu).

V tomto prístupe sme však identifikovali dva hlavné problémy:

- *anotácia trénovacej množiny* môže byť zložitý proces, pri ktorom často nevieme odhadnúť, aká veľkosť trénovacej množiny je dostačujúca. Autori tento problém diskutovali aj vo svojej práci.
- *trénovacia množina*. Ak vytvoríme anotáciu pre dokumenty súvisiace s určitou množinou tried, môže nastať problém v prípade, že sa objaví nová trieda. Pre túto triedu nebudeme mať k dispozícii žiadne trénovacie dáta, čo môže vyústiť do zlyhania klasifikácie.

Perspektívnym riešením problému je **čiastočné učenie s učiteľom** (*angl.* semi-

supervised learning), ktorého základné princípy opisujú vo svojej publikácii Chapelle a kol. [6]. Na rozdiel od učenia bez učiteľa, čiastočné učenie s učiteľom využíva v procese učenia malú množinu trénovacích dát, ktoré sú označené a súčasne využíva aj neoznačené dáta (ktorých je majorita). Aplikáciu tohto prístupu pri kategorizácii textu zrealizovali Xu el. al v [32], ktorí sa snažili zamerať práve na problematiku nízkeho počtu anotovaných dát. Na základe vykonaných experimentov dospeli k záveru, že čiastočné učenie dokázalo znížiť chybovosť voči štandardnému SVM až o 26%. Dôležité je však poznamenať, že aj v tomto prípade potrebujeme určitú množinu označených dát.

Pri skúmaní ďalších možných riešení sme sa stretli s pojmom ko-učenie (*angl.* co-learning). Predpona „ko-“ naznačuje, že môžeme hovoriť o určitom druhu kooperácie. Ko-učenie sa v skutočnosti zakladá na spolupráci, a to spravidla dvoch klasifikátorov. Výsledky prvého klasifikátora sa využívajú ako vstup pre učenie sa druhého klasifikátora. Na to, aby nám tento prístup mohol pomôcť, potrebujeme vhodne navrhnúť oba klasifikátory. Príkladom môže byť nasledujúci scenár:

- prvý klasifikátor nebude vyžadovať na svoju činnosť označenú trénovaciu množinu,
  - týmto klasifikátorom sa budeme snažiť pokryť čo najväčšie množstvo prípadov, ktoré dokáže klasifikovať,
  - výstupom bude klasifikácia, o ktorej vieme s veľkou istotou prehlásiť, že je korektná (blízka zlatému štandardu),
- druhý klasifikátor bude vyžadovať pre svoju činnosť označenú trénovaciu množinu,
  - označenú trénovaciu množinu bude tvoriť výstup prvého klasifikátora.

Ko-učenie použili na problematiku rozpoznávania pomenovaných entít vo svojej práci [8] Chung a kol. Svoju metódu navrhli so zameraním na kórejčinu, kde je identifikácia pomenovaných entít zložitejší problém ako napríklad v anglických textoch.

Použitie ko-učenia nám teda pomôže vyriešiť problém nedostatočnej, prípadne chýbajúcej označenej trénovacej vzorky. Zároveň sa v ňom skrýva potenciál, pomocou ktorého vieme nájsť klasifikáciu pre prípady, kedy jeden klasifikátor nedokáže pokryť celú testovaciu vzorku dát.

4.3 Zhrnutie

Doména prepájania multimediálneho obsahu so sociálnymi sieťami ponúka viacero riešení, ktoré sa snažia využiť textové ale aj sociálnej vlastnosti príspevkov. Textové prístupy sa snažia v texte identifikovať významné črty, ktoré pomôžu v procese mapovania. Mnohé prístupy sa tiež opierajú o znalostné bázy, ktoré však najmä pri neanglických textoch trpia neaktuálnosťou. Základom viacerých prác je tiež strojové učenie, ktoré síce pomáha pri hľadaní korektných mapovaní, avšak často si vyžaduje náročný proces učenia.

Ako hlavný cieľ našej práce sme si stanovili navrhnúť metódu, ktorá vyrieši nedostatky existujúcich prístupov: potrebu robustnej znalostnej bázy, náročného procesu učenia a jazykovú závislosť prístupov. Vychádzajúc z výsledkov našej analýzy sme si ako základné črty príspevkov sociálnej siete deklarovali pomenované entity a hashtagy. Zároveň by sme chceli odhaliť význam ďalšieho prvku - externého prepojenia, ktorý sa v príspevkoch nachádza a ktorý môžeme použiť v procese mapovania. Našou základnou hypotézou bude, že tento prvok dokáže skvalitniť štandardné prístupy mapovania.

5. Úvod k extrakcii metadát

Vyhľadanie pomenovaných entít v príspevkoch sociálnej siete nám umožní získať základnú informáciu, ktorú v sebe príspevky skrývajú. Tú môžeme následne spracovať a extrahovať z obsahu príspevku ďalšie poznatky. Keďže ho už budeme vedieť priradiť ku konkrétnej entite (prípadne entitám), môžeme tak obohatiť existujúce bázy metadát o nové, viac dynamické metadáta.

V tejto kapitole uvádzame základný prehľad o rôznych možnostiach v doméne extrakcie metadát z textov. Hoci sme sa v nami navrhnutej metóde nezamerali na extrakciu metadát, ponúkame základný prehľad stavu tejto problematiky. Zamerali sme sa na nasledujúce oblasti:

- detekcia tém,
- detekcia sentimentu (resp. emócií),
- meranie sledovanosti.

5.1 Detekcia tém

V oblasti detekcie tém existuje viacero prístupov, ktoré spravidla pracujú s textom. Využívajú pritom metódy spracovania prirodzeného jazyka. V tejto podkapitole uvádzame existujúce prístupy, ktoré umožňujú extrakciu tém z dokumentov, ale taktiež sa pozrieme aj na vybrané prístupy, ktoré môžu dopomôcť k hľadaniu podobnosti medzi termami a dokumentmi, keďže tieto dva problémy spolu úzko súvisia.

Často používaným modelom je reprezentácia dokumentu pomocou TF-IDF, ktorému sme sa podrobnejšie venovali v predchádzajúcej kapitole. Ako uvádzajú autori v [3], jednou z nevýhod je najmä nízka redukcia množstva informácií z pôvodných dokumentov a taktiež nemožnosť odhaliť štatistickú štruktúru v rámci dokumentu a medzi jednotlivými dokumentami.

Rozšírením tohto modelu bolo Latentné sémantické indexovanie (*angl.* Latent semantic indexing, LSI), s ktorým prichádzajú Deerwester a kol. v [12]. Autori sa snažili pomocou štatistických techník v dokumentoch odhadnúť latentnú sémantickú štruktúru, ktorá môže byť čiastočne skrytá z dôvodu náhodného výberu slov pri skúmaní ich frekvencie. Zamerali sa na problematiku synoným a homoným, s ktorou zápasia mnohé metódy objavovania znalostí. Využili pritom techniku jednohodnotovej dekompozície (*angl.* singlar-value decomposition) a z asociačnej matice medzi termami a dokumentami zostrojili tzv. sémantický priestor (*angl.* semantic space). V

rámci tohto priestoru sa blízke termy a dokumenty umiestňovali vedľa seba. Jednohodnotovú dekompozíciu autori využili pre zvýraznenie asociačných vzorov v dátach. V porovnaní so štandardným modelom TD-IDF, LSI zohľadňuje aj tie termy, ktoré sa v dokumente síce nenachádzajú, ale s ním súvisia.

Napriek úspešnému uplatneniu modelu LSI, tento model trpel nedostatočným štatistickým základom, ktorý sa snažil zdokonaliť Hoffman vo svojej práci [18]. Prichádza s myšlienkou Pravdepodobnostného latentného sémantického indexovania (*angl.* Probabilistic latent semantic indexing, pLSI), ktorého hlavným cieľom bolo skvalitniť existujúci model LSI a obohatiť ho o pravdepodobnostný princíp (*angl.* likelihood principle). Výsledkom je model (nazývaný aj aspektový model), ktorý umožňuje aplikovať štandardné technicky z oblasti štatistiky (napr. model fitting, model combination alebo complexity control). pLSI tiež umožňuje riešiť problém používania synoným a viacvýznamových slov (načrtnutý v [12]), ktorý sa vyskytuje nielen v dokumentoch, ale aj vo vyhľadávacích dopytoch zvedavca. Ku každému dokumentu môžeme s určitou pravdepodobnosťou pridružiť témy a k jednotlivým témam (opäť s určitou pravdepodobnosťou) slová, ktoré ich reprezentujú. Problémom tohto prístupu bolo riziko preučenia sa (dokument bol reprezentovaný ako zoznam čísiel a pri vzrastajúcom počte dokumentov lineráne rástol aj zoznam atribútov) a nemožnosti priradiť pravdepodobnosť dokumentu, ktorý sa nenachádzal v trénovacej množine.

Model latentnej Dirichletovej alokácie (*angl.* Latent Dirichlet allocation, LDA) predstavený v [3] sa snažil uvedené prístupy skvalitniť. Zakladá sa na algoritme Expectation Maximization algorithm (EM) a na hierarchickom Baysovskom modeli, v ktorom sa každý dokument modeluje ako konečná zmes (*angl.* mixture) prislúchajúcich tém. Každá téma je pritom modelovaná ako nekonečná zmes základných (pôvodných) tém, ktoré do zmesi vstupujú s určitou pravdepodobnosťou. Pravdepodobnosť tém sa tak stáva explicitnou reprezentáciou pre textový dokument.

Aplikáciu strojového učenia na kategorizáciu textov využil vo svojej práci ([20]) Thorsten Joachims. Prostredníctvom algoritmu podporných vektorov (*angl.* Support Vector Machines) sa snažil vyriešiť časté problémy v oblasti textových klasifikátorov - veľký počet atribútov (a s tým súvisiace preučenie), nízky počet irelevantných atribútov, riedkosť vektora dokumentov a lineárnu separovateľnosť týchto problémov. Súčasťou jeho výskumu bolo aj porovnanie algoritmu SVM s algoritmami Naivne Bayes, Rochio, k-najbližších susedov a rozhodovacím stromom. Vykonané experimenty nad textami z rôznych zdrojov preukázali priaznivé výsledky algoritmu SVM, pričom zároveň dosiahol aj najlepšie výsledky spomedzi ostatných skúmaných algoritmov.

Problematike dolovania tém a názorov z tweetov sa venovali Meng el at. v [25], ktorí vo svojej práci stanovili ako základný identifikátor témy **hashtag**. Využívali

pritom príspevky zo sociálnej siete Twitter. Na vzorke 200 000 tweetov identifikovali v približne 23% z nich prítomnosť aspoň jedného hashtagu. Spomedzi všetkých unikátnych hashtagov (100) získali až 75% takých, ktoré boli užitočné pre potreby extrakcie témy. Vyvinuli pravidlový klasifikátor, ktorý umožňoval identifikovať hashtagy kategórií, entít a udalostí. Využili pri tom rôzne existujúce slovníky. Samotná extrakcia prebiehala nad neorientovaným grafom, ktorého každým uzlom bol hashtag. Pomocou algoritmu priaznivej propagácie (*angl.* affinity propagation) následne zostavili klastre hashtagov spolu s centroidmi. Práve centroidy využili ako pomenovania pre jednotlivé témy. Hashtagy roztriedili do 7 základných kategórií: štandardné kategórie (napr. šport, ekonómia), ľudia a miesta, udalosti, obsahové (hashtagy, ktoré naznačujú obsah tweetu, napr. *#terazsahrá*), sentiment, zdrojové (také, ktoré poukazujú na zdroj tweetu) a diskusné.

5.2 Detekcia sentimentu

Sociálne siete sa stali nielen priestorom na zdieľanie informácií, ale taktiež aj na vyjadrovanie postoja ľudí. Postoj sa pritom navzájom dopĺňa so sentimentovou zložkou správania používateľa, ktorým zároveň autori príspevkov spravidla vyjadrujú svoje pozitívne alebo negatívne emócie na diskutovanú problematiku. Analýzu sentimentu môžeme charakterizovať ako úlohu pozostávajúcu z dvoch základných krokov:

1. Identifikácia sentimentu v texte.
2. Odhalenie polarity identifikovaného sentimentu.

V mnohých prácach sa stretávame s rôznymi prístupmi k detekcii a extrakcii sentimentu z publikovaného textu. V kontexte emócií rozlišujeme tri základné polarity: pozitívnu, negatívnu a neutrálnu. Plutchik prichádza v [27] s myšlienkou tzv. **kolesa emócií** (*angl.* wheel of emotions), ktoré pozostáva zo štyroch párov základných protichodných emócií:

- radosť - smútok (*angl.* joy - sadness),
- hnev - strach (*angl.* anger - fear),
- dôvera - znechutenie (*angl.* trust - disgust),
- očakávanie - prekvapenie (*angl.* anticipation - surprise).

Jednotlivé emócie sú (ako aj z názvu konceptu vyplýva) usporiadané v tvare kolesa, pričom ich autor navyše farebne rozlišuje. Navzájom protichodné emócie sa pritom nachádzajú oproti sebe a vyfarbujeme ich komplementárnymi farbami. Naopak, súvisiace emócie sa nachádzajú v kolese blízko seba. Plutchikovu myšlienku využili Mohammad a kol. v [26] pre zostavenie lexikónu anotovaných slov. Tieto slová boli

pritom anotované davom, a to práve do ôsmych Plutchikových základných emócií (v závislosti od emócie, ktorú vyvolávajú).

Bravo-Marquez a kol. sa v [5] zamerali na kombináciu existujúcich metód s prihliadnutím na efektívnosť. Vychádzali pritom z troch základných kategorizácií metód:

Polarita. Metódy slúžiace na kategorizáciu entity podľa polaritnej orientácie na pozitívnu, negatívnu a neutrálnu.

Emócia. Metódy extrahujúce emóciu alebo náladu z vybranej časti textu. Tieto metódy pritom spravidla pracujú s emocionálne-orientovanými lexikónmi, ktoré poskytujú zoznamy slov a k nim korešpondujúce emocionálne stavy.

Sila. Metódy rozhodujúce o miere (sile) zachytenej polarity alebo emócie, prípadne o miere subjektivity skúmaného textu.

Cieľom práce bolo najmä skvalitniť klasifikáciu subjektivity a polarity. Analyzovali sentiment dostupný z existujúcich tweetov. Keďže išlo o relatívne krátke texty, autori využili klasifikáciu na úrovni viet. Rozlišovali pritom subjektivitu (klasifikácia na subjektívne a nesubjektívne/objektívne tweety) a polaritu (klasifikácia na pozitívne a negatívne tweety, pričom obe kategórie spadajú do podkategórie subjektívnych tweetov). Kombináciou a overením uvedených aspektov sa autorom podarilo odhaliť, že polarita, emócia a sila sú navzájom komplementárne. Pri každom z prístupov dochádza k skúmaniu rovnakého problému, ale vždy z iného uhla (dimenzie) pohľadu.

Autori tiež poukázali na dôležitý poznatok vyplývajúci z trénovacích množín dát, ktoré používali. Výsledky ich experimentu sa totiž výrazne líšili v závislosti od dátovej množiny, a to najmä v dôsledku toho, že dátové množiny vznikli manuálnou anotáciou sentimentu. Manuálna anotácia sa pritom spolieha na anotátora, a ako sme už v úvode našej práce poznamenali, často je sprevádzaná vysokou mierou subjektivity. Ako môžeme vidieť, výsledkom môžu byť aj nekonzistentné výsledky.

V predchádzajúcej kapitole sme analyzovali príspevok ([25]) Menga a kol., ktorý prichádza s myšlienkou sumarizácie názorov z tweetov. Za názory v danom kontexte považujeme najmä pozitívny alebo negatívny vzťah autora tweetu k určitej entite. Autori prichádzajú s pojmom *insightful opinionated tweet* (voľným prekladom *tweet s pohľadom a názorom*), ktorým nazývajú taký tweet, ktorý okrem názoru jeho autora obsahuje aj pohľad. Príklad:

„Myslím si, že je načase povedať, že si Barack Obama zaslúži pochvalu za jeho postoj k akcii na #Libya, a to napriek rozdielnemu názoru domácich občanov.“

Autor príspevku nevyjadruje len svoj pozitívny postoj k Barackovi Obamovi, ale taktiež explicitne tvrdí, že tento postoj vznikol, ako dôsledok jeho činov v Líbii.

Autori práce klasifikovali tweety na 7 rôznych kategórií podľa pohľadu, pričom pri každej identifikovali vzory, ktorými bolo možné tweet kategorizovať. Pre detekciu sentimentu v tweete využili lexikálny slovník obsahujúci slová anotované na základe ich polarity. Vo svojom klasifikátore vylúčili zo skúmaných slov tie, ktoré pochádzali priamo z pomenovanej entity, keďže samé môžu obsahovať sentiment. Autori tiež skúmali fenomén negácie (*angl.* negation phenomenon), teda slová alebo výrazy, ktoré negujú polaritu vety (napr. „nikdy“, „eliminovať“, „znížiť“). Cieľom ich práce bola klasifikácia tweetov s pohľadom a s názorom využitá pri zostavení celkového rámca, ktorý na vytvorenie sumarizácie názorov využil tri základné prvky: tému (analyzované v predchádzajúcej podkapitole), pohľad a sentiment. Výstupom bola metóda, ktorá dokázala klasifikovať tweety na témy, ku ktorým bol navyše priradený názor, pohľad a sentiment samotných autorov tweetov.

Problémom vyššie diskutovaných prístupov môže byť najmä závislosť od lexikónu slov. S tým úzko súvisí aj ďalší problém, a to prenositeľnosť do iného jazykového prostredia. Problematike sa v [1] venovali Aisopos a kol., ktorí prichádzajú s univerzálnym riešením nezávislým od voľby jazyka spracovávaného textu. Vo svojej práci sa tiež zamerali na riešenie častých preklepov a iných chýb v textoch, ktoré sa vo vysokej miere vyskytujú v online prostredí, a to najmä v prostredí sociálnych sietí. V samotnom návrhu metódy reprezentovali tweety ako grafy n-gramov. Každý graf bol pritom charakteristický triedou sentimentu, do ktorej patril. Autori definovali tri základné triedy, do ktorých rozdelili celú trénovaciu množinu tweetov: pozitívnu, negatívnu a neutrálnu. Následne sa snažili objaviť vzory charakteristické pre každú triedu, prostredníctvom čoho by vedeli klasifikovať testovaciu množinu do tried. Experimentálne vo svojej práci potvrdili vhodnosť modelu grafu n-gramov (pre n-gramy veľkosti $n = 3$ a $n = 4$) ako referenčného modelu pre sentimentálnu analýzu v prostredí sociálnych sietí - a to nielen z hľadiska efektívnosti, ale aj z hľadiska vyššej miery presnosti v porovnaní s inými modelmi.

Ďalší spôsob detekcie sentimentu skúmali Davidov a kol. v [11], kde sa snažili dáta z Twittera klasifikovať podľa sentimentu. Využili pri tom dátovú množinu obsahujúcu 50 rôznych hashtagov (charakterizujúcich sentiment) a 15 znakov (emotikonov) priamo vyjadrujúcich emóciu (napr. :), :(alebo :D). Pri klasifikácii sa zamerali na tie tweety, ktoré neboli predtým žiadnym spôsobom anotované. Ich metóda sa zakladala na 4 základných atribútoch: (1) detekcia interpunkcie, (2) jednotlivé slova, (3) n-gramy slov a (4) vzory. Pri *vzoroch* klasifikovali slová na tie, ktorá sa v textoch vyskytovali častejšie (t.j. slová s vysokou frekvenciou) a tie, ktoré sa vyskytovali zriedkavejšie - tzv. obsahové slová (*angl.* content words). Okrem toho

sa zamerali na problém mnohonásobného a prekrývajúceho sa sentimentu v rámci jedného tweetu (napr. použité značky sa z pohľadu sentimentu prekrývali alebo - naopak - si úplne odporovali). Autori článku experimentálne overili svoje hypotézy, pričom výsledky svojej metódy porovnávali s manuálnou anotáciou (ktorú vykonali participanti). Spomedzi skúmaných atribútov dosahoval najlepšie výsledky (f-skóre) atribút (4) vzory. Klasifikácia založená na emotikonoch dosiahla v porovnaní so značkami signifikantné zlepšenie. Hoci boli aj výsledky overenia prostredníctvom manuálnej anotácie participantami priaznivé, obávame sa, že ich počet bol príliš nízky (4 participanti) na to, aby sme ich považovali za referenčné.

5.3 Meranie sledovanosti

Spoločnosti produkujúce audiovizuálny obsah sa zaujímajú o jeho úspešnosť a popularitu medzi koncovými divákmi, a to z rôznych dôvodov (reklama, úspech, budovanie značky, a pod.). V súčasnosti môžeme za najpoužívanejšiu metriku pokladať sledovanosť televízneho programu. Jedným z celosvetovo najpoužívanejších systémov je systém Nielsenovho merania sledovanosti[1], ktorý sa pre účely merania sledovanosti televízie využíva už od roku 1950. Tento systém sa pritom zakladá na pozorovaní správania diváka - či už manuálnym zapisovaním alebo pomocou špeciálneho zariadenia pripojeného k televíznemu prijímaču (známym tiež pod pojmom *píplmeter*), ktoré zbiera tieto údaje elektronicky.

Na Slovensku bol klasický systém založený na meraní prostredníctvom zistení agentúr (napr. formou dotazníkov a ankiet) v roku 2002 nahradený elektronickým meraním (pomocou píplmetra). V súčasnosti realizuje komerčne toto meranie spoločnosť PTM[2]. Hoci sa uvedené analýzy sledovanosti snažia demograficky pokryť čo najväčšie percento obyvateľov, reálny počet účastníkov je stále relatívne nízky (k júlu 2014 - 1200 domácností s približne 3500 osobami [10]). Sociálne siete pritom môžu predstavovať skrytý potenciál, ako obohatiť a skvalitniť existujúce prístupy.

Wakamiya a kol. navrhli v [31] metódu merania sledovanosti na základe aktivity na sociálnej sieti Twitter. Okrem detekcie samotných televíznych programov v tweete sa zaoberali aj problematikou detekcie miesta odoslania tweetu, keďže sa výskum odohrával aj nad rôznymi regionálnymi televíznymi stanicami v Japonsku. Výstupom ich prístupu bola preto kombinácia troch rôznych relevancií:

- **textovej** - pre nájdenie prepojenia medzi televíznym programom (v zmysle zoznamu relácií a času ich vysielania) a tweetom,
- **priestorovej** - nájdete konkrétnej televíznej stanice, ktorá daný program vy-

[1] http://www.nielsen.com
[2] http://www.pmt.sk

siela (jedna televízna relácia sa vysiela opakovane na viacerých staniciach),

- **časovej** - autori vychádzali z predpokladu, že tweety sú odosielané približne v podobnom čase, ako sa daný program vysiela.

Sociálnou sieťou Facebook sa v tejto oblasti zaoberali v [7] Cheng a kol. Navrhli model, ktorý sa zakladal na dostupných informáciách z fanúšikovských stránok televíznych relácií, medzi ktoré patrili: príspevky, komentáre, označenia „Páči sa mi to“ a zdieľania. Tieto informácie ďalej využili na účely predpovede sledovanosti. Predpoveď sa realizovala pomocou neurónovej siete natrénovanej spätnou propagáciou (*angl.* back-propagation). Svoje hypotézy overili experimentom, pri ktorom porovnali predpovede získané z neurónovej siete a výsledky sledovanosti pochádzajúce z Nielsenovho merania sledovanosti. Vykonané experimenty preukázali, že aktivita na sociálnej sieti signifikantne koreluje s mierou sledovanosti televízneho programu, a môže preto predstavovať ďalšiu zaujímavú informáciu, ktorú vieme pre potreby rozšírenia metadát získať.

6. Prepájanie sociálneho a televízneho obsahu

V predchádzajúcich kapitolách sme analyzovali existujúce prístupy, ktoré sa snažili o automatické prepojenie sociálneho a televízneho obsahu. Túto úlohu môžeme tiež charakterizovať ako potrebu mapovania príspevkov zo sociálnych sietí na inštancie, ktoré môžeme nájsť v znalostných bázach televízneho obsahu. Mnohé riešenia používajú ako základnú znalostnú bázu Wikipediu alebo DBpediu. Ako sme diskutovali už v úvode našej práce, v prípade, že sa v problémovej doméne vyskytne nová položka, často musí manuálne zasiahnuť dav, ktorý túto položku do znalostnej bázy pridá. Problém sa navyše prehlbuje, ak uvažujeme špecifickú doménu, akou je napríklad oblasť televízneho vysielania.

Televízny program obsahuje informácie o jednotlivých televíznych reláciách - od základných (napr. názov relácie, anotácia, zoznam hercov) až po podrobnejšie, menej zriedkavejšie (napr. hodnotenie, recenzie). Tieto informácie však môžeme charakterizovať aj z hľadiska času - pre každú televíznu reláciu vieme prezradiť, kedy sa vysiela. Zároveň, v prípade, že sa objaví nová položka, vieme ju pomocou televízneho programu včas identifikovať [1] ako inštanciu v báze dát. Na druhej strane, položky, ktoré sa po čase z televízneho programu vytratia, už nemusia byť pre zvedavca až natoľko atraktívne.

Ako sme odhalili pri analýze sociálneho obsahu, obdobné správanie môžeme nájsť aj na sociálnych sieťach, kde používatelia diskutujú televízny obsah. Platí pritom, že najčastejšie diskutujú práve tie televízne relácie, ktoré sa aktuálne vysielajú, prípadne sa čoskoro plánujú vysielať. Zároveň sa obsah na sociálnej sieti vyznačuje vysokou mierou dynamiky.

Uvedené dôvody nás priviedli k rozhodnutiu zamerať sa práve na problematiku prepojenia obsahu medzi **televíznym programom** (v úlohe referenčnej bázy dát) a **sociálnou sieťou** (ako zdrojom pre extrakciu dynamických metadát). Pri návrhu nášho riešenia budeme vychádzať z cieľov, ktoré sme si stanovili:

1. Vytvoriť mapovanie medzi existujúcou bázou dát (televíznym programom) a zdrojom, z ktorého môžeme extrahovať dynamické dáta (sociálnou sieťou).
2. Extrahovať dynamické metadáta (zo sociálnej siete) s využitím vytvoreného mapovania.

V súčasnej fáze sme sa zamerali najmä na procesy spojené s **vytvorením mapovania**. Naše riešenie vychádza zo slovenských príspevkov, avšak našou snahou bolo ho navrhnúť tak, aby bolo jazykovo-nezávislé.

[1] Za predpokladu, že máme k dispozícii televízny program na dostatočný počet dní vopred.

6.1 Úvod k mapovaniu

Úlohu mapovania môžeme charakterizovať ako potrebu identifikácie toho, na ktorú televíznu reláciu (prípadne relácie) sa vzťahuje obsah konkrétneho príspevku zo sociálnej siete. Tento vzťah môžeme opísať aj ako stav, kedy **príspevok zo sociálnej siete diskutuje obsah, postavy alebo obsadenie z určitej televíznej relácie**.

Nech máme príspevok P_i z množiny všetkých relevantných príspevkov sociálnej siete (P_i patrí do P, kde P je množina príspevkov sociálnej siete). Potom pod úlohou **mapovania príspevkov sociálnej siete na televízne relácie** budeme rozumieť proces nájdenia množiny televíznych relácií R, pre ktoré platí, že je medzi nimi vzťah súvislosti s príspevkom sociálnej siete. Potrebujeme teda poznať odpovede na nasledujúce otázky:

1. Patrí príspevok P do televíznej oblasti?
2. Akú množinu televíznych relácií R môžeme identifikovať v príspevku P?
3. Do akej miery platí vzťah medzi príspevkom P_i a reláciou $R_j \in R$?

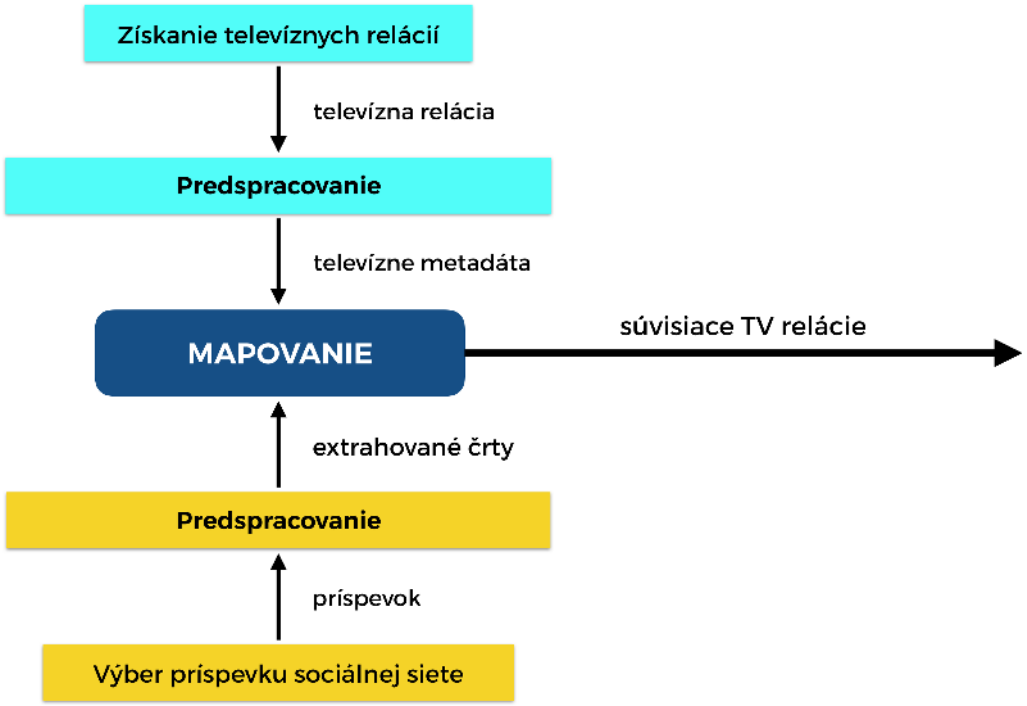
Proces predspracovania rozdeľujeme na dve fázy, v ktorých pracujeme na jednej strane s televíznymi reláciami a na druhej strane s príspevkom sociálnej siete (pre ktorý hľadáme súvisiace televízne relácie). V oboch fázach sa snažíme vytvoriť také množiny dát, ktoré budeme môcť **v procese mapovania porovnať a nájsť medzi nimi vzťah**.

Je dôležité poznamenať, že príspevok sociálnej siete je dynamickým prvkom, a teda jeho predspracovanie nastane až v okamihu, keď ho máme k dispozícii. Opačným prípadom sú televízne relácie, ktoré máme k dispozícii nezávisle od sociálnej siete, a preto ich predspracovanie môžeme zrealizovať vopred. Táto skutočnosť môže mať pozitívny vplyv na časovú efektívnosť mapovania.

Nasledujúcim krokom je realizácia mapovania, ktorá pracuje s výstupmi predspracovania. Výsledkom mapovania je **ohodnotený** a usporiadaný **zoznam** relácií - **kandidátnych televíznych relácií**, ktoré potenciálne súvisia s príspevkom sociálne siete. **Hodnota** predstavuje skóre, ktoré relácia dosiahla v procese hodnotenia. Z toho zoznamu následne získame požadovaný zoznam najsúvisiacejších relácií (t.j. N relácií s najvyšším skóre súvislosti).

V nasledujúcich podkapitolách sa bližšie pozrieme na všetky procesy, ktoré súvisia s predspracovaním, mapovaním a napokon získaním zoznamu kandidátnych televíznych relácií.

Uvedený postup ilustrujeme na obrázku č. 6.1.



Obr. 6.1: Všeobecný pohľad na proces mapovania

6.2 Vytvorenie televíznych metadát

Pri práci s televíznym programom vychádzame z dvoch kľúčových informácií:

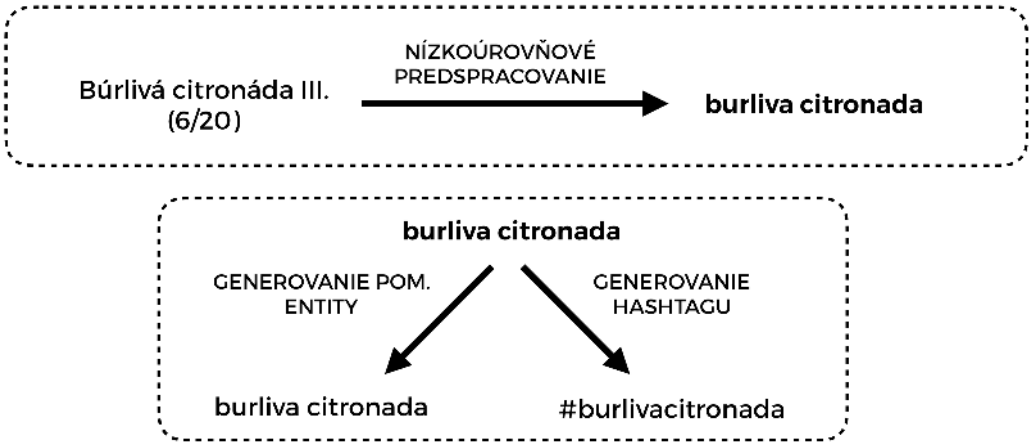
- názov televíznej stanice,
- názov televíznej relácie.

Názov televíznej stanice pritom vystupuje v roli kategórie, do ktorej televízna relácia patrí a ďalej s ním už nepracujeme. **Názov televíznej relácie** podlieha dvom typom predspracovania:

- *nízkoúrovňové predspracovanie* - pracujeme s názvom relácie ako s textom pre účely vygenerovania slovného **termu** (snažíme sa teda nájsť jednotný tvar, ktorý by bol vhodný pre ďalšie spracovanie),
- *vysokoúrovňové predspracovanie* - na term vygenerovaný z názvu relácie aplikujeme rôzne pravidlá, pomocou ktorých generujeme **televízne metadáta**: kandidátne pomenované entity a kandidátne hashtagy.

Pri vytváraní termu (**nízkoúrovňové predspracovanie**) aplikujeme štandardné techniky z prostredia spracovania prirodzeného jazyka:

1. Všetky veľké písmená prevedieme na malé.



Obr. 6.2: Pohľad na nízkourovňové a vysokoúrovňové predspracovanie

2. Z názvu relácie odstránime diakritiku a všetky znaky okrem písmen, číslic a medzier.
3. Z názvu relácie odstránime informácie o sérii (napr. informácie o čísle vysielanej časti a čísle aktuálnej série)[2].

Keďže názvy televíznych relácií v programe nie sú súčasťou viet, ale chápeme ich ako samostatné pomenované entity, môžeme vynechať proces lematizácie a hľadania koreňa slova (*angl.* stemming). Hoci tým nezískame základný tvar slov, pre účely mapovania nám tento stav plne postačuje.

Pri **vysokoúrovňovom predspracovaní** aplikujeme viaceré nami navrhnuté pravidlá, pomocou ktorých zostavujeme **bázu televíznych metadát**, ktorá pozostáva z (1) *kandidátnych pomenovaných entít*, (2) *kandidátnych hashtagov* a (3) *prefixového indexu*. Ako sme už uviedli v predchádzajúcej kapitole, výhodou televíznych metadát je ich statickosť - v samotnom procese mapovania sú k dispozícii a nie je potrebné ich pravidelne generovať. To má pozitívny vplyv najmä z hľadiska výpočtovej náročnosti.

Báza televíznych metadát je zároveň celkovým výstupom fázy predspracovania televízneho programu. Pre každú televíznu stanicu pritom generujeme samostatnú bázu televíznych metadát. Na obrázku č. 6.2 sa nachádza ukážka generovania pre televíznu reláciu *Búrlivá citronáda III. (6/20)*. Uvedený názov reflektuje šiestu časť tretej série seriálu.

Poznámka: Na obrázku č. 6.2 sa nenachádza tretia zložka bázy televíznych metadát - prefixový index.

[2]Úplné detaily implementácie uvádzame v kapitole *Implementácia*, časť *Extrahovač metadát o seriáli*.

Obr. 6.3: Príklad invertovaného indexu pomenovaných entít

V nasledujúcich podkapitolách sa postupne pozrieme na to, ako získame jednotlivé televízne metadáta a ako ich využijeme v procese mapovania.

6.2.1 Kandidátne pomenované entity

Pri získavaní kandidátnych pomenovaných entít z TV programu vychádzame z jednoduchej myšlienky - za pomenovanú entitu budeme považovať **názov televíznej relácie**. Keďže tento proces nasleduje až po procese nízkoúrovňového spracovania, v samotnej báze televíznych metadát budeme pomenovanú entitu reprezentovať malými písmenami, bez diakritiky a bez informácií o sérii.

So vzrastajúcou veľkosťou bázy televíznych metadát vznikajú problémy pri jej prehľadávaní. Práve z tohto dôvodu sme zaviedli invertovaný index, ktorého kľúč je prvé písmeno pomenovanej entity a hodnota je zoznam všetkých pomenovaných entít s daným prvým písmenom. Ku každej pomenovanej entite v indexe je zároveň priradená TV relácie, z ktorej vznikla.

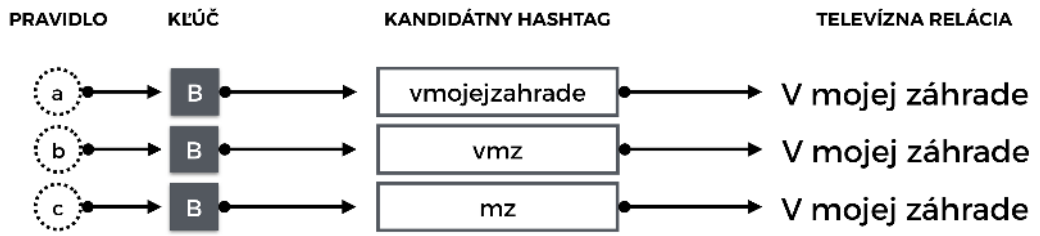
Invertovaný index, ktorý vznikne z pomenovaných entít (ilustruje obrázok 6.3), budeme nazývať index pomenovaných entít.

6.2.2 Kandidátne hashtagy

Ako sme už diskutovali v predchádzajúcich kapitolách, v sociálnom obsahu sa často stretávame s hashtagmi. Tie sú spravidla súčasťou príspevkov a nevznikajú náhodne, ale systematicky. Používatelia sociálnych sietí sa pri ich tvorbe totiž riadia určitými pravidlami, ktoré priamo vychádzajú z ich vlastností. Táto skutočnosť nám preto umožňuje navrhnúť algoritmus generovania hashtagov, ktorý bude pracovať s názvami televíznych relácií. Z názvov televíznych relácií tak zostavíme množinu kandidátnych hashtagov, ktorú využijeme v samotnom procese mapovania.

Pri generovaní kandidátnych hashtagov vytvárame 3 množiny hashtagov, ktoré vznikajú na základe nasledujúcich pravidiel:

(a) použijeme celý názov televíznej relácie, z ktorého odstránime medzery (napr.



Obr. 6.4: Príklad invertovaného indexu hashtagov

„V siedmom nebi“ → vsiedmomnebi),

(b) hashtag zostavíme zo začiatočných písmen jednotlivých slov slovného spojenia (napr. „V siedmom nebi“ → vsn),

(c) hashtag zostavíme ako začiatočné písmená jednotlivých slov slovného spojenia, ale iba tých, ktoré obsahujú aspoň 2 znaky (napr. „V siedmom nebi“ → sn).

Poznámka: *Za hashtag považujeme slovo s minimálnou dĺžkou 2 znaky. Z množiny kandidátnych hashtagov preto vždy odstránime všetky slová, ktorých dĺžka je 1.*

V našich prvotných úvahách sme uvažovali aj 4. pravidlo: za hashtag považovať každé slovo z názvu televíznej relácie (napr. „V siedmom nebi“ → „v“, „siedmom“ a „nebi“). Priebežné overovanie metódy ale odhalilo, že tento prístup mal za následok zvýšenie počtu nekorektných výsledkov mapovania. Práve z tohto dôvodu sme toto pravidlo napokon vylúčili z našej metódy.

Podobne, ako kandidátne pomenované entity, aj hashtagy umiestňujeme do invertovaného indexu, ktorého kľúč je prvé písmeno hashtagu a hodnota je zoznam všetkých hashtagov s daným prvým písmenom. Ku každému hashtagu v indexe je zároveň priradená TV relácia, z ktorej vznikol, a pravidlo (z pravidiel a-c), ktoré sme použili. Invertovaný index, ktorý vznikne z hashtagov (ilustruje obrázok 6.4), budeme nazývať **index hashtagov**.

6.2.3 Prefixový index

Identifikácia pomenovaných entít v texte je netriviálna úloha, a to aj v prípade, že vieme, aké entity hľadáme. Pri mapovaní na sociálny obsah budeme pracovať s pomenovanými entitami - názvami televíznych relácií. Budeme pritom vychádzať z hypotézy, že názvy relácií sa v texte budú vyskytovať ako slová (alebo slovné spojenia) začínajúce veľkým písmenom. Ak však uvažujeme spisovný text, každá veta začína veľkým písmenom, čo môže zvýšiť nepresnosť. Zároveň sa v texte môžu objaviť aj iné pomenované entity, ktoré nie sú názvami TV relácií. Pre odhalenie nekorektných pomenovaných entít môžeme využiť ďalšie televízne metadáta - **prefixový index**.

Prefixový index televíznych relácií pozostáva z prefixových slov, ktoré definujeme ako **prvé 4 znaky z názvu pomenovanej entity**[3]. Napríklad z názvu televíznej relácie *„Teória veľkého tresku“* získame prefixové slovo *„teor“*.

Pri tvorbe prefixového indexu vynechávame z pomenovanej entity nasledujúce druhy slov:

- názov spracovávanej televízie (napr. CBS),
- vybranú skupinu stopslov (napr. anglické členy „the“, „a“ a „an“).

Dôvodom tejto voľby je, že v názvoch TV relácií sa vyskytujú pomerne často (a to spravidla hneď na začiatku názvov). Ak by sme ich uvažovali pre účely prefixového indexu, znížili by sme jeho spoľahlivosť pri odlišovaní korektných a nekorektných pomenovaných entít.

Tabuľka 6.1: Príklad prefixových slov

Pomenovaná entita	**Prefixové slovo**
The Small Bang Theory	smal
CSS News *(vysielané v televízii CSS)*	news
Búrlivá citronáda	burl
Top novinky	topn

Tabuľka č. 6.1 obsahuje príklady prefixových slov, ktoré vzniknú aplikovaním nášho postupu pre ich generovanie. Ako môžeme vidieť, prefixové slová neobsahujú diakritiku ani veľké písmená (hoci pomenovaná entita ich obsahuje). Tvorbe prefixového slova (ako vysokoúrovňového predspracovania) predchádza nízkoúrovňové predspracovanie, ktoré práve tieto vlastnosti prefixových slov spôsobuje.

Prefixové slová spoločne tvoria prefixový index, ktorý využijeme pri hľadaní pomenovaných entít v príspevku zo sociálnej siete. Problematike sa bližšie venujeme v podkapitole *Extrakcia pomenovaných entít.*

6.3 Predspracovanie príspevku

Po vytvorení televíznych metadát dochádza k samotnému procesu mapovania. Prvým krokom je predspracovanie príspevku zo sociálnej siete, čím z neho získavame črty, ktoré môžeme porovnávať s televíznymi metadátami. Na rozdiel od predspracovania televízneho obsahu, ktoré sa odohráva vopred a nad celým TV programom, predspracovanie príspevku prebieha pre každý príspevok zvlášť, a to až v momente, keď požiadame o výsledky mapovania.

[3]Za znaky považujeme písmená, číslice a medzery. Hodnotu 4 sme zvolili experimentálne.

Predspracovanie príspevku môžeme rozdeliť na tri paralelné procesy:

- extrakcia pomenovaných entít (z textu príspevku),
- extrakcia hashtagov (z textu príspevku),
- extrakcia hashtagov z prepojení.

6.3.1 Extrakcia pomenovaných entít

Analýzou správania televízii (ktorú sme diskutovali v kapitole *Analýza obsahu zo sociálnych sietí*) sme odhalili, že tvorcovia sociálneho obsahu často využívajú pomenované entity na kategorizáciu príspevkov. Takáto kategorizácia umožňuje používateľovi sociálnej siete nájsť prirodzenú súvislosť medzi príspevkom a televíznou reláciou. V kontexte mapovania sa preto snažíme extrahovať názvy televíznych relácií a zostaviť tak **extrahované pomenované entity**.

V tomto procese využívame slovné n-gramy, ktoré získavame z textu príspevkov (na sociálnej sieti). Problematike n-gramov sme sa bližšie venovali v kapitole *Reprezentácia a predspracovanie textu*.

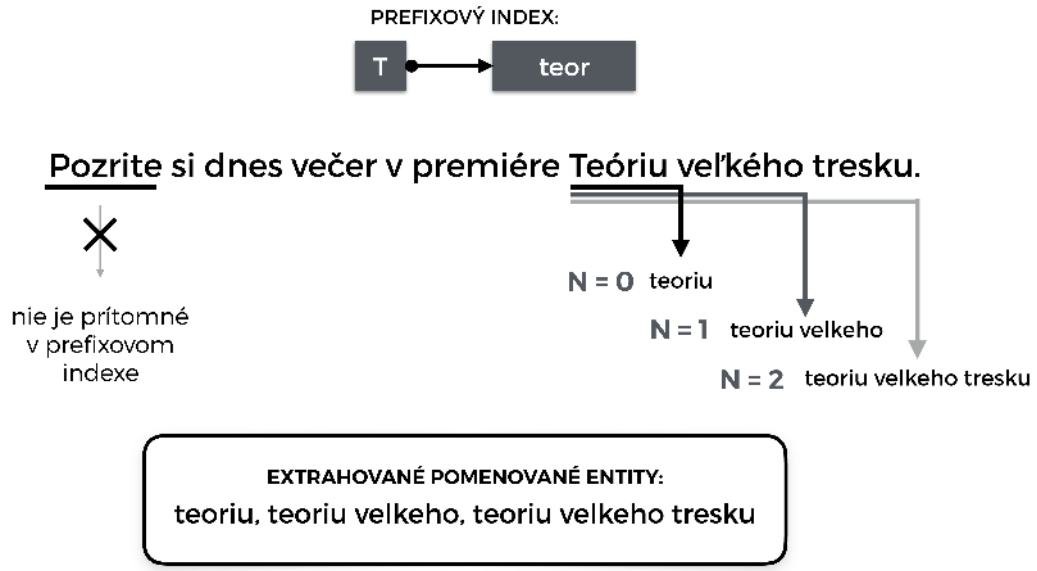
Pri extrakcii pomenovaných entít postupujeme nasledovne:

1. Z textu príspevku odstránime diakritiku.
2. Text príspevku rozdelíme na vety a jednotlivé vety na súvetia[4].
3. Postupne prechádzame súvetie po slovách a snažíme sa detegovať tie, ktoré začínajú veľkým písmenom[5]. Každé takéto slovo vyhľadáme v prefixovom indexe. Ak sa v prefixovom indexe nachádza, pridáme ho s spolu s ďalšími N nasledujúcimi slovami do množiny kandidátnych entít.
4. Uvedeným postupom získame množinu slovných $(1+N)$-gramov kandidátnych entít.

V kroku č. 3 preskakujeme tých kandidátov, ktorí v názve obsahujú názov aktuálne spracovávanej TV stanice. Hodnota N je číslo z množiny $N = \{0, 1, 2, 3\}$ - vždy teda vytvárame aj slovné bigramy, trigramy a štyrigramy. Pri návrhu tohto prístupu sme vychádzali z príspevku [1], v ktorom autori uvádzajú ako ideálny počet n-gramov $N = 3$.

[4] V tomto kroku sa za súvetie považujú časti vety oddelené akýmkoľvek nealfanumerickým znakom (napr. čiarkou). Toto pravidlo sa neuplatňuje v prípade, že je nealfanumerický znak súčasťou slova (napr. desatinné číslo 3,14).

[5] Vychádzame z predpokladu, že aj v televíznom programe sa názvy relácií začínajú vždy veľkým začiatočným písmenom. Analýzou príspevkov na sociálnej sieti Facebook sme zistili, že tvorcovia sociálneho obsahu sa správajú obdobne a vo vetách používajú taktiež veľké písmeno na pomenovanie názvu TV relácie. Výnimku tvoria hashtagy, čo sme však pri návrhu zohľadnili.



Obr. 6.5: Príklad extrakcie pomenovaných entít z príspevku. Uvedená ukážka ilustruje hľadanie slov začínajúcich na veľké písmená a následné porovnanie s prefixovým indexom.

Extrakciu pomenovaných entít ilustrujeme na príklade na obrázku č. 6.5. Ako môžeme vidieť, z príspevku vzniknú tri extrahované pomenované entity pre $N = 0, 1, 2$. Pre $N = 3$ už nemôžeme vygenerovať žiadne slovné n-gramy (vzhľadom na to, že sme na konci vety).

6.3.2 Extrakcia hashtagov

Vytvorené kandidátne hashtagy z televízneho programu potrebujeme porovnať s podobnou množinou, ktorú môžeme získať zo sociálnej siete. Vychádzajúc zo základnej charakteristiky hashtagu (začína znakom „#“) bude celý proces pozostávať iba z prehľadávania textu príspevku a vyhľadania všetkých slov začínajúcich znakom „#“. Tieto slová budeme pre účely mapovania považovať za hashtagy získané z príspevkov sociálnych sietí alebo tiež **extrahované hashtagy**.

6.3.3 Spracovanie prepojení

Súčasťou mnohých príspevkov na sociálnej sieti je prepojený obsah. Hoci sa spravidla stretávame s prepojeniami na textový obsah (napr. články), získanie textu a hlavne jeho spracovanie môže byť z hľadiska procesu mapovania časovo náročné a vzhľadom na dynamiku webu často netriviálne.

Prepojený obsah jednoznačne charakterizuje hypertextové prepojenie v podobe

Obr. 6.6: Ukážka rozdelenia URL na doménu a subdoménu

URL (*angl.* Uniform Resource Locator) adresy – v našom prípade prepojenie na webovú stránku. Toto prepojenie v sebe môže niesť časť alebo priamo celý názov televíznej relácie. Ako preukázali výsledky našej analýzy, prepojenia sa vyskytujú ako v slovenskom, tak aj v zahraničnom obsahu, a preto sme sa rozhodli využiť aj analýzu URL adresy.

Pri spracovávaní URL adries pracujeme v našom riešení s doménami druhého (skrátene doména) a tretieho rádu (skrátene subdoména). Rozdelenie URL adresy na doménu a subdoménu ilustruje obrázok č. 6.6.

Pomocou vopred definovaných pravidiel (regulárnych výrazov) získame z jednotlivých URL adries domény a subdomény, ktoré budeme následne považovať za kandidátov na názov televíznej relácie. Vychádzame pritom z nasledujúcich pravidiel:

- z adresy URL odstránime znak spojovníka „-“[6],
- ak sa subdoména zhoduje s „www“, tak ju ďalej neuvažujeme,
- ak sa subdoména alebo doména nachádza v zozname ignorovaných názvov, taktiež ju ďalej neuvažujeme.

Zoznam *ignorovaných názvov* sme zostavili manuálne analýzou celej spracovávanej dátovej množiny zo sociálnej siete. Do tohto zoznamu patria nasledujúce slová:

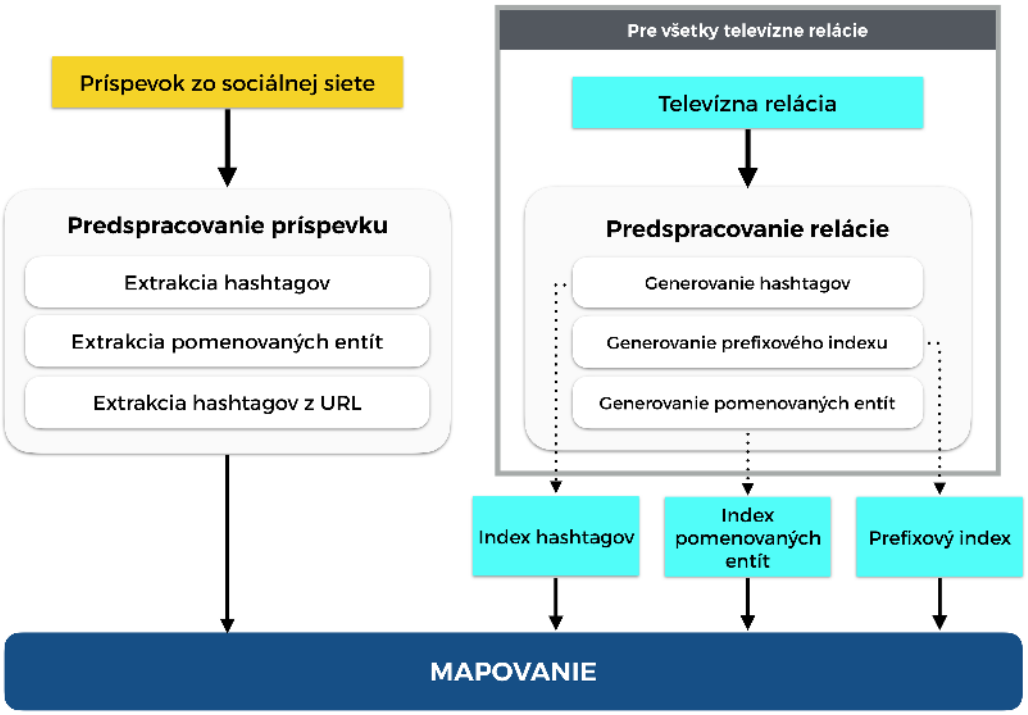
$$wwe,\ vwww,\ youtu,\ youtube,\ \{názov\ TV\ stanice\}$$

kde $\{názov\ TV\ stanice\}$ je názov aktuálne spracovávanie televíznej stanice.

Aplikovaním uvedených pravidiel na každú z URL adries získame kandidáta na názov televíznej relácie (t.j. kandidátnu reláciu). Vzhľadom na charakter doménových mien však názov neobsahuje medzery, veľké písmená ani diakritiku. Do veľkej miery teda spĺňa charakteristiky hashtagov, a preto ho zaradíme medzi **extrahované hashtagy z prepojenia**.

V prípade, že doména ani subdoména neindikujú TV reláciu, ku ktorej by mohol príspevok patriť, využijeme všetky časti URL adresy. Celú URL adresu (za prvou

[6]Znak spojovníka je jediný nealfanumerický znak, ktorý je povolený v názvoch domén a subdomén.



Obr. 6.7: Pohľad na celý proces predspracovania

lomkou) rozdelíme na podčasti (oddelené lomkou), pričom každú podčasť budeme považovať za extrahovaný hashtag. V procese mapovania následne vylúčime tie hashtagy, ktoré nie sú vhodné. Pri návrhu nášho riešenia sme vychádzali z práce [34], kde autori analyzovali URL adresu a vyhľadávali v tej signifikantné časti pre účely extrakcie pomenovaných entít.

Na obrázku č. 6.7 sa nachádza sumárny pohľad na proces predspracovania príspevku zo sociálnej siete a predspracovania názvov televíznych relácií z televízneho programu.

6.4 Realizácia mapovania

Po vytvorení televíznych metadát a predspracovaní príspevku zo sociálnej siete prechádzame do fázy mapovania. V tejto fáze vyhľadávame črty extrahované z príspevku v televíznych metadát a snažíme sa odhaliť mieru súvislosti medzi nimi.

Vychádzame zo základného predpokladu, že **pre každý príspevok vieme povedať, ktorá televízna stanica** ho na sociálnej sieti **publikovala**. Túto informáciu máme totiž prirodzene k dispozícii - každý príspevok odoslal oficiálny (a teda

overený) profil televíznej stanice[7]. V našom riešení sme sa nezaoberali možnosťou využiť neoficiálne profily na sociálnej sieti. Dôvodom tejto voľby bola najmä nespoľahlivosť obsahu, ktorý by neobsahoval len relevantné príspevky, ale často aj šum (napríklad reklamu odosielanú vlastníkmi neoficiálnych profilov).

Nech R_j je množina televíznych relácií, ktoré sa snažíme extrahovať z príspevku P_i sociálnej siete. Platí pritom, že každá relácia z R_j patrí do množiny relácií R_t zvolenej televízie t ($t \in T$, kde T je množina analyzovaných televízií). Analogicky aj príspevok P_i patrí do množiny príspevkov P_t z oficiálnej skupiny televízie t na sociálnej sieti. Cieľom mapovania je získať zoznam televíznych relácií $R_j \in R_t$, ktoré súvisia s príspevkom $P_i \in P_t$. Formálne sa snažíme vyčísliť hodnotu S_{ij}, ktorú môžeme vyjadriť ako mieru vzťahu súvislosti medzi televíznou reláciou R_j a príspevkom na sociálnej sieti P_i. S budeme nazývať **koeficient súvislosti relácie a príspevku**. Súvislosť S_{ij} pritom dosahuje hodnotu 0 práve vtedy, keď televízna relácia R_j nesúvisí s príspevkom P_i. Ak televízna relácia s príspevkom súvisí (aspoň čiastočne), vtedy dosahuje hodnoty v intervale 0 až 1 (okrem 0), pričom platí, že vyššia hodnota S_{ij} implikuje aj vyššiu mieru súvisu. Formálne platí 6.1 a 6.2.

$$\forall t \in T : \forall R_j \in R_t, \forall P_i \in P_t : \quad \exists S_{ij} = suvislost(R_j, P_i) \tag{6.1}$$

$$S_{ij} = \begin{cases} 0 & \text{ak } R_j \text{ nesúvisí s } P_i \\ (0, 1\rangle & \text{ak } R_j \text{ súvisí s } P_i \end{cases} \tag{6.2}$$

Predpokladom mapovania je, že máme k dispozícii bázu televíznych metadát a z príspevku sociálnej siete sme extrahovali pomenované entity, hashtagy aj hashtagy z prepojení. Všeobecný postup mapovania potom pozostáva z nasledujúcich základných krokov:

1. Pre každú extrahovanú pomenovanú entitu vyhľadáme v indexe pomenovaných entít **kandidátne televízne relácie**, ktoré pridáme do množiny kandidátnych televíznych relácií (označujeme C).
2. Pre každý extrahovaný hashtag vyhľadáme v indexe hashtagov **kandidátne televízne relácie**, ktoré pridáme do množiny kandidátnych televíznych relácií C.
3. Pre každý extrahovaný hashtag z prepojenia vyhľadáme v indexe hashtagov **kandidátne televízne relácie**, ktoré pridáme do množiny kandidátnych televíznych relácií C.

[7]Zoznam oficiálnych profilov môžeme získať návštevou oficiálnych stránok televíznych staníc. Oficiálne stránky televíznych staníc spravidla odkazujú na svoje profily na sociálnych sieťach.

4. Na základne obsahu množiny **kandidátnych televíznych relácií** C vypočítame jednotlivé koeficienty súvislosti S.

Pri **vyhľadávaní v indexoch** najskôr nájdeme všetky položky (kandidátov) z indexu na základe prvého písmenka extrahovanej črty z príspevku. Následne vypočítame skóre súvislosti medzi každým kandidátom a samotnou extrahovanou črtou. Skóre súvislosti vypočítame na základe príslušnej metriky:

- pre nájdenie súvislosti medzi kandidátnymi a extrahovanými pomenovanými entitami používame ako hlavnú metriku **n-gramovú podobnosť** (z práce [22]),
- pre nájdenie súvislosti medzi kandidátnymi a extrahovanými hashtagmi používame ako hlavnú metriku **binárnu podobnosť**,
- pre nájdenie súvislosti medzi kandidátnymi a extrahovanými hashtagmi získanými z externých prepojení používame ako hlavnú metriku **n-gramovú podobnosť s prefixom**.

Príklad: Z príspevku „Pozerajte dnes večer #kredenc.“ získame extrahovaný hashtag „kredenc“. Z indexu hashtagov získame všetky hashtagy nachádzajúce sa pod písmenkom „k“. Medzi každým hashtagom z indexu a hashtagom „kredenc“ následne vypočítame binárnu podobnosť.

Ako sme uviedli v kapitole *Extrakcia pomenovaných entít*, pre každú kandidátnu pomenovanú entitu generujeme slovné n-gramy rôznej dĺžky. Pri **vyhľadávaní pomenovaných entít** v indexe pomenovaných entít však vyberáme z kandidátov iba tých, ktorí majú dĺžku zhodnú s dĺžkou samotnej pomenovanej entity[8]. N-gramová podobnosť nám následne zabezpečí vyššiu odolnosť voči preklepom a nepriaznivým gramatickým vlastnostiam rôznych jazykov (napr. slovenčiny).

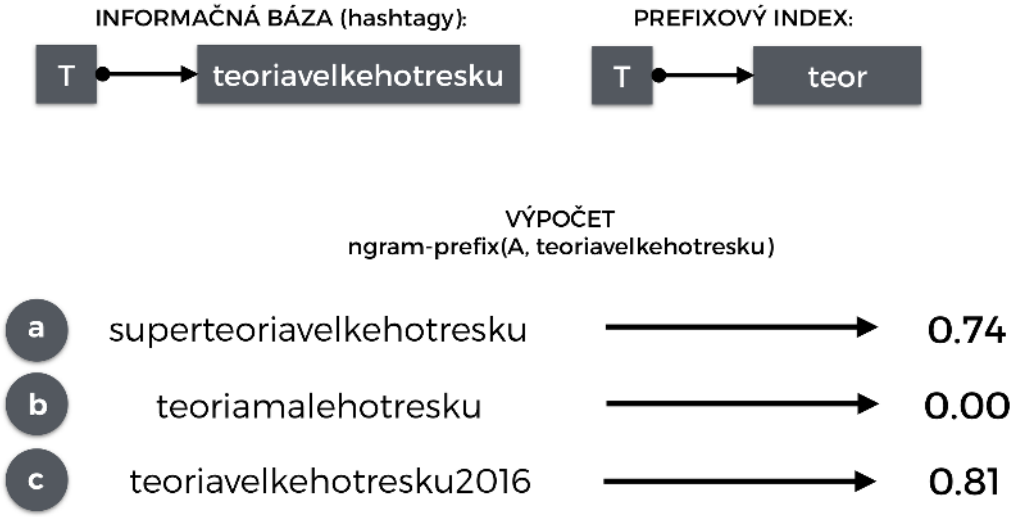
Binárnu podobnosť definujeme nasledovne:

$$BIN\_SIM(x, y) = \begin{cases} 1 & \text{if } LEVENSHTEIN(x, y) \leq 1 \\ 0 & \text{if } LEVENSHTEIN(x, y) > 1 \end{cases} \tag{6.3}$$

kde $LEVENSHTEIN(x, y)$ je Levenshteinova vzdialenosť reťazca x od reťazca y.

Binárna podobnosť je pre hashtagy vhodnejšia voľba ako n-gramová podobnosť, nakoľko sa v nich ohýbanie slovných druhov spravidla nevyskytuje. Hashtagy majú navyše kratšiu dĺžku a použitie n-gramovej podobnosti by mohlo spôsobiť nekorektné výsledky.

[8]Pripomíname, že pri generovaní kandidátnych entít vynechávame stopslová (napr. názov TV stanice a anglické členy). Do celkovej dĺžky pomenovanej entity a kandidátnej pomenovanej entity preto stopslová nezapočítavame.



Obr. 6.8: Ukážka výpočtu n-gramovej podobnosti s prefixom.

Pri porovnávaní hashtagov získaných z URL prepojenia využívame **n-gramovú podobnosť s prefixom**. Nech A je hashtag extrahovaný z externého prepojenia a B je hashtag nachádzajúci sa v báze televíznych metadát. Potom n-gramová podobnosť s prefixom (označujeme $NGRAM\_PREFIX$) je definovaná nasledovne:

- ak B nie je prefixom alebo sufixom A, potom je $NGRAM\_PREFIX(A, B) = 0$,
- inak $NGRAM\_PREFIX(A, B) = NGRAM\_SIM(A, B)$.

kde $NGRAM\_SIM(A, B)$ je n-gramová podobnosť medzi reťazcami A a B (z práce [22]).

Na obrázku č. 6.8 sa nachádza ukážka výpočtu n-gramovej podobnosti s prefixom. Túto podobnosť počítame pre tri extrahované hashtagy (a) *superteoriavelkehotresku*, (b) *teoriamalehotresku* a (c) *teoriavelkehotresku2016*. Porovnávame sa pritom s hashtagom z bázy televíznych metadát - *teoriavelkehotresku*. V prvom prípade (a) je hodnota n-gramovej podobnosti nenulová, keďže kandidát (a) je suffixom kandidáta z bázy televíznych metadát. V druhom prípade referenčný hashtag *teoriavelkehotresku* nie je prefixom ani sufixom *teoriamalehotresku* (rozdiel môžeme vidieť v podreťazci *male*). Posledným analyzovaným hashtagom je *teoriavelkehotresku2016*. V tomto prípade sú opäť splnené obe podmienky a pre získanie hodnoty n-gramovej podobnosti s prefixom aplikujeme štandardný výpočet n-gramovej podobnosti podľa práce Kodraka ([22]).

6.5 Určenie výsledného mapovania

V procese mapovania sa snažíme získať kandidátne televízne relácie. Pracujeme pritom s televíznymi metadátami. Tie vznikli generovaním z televízneho programu, pričom na generovanie boli použité viaceré prístupy. Príkladom sú napríklad hashtagy, ktoré sme vytvárali zlúčením slov (napr. „Búrlivá citronáda“ → „burlivacitronada“), ale aj s využitím začiatočných písmen slov (napr. „Búrlivá citronáda“ → „bc“). Výber postupu môže mať zásadný vplyv aj na mieru, do akej je kandidátna televízna relácia v kontexte nášho riešenia správna resp. žiadúca. To musíme zohľadniť pri výpočte koeficienta súvislosti S_{ij}, keďže pri identifikácii kandidátnych relácií sa môže stať, že získame jednu kandidátnu reláciu viackrát a nebudeme vedieť jednoznačne určiť jej výslednú podobnosť.

Tabuľka 6.2 ilustruje kroky 1 - 3 nami navrhnutého algoritmu mapovania (strana 42) s prihliadnutím na to, akými rôznymi postupmi môžeme zrealizovať mapovanie. Stĺpec *Získanie kandidáta z príspevku* vyjadruje spôsob, akým z príspevku sociálnej siete vygenerujeme kandidáta (či už vo forme pomenovanej entity alebo hashtagu). Stĺpec *Získanie kandidáta z TV relácie* analogicky vyjadruje spôsob, akým z názvu televíznej relácie zostavíme kandidáta (opäť vo forme pomenovanej entity alebo hashtagu). Platí pritom, že každý riadok predstavuje jeden postup **vyhľadávania kandidátnych televíznych relácií**.

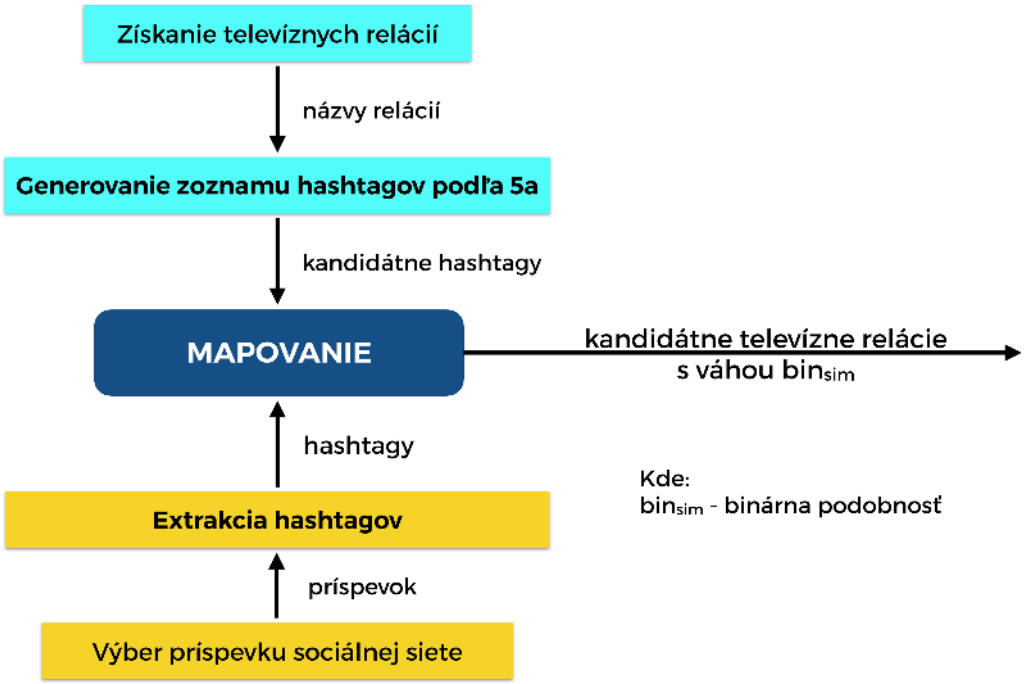
Tabuľka 6.2: Prehľad použitých kombinácií postupov pri mapovaní

k	Získanie kandidáta z príspevku	Získanie kandidáta z TV relácie
1-3	získanie hashtagov z URL prepojeného obsahu	generovanie hashtagov (a - c)
4-7	extrakcia hashtagov	generovanie hashtagov (a - c)
8	získanie entít (s rozdelením súvetí)	generovanie entít (s rozdelením súvetí)
9	získanie entít (bez rozdelenia súvetí)	generovanie entít (bez rozdelenia súvetí)

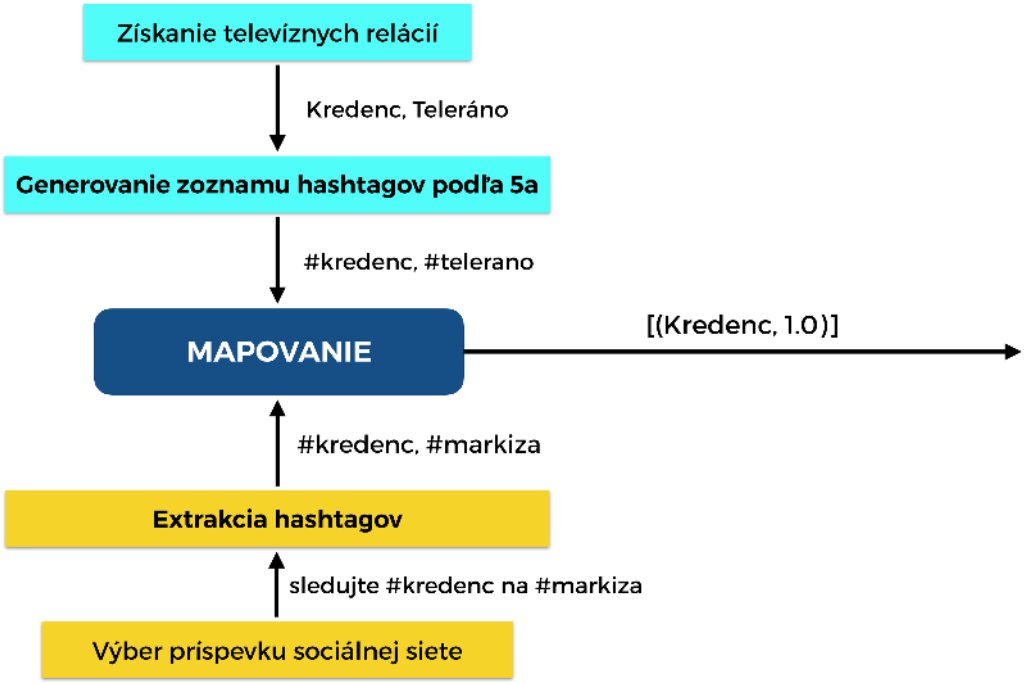
Poznámka ku generovaniu hashtagov: V tabuľke 6.2 používame skratku a - c, ktorá vyjadruje, že boli použité kroky a, b a c - jednotlivo a samostatne. Jednotlivé písmenká *a-c* korešpondujú postupu generovania hashtagov uvedenom na strane 35.

Obrázok 6.9 zobrazuje príklad mapovania založený na porovnaní hashtagov z príspevku na sociálnej sieti a hashtagov vygenerovaných z názvu televíznej relácie podľa spôsobu 5a.

Obrázok 6.10 ilustruje konkrétny príklad namapovania príspevku sociálnej siete s textom *„sledujte #kredenc na #markiza“* na televíznu reláciu s názvom *„Kredenc“*.



Obr. 6.9: Ukážka mapovania pomocou hashtagov



Obr. 6.10: Príklad mapovania pomocou hashtagov

Obdobne postupujeme aj pri ďalších spôsoboch mapovania, pričom vo výsledku získame ohodnotený zoznam kandidátnych televíznych relácií. Ak sa v rámci jedného

spôsobu mapovania (riadok tabuľky) vyskytne televízna relácia opakovane, jej výsledné skóre súvislosti S_{ijk} získame ako priemer jej súvislostí pre daný spôsob vytvárania mapovania.

Výsledné skóre S_{ij} televíznej relácie R_j a príspevku P_i získame tak, že prechádzame celú množinu kandidátnych televíznych relácií pre príspevok P_i a nájdeme všetky výskyty televíznej relácie R_j s príslušnými čiastkovými skóre S_{ijk} (kde k reprezentuje použitý spôsob mapovania, viď prvý stĺpec tabuľky 6.2). Postup uvedený na strane 42 rozšírime o výpočet skóre, čím získame finálny postup pre vytvorenie štandardného mapovania:

1. Inicializujeme skóre S_{ij} pre televíznu reláciu R_j a príspevok P_i na hodnotu 0.
2. Každú položku zo zoznamu extrahovaných pomenovaných entít príspevku P_i vyhľadáme v indexe pomenovaných entít a určíme n-gramovú podobnosť, čím získame kandidátne televízne relácie.
3. Každú položku zo zoznamu extrahovaných hashtagov príspevku P_i vyhľadáme v indexe hashtagov a určíme binárnu podobnosť, čím získame kandidátne televízne relácie. Tento postup opakujeme pre každý zo spôsobov generovania hashtagov (a až c).
4. Každú položku zo zoznamu extrahovaných hashtagov príspevku P_j (z URL prepojeného obsahu) vyhľadáme v indexe hashtagov a určíme n-gramovú podobnosť s prefixom, čím získame kandidátne televízne relácie. Tento postup opakujeme pre každý zo spôsobov generovania hashtagov (a až c).
5. Získame 9 rôznych skupín s kandidátnymi televíznymi reláciami. Ak sa v skupine vyskytujú televízne relácie viackrát, ich viacnásobný výskyt nahradíme jedinou kandidátnou televíznou reláciou, ktorej skóre podobnosti bude predstavovať *priemerné skóre podobností* danej kandidátnej televíznej relácie v rámci skupiny[9] (vzťah 6.4).
6. Pre každú kandidátnu televíznu reláciu vyberieme skupiny, v ktorých majú jednotliví kandidáti z príspevku hodnotu súvislosti vyššiu ako je prahová hodnota $\boldsymbol{rel_{thres}}$ (prahovú hodnotu objasníme na záver kapitoly).
7. Výsledné skóre pre televíznu reláciu R_j a príspevok P_i vypočítame ako priemer skóre televíznej relácie z vybraných skupín (vzťah 6.5).

Poznámka: *V krokoch 2-4 pridávame do skupiny kandidátnych televíznych relácií iba tie televízne relácie, ktorých skóre súvislosti je vyššie ako* 0.

Pre výpočet súvislosti relácie R_j a príspevku P_i podľa postupu/skupiny k (z

[9]Intuitívne môžeme predpokladať, že v rámci skupiny sa budú nachádzať iba kandidátne televízne relácie, ktoré budú mať skóre vyššie ako 0.

tabuľky 6.2) použijeme vzťah 6.4:

$$S_{ijk} = \sum_{x=1}^{k_i} \sum_{y=1}^{k_j} \frac{sim(x,y)}{COUNT\_IF(sim(x,y) > 0)} \tag{6.4}$$

kde:

- $\boldsymbol{k_i}$ je počet kandidátov v skupine k získanej z televíznej relácie s indexom i (alebo tiež veľkosť invertovaného indexu),
- $\boldsymbol{k_j}$ je počet kandidátov v skupine k získanej z príspevku sociálnej siete s indexom j,
- $\boldsymbol{sim(x,y)}$ je príslušná metrika pre výpočet podobnosti medzi kandidátom x z bázy televíznych metadát a kandidátom y z príspevku.

Pre výpočet výslednej súvislosti relácie R_j a príspevku P_i použijeme vzťah 6.5:

$$S_{ij} = \frac{\sum_{k=1}^{9} w_k * S_{ijk}}{\sum_{k=1}^{9} w_k} \tag{6.5}$$

kde:

- $\boldsymbol{w_k}$ sa rovná 1, ak z postupu k vznikli nejaké kandidátne televízne relácie.

Pri mapovaní pomenovaných entít a hashtagov z prepojení využívame metriku n-gramovej podobnosti, ktorej hodnoty sa nachádzajú v intervale $< 0, 1 >$. V zozname kandidátnych relácií sa preto môžu nachádzať aj také relácie, ktoré majú nízku mieru súvislosti s príspevkom sociálnej siete. Záverečným krokom postupu je preto vylúčenie tých relácií, ktoré nepovažujeme za súvisiace vzhľadom na ich hodnotu skóre S_{ij}. Na vylúčenie použijeme nami definovanú prahovú hodnotu $\boldsymbol{rel_{thres}}$, ktorá je zároveň parametrom nášho algoritmu mapovania. Tento krok sme uviedli ako krok č. 6 nášho postupu mapovania na strane č. 47.

Stanovenie konkrétnej prahovej hodnoty sme realizovali formou experimentu. Na vybranej množine príspevkov sme vykonali mapovanie. Postupne sme menili hodnotu parametra rel_{thres}, pre každú hodnotu vykonali experiment a následne overili výsledky mapovania. Z každého experimentu sme tak získali hodnoty metrík, ktoré sme sledovali (napr. F1-skóre).

Výslednú prahovú hodnotu sme následne určili ako takú **hodnotu parametra** $\boldsymbol{rel_{thres}}$, pri ktorej sme dosiahli maximálne hodnoty F1-skóre. Podrobnosti vykona-

ného experimentu spoločne s výsledkami uvádzame v kapitole *Overenie mapovania*.

6.6 Rozšírenie štandardného mapovania

Doposiaľ sme uvažovali, že v každom príspevku sociálnej siete sa nachádzajú pomenované entity, hashtagy alebo externé prepojenia, ktoré nám pomôžu pri určení výsledného mapovania. Niekedy však tieto črty nemusia byť užitočné alebo, v tom horšom prípade, vôbec prítomné. Náš algoritmus sa tak môže dostať do situácie, kedy nevie rozhodnúť o korektnom mapovaní, prípadne nevie určiť žiadne mapovanie.

Riešením tejto situácie je použitie *dynamických televíznych metadát*. Dynamické televízne metadáta získame z výsledkov mapovania, ktoré už algoritmus vykonal. Týmito výsledkami sú dvojice *príspevok - televízna relácia*. Využijeme pritom práve črty, ktoré sa v príspevku nachádzali a ktoré algoritmus mapovania extrahoval. Tento koncept vychádza z myšlienky ko-učenia, ktorej sme sa venovali v kapitole Ko-učenie.

Pri zostavovaní dynamických televíznych metadát vychádzame z výstupov **štandardného mapovania**. Za štandardné mapovanie považujeme metódu mapovania, ktorá využíva televízne metadáta a črty extrahované z príspevku. Opisu štandardnej metódy mapovania sme sa venovali v predchádzajúcich kapitolách.

Výstupom štandardnej metódy mapovania sú príspevok a príslušné súvisiace televízne relácie. Dynamické televízne metadáta budú pozostávať z televíznych relácií a k nim prislúchajúcim črtám, ktoré sme získali z príspevku. Zamerali sme sa pritom na nasledujúce črty:

- text príspevku,
- hashtagy,
- hashtagy extrahované z externého prepojenia.

V porovnaní so štandardným mapovaním využívame namiesto pomenovaných entít celý text príspevku. V prípade, že by sme si zvolili iba pomenované entity, nepokryli by sme tým prípad, kedy sa v príspevku nenachádza žiadna z čŕt.

Ak pracujeme s príspevkom ako celkom, môžeme sa naň pozerať ako na dokument. Pri hľadaní podobnosti dvoch dokumentov nemusí byť n-gramová podobnosť postačujúca. V súčasnej fáze riešenia sme sa preto rozhodli aplikovať model TF-IDF, ktorý zohľadňuje aj frekvenciu termov v dokumente a frekvenciu termov naprieč dokumentmi. Tento koncept sme použili aj na hashtagy, čím nám vznikli dva rôzne TF-IDF modely, ktoré si v nasledujúcej podkapitole priblížime.

6.6.1 Vytvorenie dynamických televíznych metadát

Nech P_i je príspevok, z ktorého sme pomocou štandardného mapovania získali televízne relácie R_i (relácia R_{ij} patrí do R_i). Potom dynamické televízne metadáta môžeme zostaviť použitím nasledujúceho postupu:

1. Z textu príspevku P_i získame tokeny - odstránime všetky znaky okrem písmen a číslic.
2. Ak sa relácia R_{ij} z množiny relácií R_i nenachádza v báze dynamických televíznych metadát, pridáme ju do nej (inak pokračujeme krokom č. 3). Vytvoríme zároveň dva vektory: **vektor tokenov entít** obsahujúci tokeny vznikajúce v prvom kroku postupu a **vektor tokenov hashtagov** obsahujúci hashtagy extrahované z príspevku aj z externého prepojenia.
3. Do vektora tokenov entít pridáme tokeny, ktoré vznikli v prvom kroku.
4. Do vektora tokenov hashtagov pridáme hashtagy, ktoré sme získali z príspevku a ktoré sme extrahovali z URL adresy externého prepojenia.
5. Z oboch vektorov zostavíme dva modely TF-IDF: **model textu príspevku** a **model hashtagov**. Ak tieto modely už existovali, aktualizujeme ich. V modeli TF-IDF budeme pritom za dokument považovať televíznu reláciu a jeho obsahom budú tokeny z vektorov.

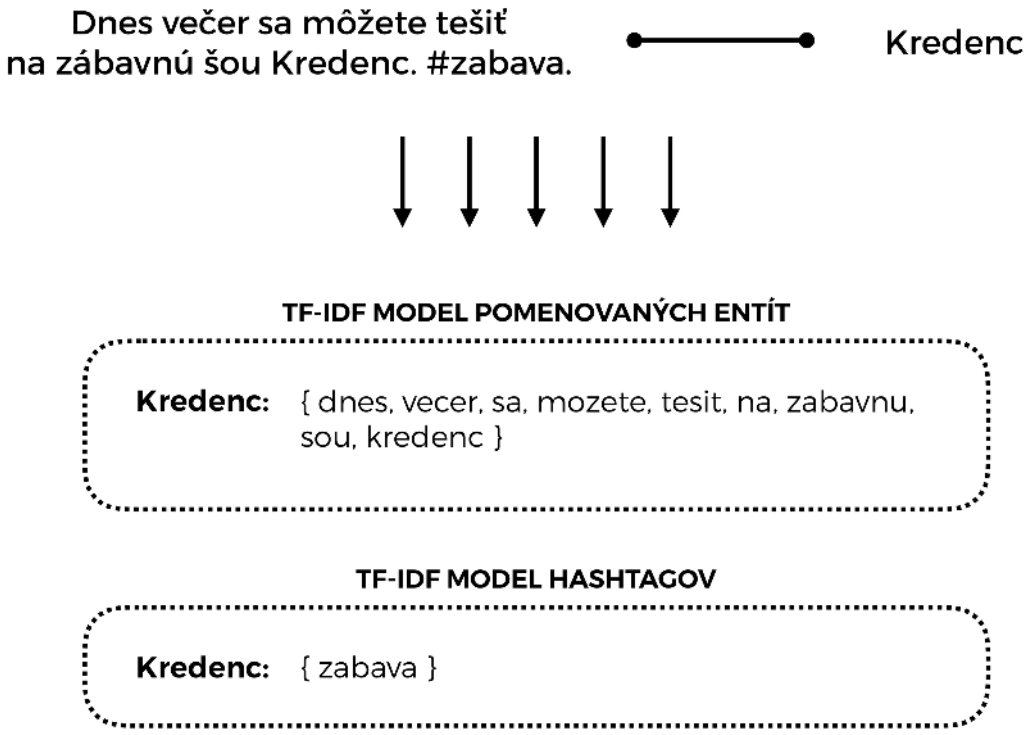
Vytvorenie oboch TF-IDF modelov ilustrujeme príkladom na obrázku č. 6.11. Pre zjednodušenie neuvádzame pri príslušných vektoroch vypočítané hodnoty pre jednotlivé termy.

Dynamická informačná báza sa teda skladá z dvoch TF-IDF modelov: *modelu tokenov* (obsahujúci tokeny z textu príspevku) a *modelu hashtagov* (obsahujúci hashtagy z príspevku a z externých prepojení). Tieto modely môžeme použiť pri mapovaní príspevku, pri ktorom by štandardné mapovanie zlyhalo. Na druhej strane, ak štandardné mapovanie uspeje, dynamickú informačnú bázu môžeme aktualizovať o novozískané črty z príspevku.

6.6.2 Použitie dynamických televíznych metadát

Príspevok, pre ktorý sme nevedeli pomocou štandardného mapovania určiť súvisiace televízne relácie, sa pokúsime namapovať s využitím dynamickej informačnej bázy. Postupujeme podobne, ako pri vytváraní dynamickej informačnej bázy.

1. Z textu príspevku P_i získame tokeny: odstránime všetky znaky okrem písmen a číslic.



Obr. 6.11: Ukážka vytvorenia TF-IDF modelov

2. Z tokenov zostavíme vektor modelu TF-IDF, ktorý pomocou **kosínovej podobnosti** porovnáme s ostatnými vektormi z modelu textu príspevku (viď. predchádzajúca podkapitola).
3. Z príspevku P_i a externého prepojenia extrahujeme hashtagy.
4. Hashtagy usporiadame do vektora TF-IDF, ktorý pomocou **kosínovej podobnosti** porovnáme s ostatnými vektormi z modelu hashtagov (viď. predchádzajúca podkapitola).
5. Použitím oboch TF-IDF modelov získame zoznam dokumentov (v skutočnosti zoznam televíznych relácií) s vypočítanými kosínovými podobnosťami. Ak sa televízna relácia vyskytne v zozname dvakrát (našli sme ju pomocou oboch modelov), pracujeme s vyššou kosínovou podobnosťou.

Poznámka: *Od vysvetlenia porovnávania vektorov a zostavovanie TF-IDF modelu abstrahujeme, nakoľko sme použili štandardné postupy používané v tejto oblasti.*

Výsledkom uvedeného postupu je zoznam televíznych relácií, ktoré do určitej miery súvisia s daným príspevkom. V súčasnej fáze riešenia sme ešte nezaviedli vhodný spôsob, ako zo zoznamu vybrať tie televízne relácie, ktoré môžeme s istotou prehlásiť ako súvisiace s príspevkom. Pre účely vyhodnotenia bázy dynamických

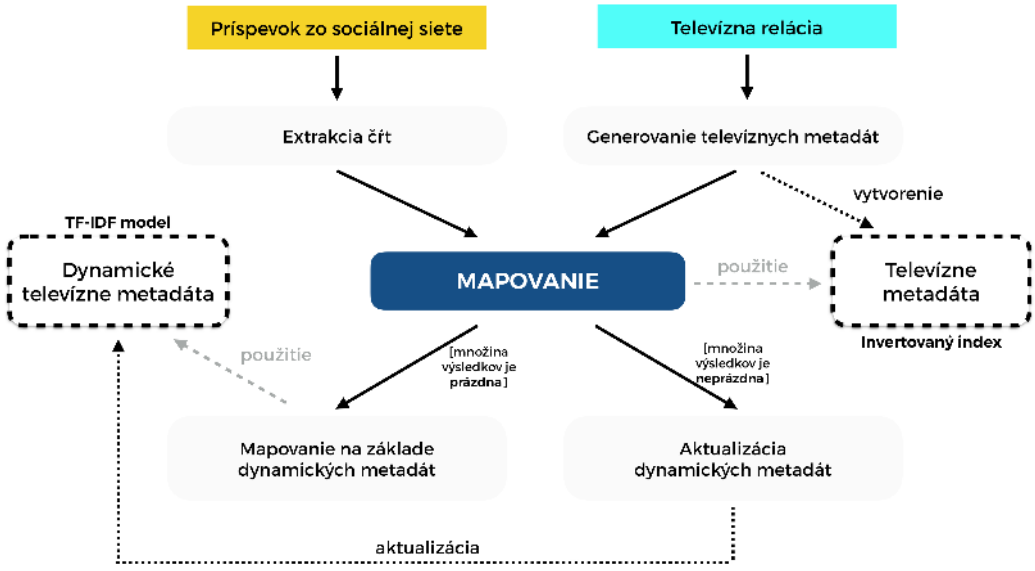
televíznych metadát považujeme za výsledok mapovania tú televíznu reláciu, ktorá má najvyššie skóre kosínovej podobnosti.

6.7 Zhrnutie

Na obrázku č. 6.12 sa nachádza celkový pohľad na proces mapovania s prihliadnutím na využitie dynamických televíznych metadát. Pri predspracovaní televíznych relácií vytvárame televízne metadáta. Po príchode príspevku zo sociálnej siete ho predspracujeme a následne vstupuje do procesu mapovania. Následne sa tok rozdeľuje podľa toho, či je množina výsledkov štandardného mapovania prázdna alebo neprázdna. Ak sa v množine výsledkov nachádzajú televízne relácie, aktualizujeme (čierna prerušovaná čiara) dynamické televízne metadáta (reprezentované TF-IDF modelom). V prípade, že je množina výsledkov štandardného mapovania prázdna, používame existujúce dynamické metadáta (sivá prerušovaná čiara). Pripomíname, že mapovanie na základe dynamických metadát vychádza iba z výsledkov predchádzajúcich mapovaní.

Vo všeobecnosti platí, že nami navrhnutá metóda si nevyžaduje žiadnu prechádzajúcu anotáciu príspevkov zo sociálnej siete a vychádza iba z dvoch základných informácií:

- informácia o tom, ktorá televízna stanica publikovala mapovaný príspevok na sociálnej sieti,
- zoznam všetkých televíznych relácií príslušnej televíznej stanice.



Obr. 6.12: Celkový pohľad na proces mapovania

7. Implementácia mapovania

Navrhnuté riešenie mapovania televízneho programu na príspevky sociálnej siete bolo implementované v jazyku **Ruby**. Ako primárnu sociálnu sieť sme si zvolili sociálnu sieť **Facebook**. Pre získavanie televízneho programu a príspevkov zo sociálnej siete sme vytvorili automatizovaný sťahovač, ktorý pravidelne získaval potrebné dáta. Tento sťahovač bol implementovaný v jazyku **PHP** a predstavoval samostatný komponent aplikácie nezávislý od mapovania.

V prílohe *Inštalácia a návod na použitie* uvádzame inštalačnú príručku a ďalšie užitočné inštrukcie, ktoré môžu dopomôcť k spusteniu nášho riešenia.

7.1 Architektúra riešenia

Riešenie pracuje ako súbor viacerých knižníc, ktoré spoločne poskytujú požadovanú funkcionalitu - mapovanie. V našom riešení sme identifikovali 5 základných komponentov: *Sťahovač TV Programu*, *Sťahovač dát z Facebooku*, *Úložisko*, *Mapovanie* a *Konzola*.

Sťahovač TV Programu. Umožňuje získavanie a ukladanie TV programu z externého zdroja TV programu (vo formáte XML).

Sťahovač dát z Facebooku. Pomocou rozhrania Graph API[1], ktoré poskytuje služba Facebook, získava v pravidelných intervaloch verejné príspevky zo sociálnej siete Facebook. Príspevky sú ukladané vo forme súborov formátu JSON. Opis štruktúry, v ktorej ich uchovávame, uvádzame v prílohe *A.1 Štruktúra dát príspevkov sociálnej siete*.

Úložisko. Umožňuje prácu s databázou tvorenou sťahovanými TV programami (v podobe televíznych metadát) a príspevkami zo sociálnej siete Facebook. Úložisko túto databázu získava prostredníctvom rozhrania, ktoré vystavujú komponenty riadiace sťahovanie. Z pohľadu architektúry reprezentujeme úložisko pomocou NoSQL databázy (JSON). Na základe toho, s ktorými dátami pracujeme, poskytuje komponent *Úložisko* dve základné rozhrania:

Načítavanie televíznych metadát. Predstavuje rozhranie pre prácu s TV programom. Umožňuje jeho získanie (z úložiska), analýzu, filtrovanie a ukladanie. Pre ostatné komponenty poskytuje rozhranie, vďaka ktorému je možné získať z programu iba potrebné údaje (v kontexte nášho riešenia televíznu stanicu a názvy jednotlivých relácií).

[1] https://developers.facebook.com/docs/graph-api

Načítavanie príspevkov. Predstavuje rozhranie pre prácu s dátami zo sociálnej siete Facebook. Podobne ako komponent DB TV Programu, umožňuje získať príspevky uložené v úložisku, ich filtráciu a agregáciu na základe zvolených parametrov.

Mapovanie. Realizuje funkcionalitu mapovania príspevkov zo sociálnej siete a inštancií televízneho programu (televíznych relácií). Ide o komponent, ktorý sa skladá z viacerých menších súčiastok, ktoré navzájom spolupracujú pre dosiahnutie potrebného výsledku.

Konzola. Používateľské rozhranie umožňujúce spúšťanie príkazov aplikácie (a to najmä samotného predspracovania a mapovania).

Uvedené komponenty predstavujú vysokoúrovňový pohľad na celkovú architektúru systému. Vzájomné prepojenia medzi nimi môžeme vidieť na diagrame komponentov (obrázok 7.1).

Považujeme za dôležité poznamenať, že uvedený model nezohľadňuje ukladanie výsledkov mapovania ani ich následné využitie pre potreby extrakcie metadát, nakoľko ich náš návrh a implementácia v tejto fáze projektu ešte bližšie nešpecifikujú.

Medzi externé knižnice, ktoré sme pri implementácii použili patria:

- **Facebook SDK for PHP**[2] (jazyk PHP) - pre získanie a spracovanie dát zo sociálnej siete Facebook,
- **nokogiri**[3] (jazyk Ruby) - pre spracovanie XML súborov.

7.2 Moduly nižšej úrovne

Súčasťou našej implementácie sú aj moduly nižšej úrovne, ktoré sme používali pri riešení mnohých čiastkových úloh. Z dôvodu redundancie ich neuvádzame v súhrnnom diagrame komponentov. Tieto moduly tvorili súčiastky komponentu *Mapovanie*.

Naša implementácia zahŕňa nasledujúce moduly nižšej úrovne:

- generátor n-gramov,
- generátor hashtagov z (pomenovanej) entity,
- generátor (pomenovaných) entít,
- extrahovač metadát o seriáli,
- lematizér (v súčasnosti nepoužívaný).

[2] https://developers.facebook.com/docs/reference/php
[3] http://www.nokogiri.org

Extrahovač metadát o seriáli umožňuje z názvu seriálu získať informáciu o čísle aktuálnej série a aktuálnej časti série. Pri implementácii tohto modulu sme využili skutočnosť, že číslo série sa spravidla vyskytuje v TV programe v podobe rímskej číslice. Aktuálna časť série sa najčastejšie reprezentuje ako číslica nasledovaná bodkou uzatvorená do dvoch okrúhlych zátvoriek. Napríklad z názvu televíznej relácie „Alf IV. (6.)“ by sme zistili, že bola odvysielaná *6.* časť *4.* série seriálu *Alf*.

7.3 Dátový model

V súčasnej verzii implementácie pracujeme so zjednodušeným dátovým modelom, ktorý obsahuje iba tie položky, ktoré si naše riešenie vyžaduje. Vychádzali sme pritom z dostupných dát, ktoré sme získali pri sťahovaní TV programu. V kapitole *Technická dokumentácia* uvádzame podrobný opis dát, ktoré sme získali z jednotlivých dátových množín TV programu a sociálnej siete Facebook.

Grafickú reprezentáciu dátového modelu predstavuje diagram tried (obrázok 7.2).

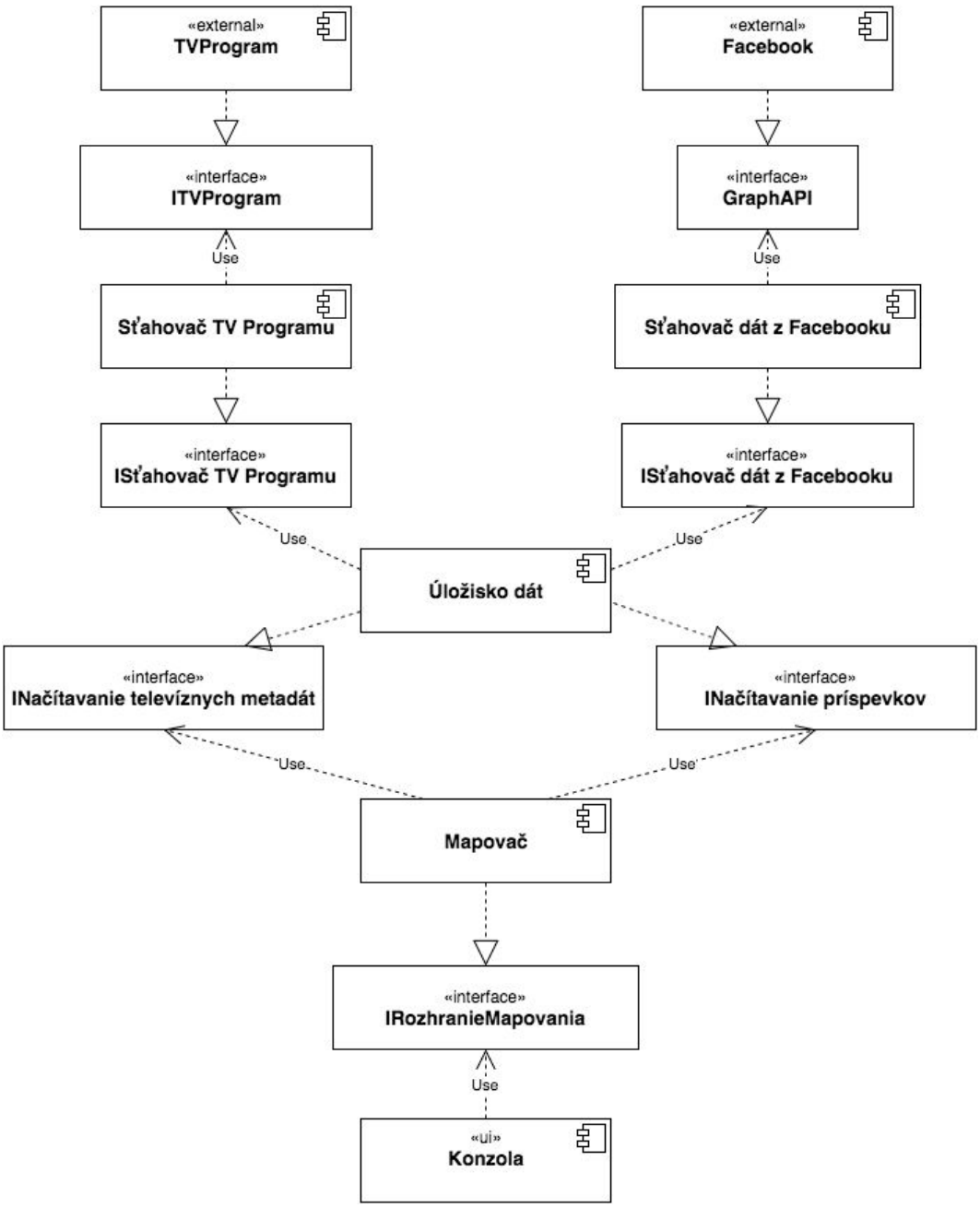
V dátovom modeli môžeme identifikovať nasledujúce entity:

Stanica. Obsahuje informácie o televíznych staniciach, ktorých TV program získavame a ktorých príspevky na sociálnej sieti skúmame.

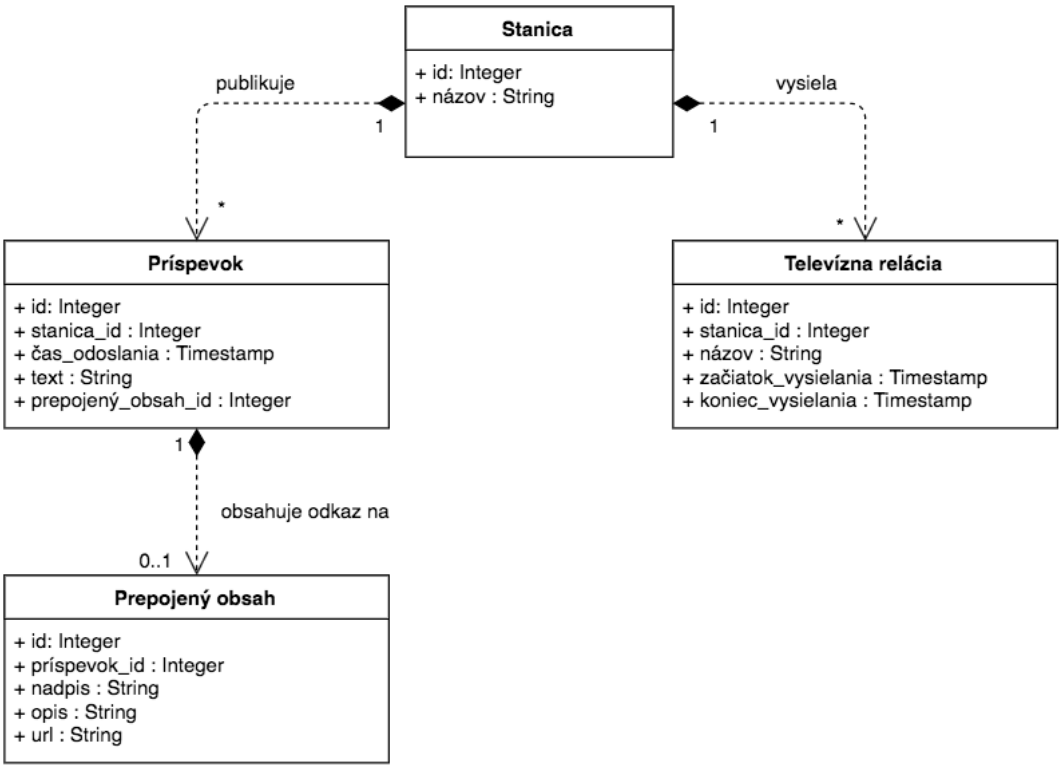
Príspevok. Reprezentuje príspevky zo sociálnej siete. Platí pritom, že o každom príspevku vieme povedať (bez potreby mapovania), ktorá televízna stanica ho publikovala.

Televízna relácia. Informácia o televíznej relácii získaná z TV programu. Hoci v aktuálnom návrhu riešenia nepracujeme so začiatkom a koncom vysielania, tieto údaje plánujeme využiť v ďalšej fáze.

Prepojený obsah. Obsahuje informácie o odkaze na prepojený obsah, ktorý je súčasťou príspevku na sociálnej sieti. Prepojený obsah je priamo previazaný na príspevok, avšak zároveň platí, že nie každý príspevok sociálnej siete ho obsahuje.



Obr. 7.1: Diagram komponentov implementovaného riešenia



Obr. 7.2: Diagram tried implementovaného riešenia

8. Overenie mapovania

Pri overovaní nášho riešenia sme postupovali inkrementálne a iteratívne. Jednotlivé časti riešenia sme implementovali postupne a na záver každej iterácie overili. Takýto postup nám umožnil nielen overiť správnosť nášho riešenia, ale zároveň sme získali cenné priebežné výstupy, ktoré nám pomohli náš návrh neustále zlepšovať.

V tejto kapitole sa bližšie pozrieme na opis použitej dátovej množiny, návrh experimentov a – v neposlednom rade – dosiahnuté výsledky. Zamerali sme sa pritom na overenie čiastkových príspevkov a vylepšení, ktoré sme navrhli, ale zároveň ponúkame pohľad aj na overenie metódy ako celku. Na záver kapitoly ponúkame porovnanie nami navrhnutej metódy mapovania s inými, už existujúcimi riešeniami.

Naše riešenie sme overovali s použitím nasledujúcich techník:

- **jednotkové testovanie** (*angl.* unit testing) - overovali (verifikovali) sme funkcionalitu jednotlivých modulov pomocou testovacieho rámca Rspec[1],
  - pri jednotkovom testovaní sme sa snažili postupovať vývojom riadeným testami (*angl.* test driven development), pričom sme tiež využívali externé služby na monitorovanie testov (Travis CI[2]),
  - využívali sme malú, vopred definovanú dátovú množinu,
- **overovanie návrhu** - overovali (validovali) sme správnosť riešenia ako celku,
  - zamerali sme sa tiež na to, či a do akej miery môže naše rozšírenie skvalitniť existujúce metódy,
  - využívali sme rozsiahlu dátovú množinu, ktorá sa snažila riešenie overiť nielen do šírky ale aj do hĺbky,
  - pri overovaní návrhu sme sa zamerali na vopred definované metriky.

V nasledujúcich kapitolách sa bližšie pozrieme na **Overovanie návrhu**.

8.1 Použité dátové množiny

Dôležitým krokom vedúcim k overeniu nášho riešenia bolo získanie dátovej množiny. Vzhľadom na charakter riešenia sme potrebovali mať informácie o televíznom programe a taktiež obsah zo sociálnej siete.

V období 1.5.2015 - 1.5.2016 sme získavali televízny program vybraných sloven-

[1] http://rspec.info/
[2] https://travis-ci.org/

ských a medzinárodných TV staníc. Z televízneho programu sme extrahovali názvy TV relácií, čím sme pre každú TV stanicu získali zoznam TV relácií, ktoré v danom období vysielala. V rovnakom období sme zároveň získavali príspevky zo sociálnej siete Facebook, pričom sme sa zamerali na oficiálne profily jednotlivých televízií.

Analyzovali sme pritom iba príspevky, ktoré zasielal oficiálny profil TV stanice. Dôvodom tejto voľby bol charakter komentárov používateľov, v ktorých sa takmer vôbec nevyskytovali žiadne z čŕt, ktoré boli základom nášho prístupu. Zároveň sme sa často stretávali s odbočovaním od témy (*angl.* offtopic) alebo s komentármi, ktoré neobsahovali žiadne relevantné informácie (napr. obsahovali iba obrázok alebo emotikony). Tejto problematike sme sa v našej práci nevenovali, a preto sme sa rozhodli zamerať sa iba na príspevky.

Tabuľka č. 8.1 obsahuje zoznam televíznych staníc, ktorých televízny program sme získavali, spolu s prepojením na zodpovedajúce profily na sociálnej sieti Facebook. Pri jednotlivých televíziách sa počty príspevkov rôznia. Keďže slovenských televíznych staníc (Markíza, JOJ) je menej, anotovali sme viac príspevkov. Pri samotnom vyhodnotení sme každú televíziu vyhodnocovali najskôr samostatne a až takéto čiastkové výsledky sme priemerovali do výsledných hodnôt metrík.

Tabuľka 8.1: Štatistiky dátovej množiny

Televízia	**Facebook profil**	**Počet unikátnych relácií**	**Počet anotovaných príspevkov**
Markíza	/TeleviziaMarkiza	184	250
JOJ	/tvjoj	165	250
CBS	/cbs	56	60
NBC	/nbc	54	60
The CW	/thecw	28	60
Fox	/foxtv	48	60
Spolu		535	740

Poznámka: Pred uvedené adresy Facebook profilov je potrebné zadať predponu http://www.facebook.com.

Ako môžeme vidieť, v prípade slovenských televízií je počet unikátnych televíznych relácií vyšší, keďže tieto televízie vysielajú široké spektrum produkcie: spravodajské relácie, filmy, seriály a aj vlastnú tvorbu. Takéto rozdiely však v našom prípade nie sú nevýhodou, keďže si môžeme overiť účinnosť nášho prístupu na rôznorodých typoch televízií.

Aby sme mohli vykonať overenie, potrebovali sme vytvoriť zlatý štandard – korektné mapovanie, voči ktorému môžeme porovnať našu metódu. Zlatý štandard sme vytvorili manuálnou anotáciou príspevkov zo sociálnej siete. V rámci anotovania

sme pre každý príspevok určili množinu televíznych relácií, s ktorou súvisí. K tejto množine sme dospeli manuálnou analýzou obsahu príspevku a prepojeného obsahu (externých odkazov). Uplatnili sme pritom nasledujúce pravidlá:

- televíznu reláciu sme k príspevku priradili iba v prípade, že sa vysiela na televíznej stanici, ktorá príspevok publikovala,
- ak príspevok obsahoval fotografiu bez uvedenia akéhokoľvek textového opisu alebo externého prepojenia, takýto príspevok sme vyhodnotili ako príspevok bez mapovaných TV relácií,
- ak príspevok obsahoval externé prepojenia, na základe ktorého URL adresy sa nedalo jednoznačne určiť mapovanie, toto externé prepojenie sme pri anotácií navštívili a analýzou obsahu sme získali korektné mapovanie[3],
- ak príspevok obsahoval zdieľaný príspevok (autor príspevku zdieľal príspevok z iného profilu), pri analýze sme zahrnuli aj obsah tohto príspevku (vrátane všetkých externých prepojení, ak boli prítomné).

8.2 Priebeh experimentov

Cieľom experimentov bolo kvantitatívne overiť vlastnosti nami navrhnutej metódy mapovania. Na základe výsledkov sme následne mohli vyhodnotiť prínos nielen metódy ako celku, ale aj jednotlivých čiastkových príspevkov, ktorými sme sa snažili zlepšiť existujúce riešenia v nami analyzovanej doméne.

Všeobecný postup pri realizovaní experimentov vychádzal z nasledujúcich krokov:

1. Príprava dátovej množiny (TV program stanice, anotované príspevky zo sociálnej siete).
2. Nastavenie parametrov metódy (vychádzalo z definície konkrétneho experimentu).
3. Realizácia mapovania.
4. Porovnanie výsledkov mapovania a anotácií príspevkov zo zlatého štandardu.
5. Vyhodnotenie výsledkov.

V kroku č. 4 sme pre každý príspevok porovnali TV relácie, ktoré sme získali pomocou mapovacej metódy a TV relácie, ktoré sa nachádzali v anotovanej dátovej

[3]Prepojenia zdieľané na sociálnej sieti odkazovali na články z webových portálov samotných televíznych staníc. Tieto webové portály mali jednotlivé články kategorizované, pričom táto kategória spravidla predstavovala samotnú TV reláciu.

množine (a považujeme ich za zlatý štandard). Na základe tohto porovnania sme následne vedeli exaktne určiť, ktoré TV relácie mapovanie vrátilo správne. Vychádzali sme pritom zo štandardných štatistických konceptov z oblasti vyhľadávania informácií – hľadali sme skutočné a falošné pozitívne prípady (*angl.* true/false positives).

Nech P je skúmaný príspevok, COR_MAP je množina TV relácií, ktorá súvisí s daným príspevkom (na základe zlatého štandardu) a RET_MAP je množina TV relácií (TV relácia R patrí do COR_MAP), ktorú vrátila metóda mapovania pri analýze príspevku P. Potom ako *Korektne určenú reláciu* ($R_{correct}$) definujeme takú TV reláciu, ktorá sa pre daný príspevok P nachádza v množine TV relácií zo zlatého štandardu a zároveň sa nachádza v množine TV relácií, ktoré vracia metóda mapovania ako výsledok. Platí teda nasledujúci vzťah:

$$\forall R_{correct} : R_{correct} \cap COR\_MAP \cap RET\_MAP \neq \varnothing \tag{8.1}$$

Na vyhodnotenie priebehu experimentov sme využívali základné metriky z oblasti vyhľadávania informácií:

- **presnosť** (*angl.* precision) - definovaná ako pomer korektne určených (namapovaných) televíznych relácií voči všetkých reláciám, ktoré mapovanie vrátilo pre daný príspevok (pomer vrátených relevantných a vrátených),
- **úplnosť** (*angl.* recall) - definovaná ako pomer korektne určených (namapovaných) televíznych relácií voči všetkým súvisiacim reláciám pre daný príspevok zo zlatého štandardu (pomer vrátených relevantných a relevantných),
- **F1-skóre** - definované ako harmonický priemer presnosti a úplnosti (vzťah 8.2).

$$F1 = \frac{2 * P * R}{P + R} \tag{8.2}$$

kde:

- $F1$ - F1-skóre
- P - presnosť
- R - úplnosť

V rámci vyhodnotenia uvádzame všetky hodnoty uvedených metrík v intervale $< 0, 1 >$. Pri výpočte *presnosti* sme navyše stanovili nasledujúce podmienky:

- ak zlatý štandard pre vyhodnocovaný výsledok obsahuje aspoň 1 relevantnú TV reláciu a mapovanie nevráti žiadnu relevantnú TV reláciu ($COR\_MAP \neq \emptyset$ a zároveň $RET\_MAP = \emptyset$), hodnota presnosti je $P = 0$,

- ak zlatý štandard pre vyhodnocovaný výsledok neobsahuje žiadnu relevantnú TV reláciu a mapovanie nevráti žiadnu TV reláciu ($COR\_MAP = \emptyset$ a zároveň $RET\_MAP = \emptyset$), hodnota presnosti je $P = 1$.

Každú metriku počítame vždy nad celkový výsledkom z jedného príspevku. Výslednú hodnotu metriky získame ako priemernú hodnotu naprieč všetkými čiastkovými hodnotami.

8.3 Výsledky realizovaných experimentov

Vyhodnotenie výsledkov mapovania je užitočný indikátor kvality navrhnutej metódy. Ako sme však uviedli v kapitole *Prepájanie sociálneho a televízneho obsahu*, v rámci našej metódy prichádzame s viacerými vylepšeniami, ktorých cieľom je skvalitniť existujúce prístupy. V nasledujúcich podkapitolách sa preto postupne pozrieme na výsledky experimentov, ktoré sa snažili priniesť pohľad na význam jednotlivých vylepšení, ale zároveň poskytnúť prehľad o celkovej úspešnosti nášho prístupu k mapovaniu.

8.3.1 Prahová hodnota

V kapitole Prepájanie sociálneho a televízneho obsahu sme zaviedli parameter rel_{thres}, ktorý umožňuje z výsledkov mapovania vybrať iba tie, o ktorých spoľahlivo môžeme prehlásiť, že budú korektné. Cieľom experimentov bolo stanoviť hodnotu parametra a tiež zistiť, aký má vplyv na výsledné mapovanie. Stanovenie korektnej hodnoty parametra je kľúčové aj pre dynamickú informačnú bázu, ktorá využíva ako vstup práve korektné výsledky mapovania.

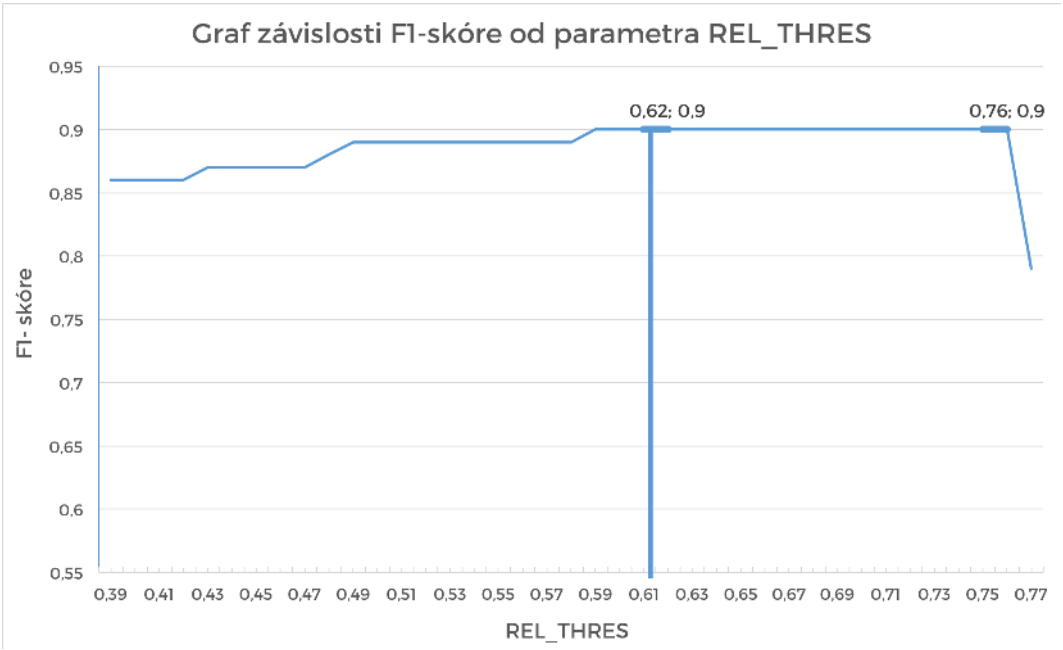
Na obrázku č. 8.1 sa nachádza graf závislosti F1-skóre od hodnoty parametra rel_tres.

Naším cieľom bolo stanoviť takú hodnotu parametra rel_thres takú, aby sme maximalizovali hodnoty skúmaných metrík. Vzhľadom na to, že metrika F1-skóre je harmonický priemer presnosti a úplnosti, snažili sme sa maximalizovať práve F1-skóre. Výslednú ideálnu prahovú hodnotu sme určili zo vzťahu:

$$REL_{THRES} = argmax f(x) \tag{8.3}$$

kde $y = f(x)$ je funkcia závislosti hodnoty F1-skóre (y) od parametra rel_{thres} (x).

Ako môžeme vidieť na obrázku č. 8.1, k maximalizácii F1-skóre dochádza v prípade, že sa hodnota parametra rel_{thres} nachádza v intervale $< 0,59; 0,76 >$. V



Obr. 8.1: Graf F1-skóre pre rôzne prahové hodnoty

tomto intervale dosahuje výsledné F1-skóre hodnotu 0,90. V prípade, že je parameter rel_{thres} nižší ako hodnota 0,59 medzi výsledky mapovania zaraďujeme aj tie televízne relácie, ktoré s daným príspevkom nesúvisia. Naopak, v prípade, že hodnotu parametra zvýšime nad hranicu 0,76, metóda vylúči z výsledkov aj tie televízne relácie, ktoré boli korektne namapované.

Pomocou experimentu sme tiež odhalili, že hodnota rel_{thres} nižšia ako 0,59 nemá negatívny vplyv na úplnosť. Táto skutočnosť iba potvrdzuje, že naša metóda dokáže vrátiť korektné súvisiace relácie aj v prípade, že by sme parameter vôbec nepoužili. Cieľom parametra rel_{thres} je vylúčiť tie relácie, ktoré s istotou vieme prehlásiť za nekorektné. Na základe výsledkov sme ako ideálnu prahovú hodnotu mapovania stanovili $rel_{thres} = 0,70$. Túto hodnotu budeme používať pri vytváraní mapovania vo všetkých nasledujúcich experimentoch[4].

V nasledujúcich podkapitolách sa pozrieme na tri kľúčové vylepšenia, ktoré sme zaviedli za účelom skvalitniť mapovanie: skúmanie externých prepojení, prefixový index a dynamickú informačnú bázu.

[4]Výpočet ideálnej prahovej hodnoty mapovania sme vykonávali nad inou dátovou množinou ako nasledujúce experimenty.

8.3.2 Črty príspevku

V našom riešení uvažujeme tri základné črty príspevku: *pomenované entity*, *hashtagy* a *externé prepojenia*. Existujúce riešenia sa pritom zameriavali iba na prvé dve črty, ktoré sa však často nevyskytovali vo všetkých príspevkoch. Taktiež, aj keď sa v príspevku nachádzali, nemuseli byť pre účely mapovania prínosné.

V kapitole *Analýza obsahu zo sociálnych sietí* sme uviedli výsledky našej analýzy príspevkov. Cieľom nasledujúcich experimentov bolo porovnať, do akej miery vplýva použitie všetkých čŕt na úspešnosť mapovania. Zamerali sme sa na to, aký vplyv na výsledok bude mať, ak použijeme črty jednotlivo a ako sa zmenia hodnoty metrík, ak použijeme všetky tri črty príspevku súčasne. V každom z experimentov sme pracovali s celou testovacou množinou (nevyberali sme cielene iba tie príspevky, ktoré obsahujú danú črtu resp. kombináciu čŕt).

Tabuľka č. 8.2 obsahuje výsledky skúmaných metrík pre prípady, kedy uvažujeme jednotlivé črty príspevku samostatne a v kombinovanej forme. Posledný riadok tabuľky predstavuje úplne nastavenie algoritmu, kedy využívame všetky tri črty súčasne.

Tabuľka 8.2: Porovnanie presnosti pri rôznych nastaveniach mapovania

Nastavenie	**Priemerná presnosť**	**Maximálna presnosť**
Pomenované entity (P)	0,64	0,87
Hashtagy (H)	0,24	0,81
Externé prepojenia (E)	0,51	0,68
P + H	0,75	0,94
P + E	0,86	0,90
H + E	0,66	0,89
Úplné nastavenie	0,92	0,98

Uvedené výsledky potvrdzujú naše úvodné hypotézy. Pridanie externých prepojení (ako ďalšej črty príspevku) dokáže zvýšiť priemernú presnosť mapovania, a to pri všetkých kombináciách. V prípade, že by sme využili výhradne externé prepojenia a neprihliadali na obsah príspevku, dosiahli by sme presnosť až 0,51. Ako si však neskôr ukážeme, zapojenie všetkých čŕt zlepšilo celkovú presnosť natoľko, že sa nám podarilo dosiahnuť porovnateľné alebo lepšie výsledky ako existujúce prístupy.

8.3.3 Prefixový index

Zavedením prefixového indexu sme sa snažili odstrániť potenciálne nekorektné mapovania, a to najmä falošné pozitívne prípady (*angl.* false positive). Ako sme už diskutovali v kapitole *Predspracovanie príspevku*, prefixový index sme využili pri

extrakcií pomenovaných entít.

Na základe našich experimentov sme zistili, že zavedenie prefixového indexu zvýši celkovú presnosť v priemere o **2%**. Pri americkej televízii Fox sme pritom zaznamenali zvýšenie presnosti až o **8%**. Hoci sme sa v rámci našej práce nezamerali na výkon, ukázalo sa, že prefixový index navyše dokáže zvýšiť rýchlosť algoritmu mapovania. Priemerné trvanie štandardného mapovania jedného príspevku sme znížili z pôvodných 40 ms na 20 ms.

Poznámka: Podrobné informácie o vyhodnotení rýchlosti mapovania uvádzame v prílohe *Príloha k overeniu riešenia*.

8.3.4 Ko-učenie

Použitím dynamických televíznych metadát sme sa snažili nájsť mapovanie v prípadoch, kedy štandardný prístup nevracal žiadne výsledky. Vychádzali sme z predpokladu, že takáto situácia môže nastať v troch prípadoch:

1. Príspevok zo sociálnej siete neobsahuje žiadne pomenované entity, hashtagy alebo externé odkazy (jadro štandardného prístupu k mapovaniu).
2. Štandardné mapovanie vráti výsledok, ale tento výsledok z hľadiska koeficienta súvislosti nepresiahne prahovú hodnotu rel_{thres} (dôsledkom čoho je výsledok vylúčený).
3. Mapovanie zlyhá z dôvodu zlého návrhu metódy.

Poznámka: V predchádzajúcich experimentoch sme testovali príspevky v poradí, v akom boli publikované na sociálnej sieti. Toto poradie však môže mať vplyv na kvalitu dynamických televíznych metadát. Preto sme sa rozhodli pre účely skvalitnenia výsledkov experimentu **zamiešať poradie príspevkov**. Experimenty so zamiešaným poradím sme vykonali 10x po sebe a výsledne hodnoty jednotlivých metrík spriemerovali.

Pri vyhodnocovaní sme sa snažili zamerať na mieru prínosu dynamických televíznych metadát k štandardnému mapovaniu. Tabuľka 8.3 obsahuje výsledné hodnoty pre skúmané metriky bez použitia dynamickej informačnej bázy a s použitím dynamickej informačnej bázy. Hoci zlepšenie nepozorujeme pri všetkých televíziách, pri niektorých môžeme hovoriť o veľmi pozoruhodných výsledkoch. Ako môžeme vidieť v tabuľke č. 8.3, významné zlepšenie nastalo pri televízii **CBS**. Štandardné mapovanie v tomto prípade zlyhalo v 12 príspevkoch. Po zavedení dynamických televíznych metadát sa podarilo korektne namapovať až 9 príspevkov, čo malo za následok aj zvýšenie presnosť až o **12%**. Dôležité je taktiež poznamenať, že zavedenie dynamickej bázy znalostí nemalo za následok nárast falošných pozitívnych prípadov

mapovania.

Tabuľka 8.3: Výsledky overenia dynamických televíznych metadát (DTM)

TV stanica	**Presnosť bez DTM**	**Presnosť s DTM**	**Zlepšenie**
CBS	0,85	0,97	**+ 12 %**
Fox	0,87	0,89	+ 2 %
NBC	0,92	0,92	0 %

Pri manuálnej analýze výsledkov sme objavili zaujímavý prípad, kde sa preukázateľne prejavila sila dynamickej informačnej bázy. Takýmto prípadom bolo správanie americkej komerčnej televízie CBS. Pri príležitosti odovzdávania hudobných cien *Academy of Country Music Awards* prebiehalo na sociálnej sieti komentovanie celého galavečera formou príspevkov na sociálnej sieti. V texte príspevkov televízia používala celý názov TV relácie (*Academy of American Country Music Awards*), ale aj skrátený názov *ACMs*. V tomto prípade zafungovalo štandardné mapovanie, ktoré sa opieralo o pomenované entity. V ďalších príspevkoch, v ktorých postupne vyhlasovala jednotlivých víťazov súťaže už používala iba skratku *ACMs*, ktorá by sa však v našej informačnej báze za normálnych okolností nenachádzala. Práve vďaka použitiu dynamickej informačnej bázy sme mali k dispozícii obsah príspevkov, pri ktorých postačovalo štandardné mapovanie a na základe toho sme vedeli korektne namapovať aj ďalšie príspevky.

8.3.5 Úplne nastavenie

V predchádzajúcich podkapitolách sme sa zamerali na overenie riešenia z hľadiska rôzneho nastavenia metódy. Skúmali sme tiež, aký vplyv má použitie vylepšení, s ktorými v rámci našej práce prichádzame. V nasledujúcich riadkoch si zhrnieme získané výsledky a budeme analyzovať, v ktorých prípadoch existuje priestor na ďalšie zlepšenie nášho riešenia.

Tabuľka č. 8.4 obsahuje výsledky jednotlivých metrík pri úplnom nastavení metódy. Pri názvoch televízií uvádzame v zátvorke jazykovú verziu príspevkov (EN = po anglicky, SK = po slovensky). Ako môžeme vidieť, hodnoty presnosti a úplnosti sú si relatívne blízke. Dôvodom bol nízky počet televíznych relácií, ktoré súviseli s príspevkom. Túto skutočnosť pripisujeme obsahu príspevkov, keďže tvorcovia obsahu (televízie) sa snažili príspevkom cieliť na 1 konkrétnu televíznu reláciu.

Ak sa bližšie pozrieme na **falošné negatívne prípady** (prípad, kedy algoritmus vrátil nekorektné mapovanie), tak spomedzi celkovo 734 namapovaných TV relácií bolo 21 namapovaných nekorektne (približne 3%). Okrem prípadov, kedy sa výsledok mapovania nenachádzal v nami vytvorenom zlatom štandarde, sme sa

Tabuľka 8.4: Výsledky overenia pri záverečnom nastavení mapovania

TV stanica	**Presnosť**	**Úplnosť**	**F1-skóre**
CBS (EN)	0,85	0,85	0,85
NBC (EN)	0,92	0,94	0,93
Fox (EN)	0,87	0,94	0,89
The CW (EN)	0,98	0,97	0,97
Markíza (SK)	0,97	0,97	0,97
JOJ (SK)	0,93	0,95	0,94
Priemer	**0,92**	**0,94**	**0,93**

tiež stretli so situáciou, kedy zlatý štandard obsahoval mapovanie, ale naša metóda mapovania neurčila žiadne vhodné súvisiace televízne relácie (5,87 % spomedzi všetkých príspevkov). Hoci takýto výsledok nemôžeme považovať za správny, v širšom kontexte možno považovať prázdnu množinu za výhodnejší prípad ako množinu obsahujúcu nekorektné mapovania.

Samostatne sme sa pozreli na **príspevky**, ktorým neboli v zlatom štandarde **priradené žiadne TV relácie**. Takéto prípady zahŕňali napríklad odkazy na spravodajské servery, kde obsah správ nesúvisel so žiadnou TV reláciou. V rámci vyhodnotenia sme takéto príspevky neuvažovali – predstavovali 6% z celej získanej dátovej množiny. Pri aplikovaní nášho mapovacieho prístupu vrátil algoritmus korektne na výstup prázdnu množinu v 81% prípadov.

Pri porovnávaní sa s existujúcimi prácami (najmä [9]) sme sa stretli s alternatívnou definíciou presnosti a úplnosti, ktorá uvažuje jednotlivé televízne relácie ako samostatné triedy. Všetky príspevky, ktoré súvisia s danou televíznou reláciou patria do rovnakej triedy T. Skúmané metriky potom definujeme nasledovne:

- **presnosť@T** - podiel počtu príspevkov, ktoré algoritmus korektne namapoval na triedu T a počtu všetkých príspevkov, ktoré algoritmus namapoval do triedy T,
- **úplnosť@T** - podiel počtu príspevkov, ktoré algoritmus korektne namapoval na triedu T a počtu príspevkov, ktoré podľa zlatého štandardu patria do triedy T.

Tabuľka č. 8.5 uvádza výsledné hodnoty metrík pre všetky nami analyzované televízne stanice. Uvedené výsledky možno považovať za vhodnejšiu reprezentáciu výsledkov v prípade, že zaraďujeme metódu mapovania do oblasti klasifikácie. Pri porovnaní s tabuľkou č. 8.4 môžeme preto pozorovať nárast presnosti ale pokles úplnosti, čo reflektuje definíciu metrík.

V našej ďalšej práci sa zameriame na skvalitnenie výberu dát, pričom sa budeme snažiť prihliadať aj na vyváženie tried. To sa v súčasných výsledkoch negatívne

Tabuľka 8.5: Výsledky overenia pri záverečnom nastavení mapovania

TV stanica	**Presnosť@T**	**Úplnosť@T**	**F1-skóre@T**
CBS (EN)	1,00	0,82	0,91
NBC (EN)	0,89	0,85	0,87
Fox (EN)	0,92	0,95	0,93
The CW (EN)	1,00	0,88	0,93
Markíza (SK)	0,98	0,77	0,86
JOJ (SK)	0,87	0,59	0,70
Priemer	**0,94**	**0,81**	**0,87**

prejavilo v značných rozdieloch medzi úplnosťou@T slovenských a amerických televíznych staníc (hlavne pri televíznej stanici JOJ, ktorá dosiahla pomerne nízku hodnotu úplnosti@T).

8.4 Porovnanie s existujúcimi prístupmi

V úvode sme sa venovali prístupom príbuzným nášmu riešeniu. Teraz si zosumarizujeme vybrané prístupy a pozrieme sa na to, aké výsledky sa im podarilo dosiahnuť.

Yerva a kol. vo svojej práci [33] klasifikovali príspevky z Twittera s cieľom priradiť ich k rôznym komerčným spoločnostiam. V experimentoch sa zamerali na porovnanie viacerých klasifikátorov, ktoré navrhli. Najvyššie skóre úspešnosti (*angl.* accuracy) pritom dosiahli s klasifikátorom Basic Profile-BP2 (BPRA1), a to 0,84.

Viacerým typom pomenovaných entít sa vo svojej práci [16] venovali Gattani a kol., ktorí sa vo svojom riešení snažili zahrnúť sociálny kontext. Pri mapovaní filmov na tweety dosiahli presnosť 1,00 (100%) a pri mapovaní televíznych relácií na tweety presnosť 0,50.

Cremonesi a kol. využívali v [9] taktiež sociálnu sieť Twitter, pričom sa zamerali na televízne relácie. Na vytvorenie mapovania využívali pravidlá aplikované na text príspevkov, ale taktiež aj strojové učenie. Pri optimálnom nastavení svojej metódy dosiahli hodnotu presnosti 0,92 a úplnosti 0,65.

Shen a kol. mapovali v [30] pomenované entity rôznych typov pre účely identifikácie tém v tweetoch. Ich metóda dosiahla výsledné skóre úspešnosti 0,85.

V tabuľke č. 8.6 sa nachádza prehľad výsledkov sumarizovaných prístupov. Čísla v stĺpci *Presnosť* označené hviezdičkou (*) predstavujú hodnoty metriky *úspešnosť* (*angl.* accuracy), keďže autori vo svojej práci metriku *presnosť* nepoužívali. Naše riešenie dosahuje porovnateľné výsledky z hľadiska presnosti, pričom v niektorých prípadoch (práce Gattahi a kol.) sú tieto výsledky výrazne lepšie. Priame porovnanie však nie je úplne korektné aj z toho dôvodu, že existujúce riešenia sa

Tabuľka 8.6: Výsledky overenia pri záverečnom nastavení mapovania

Prístup	**Sociálna sieť**	**Typ entít**	**Presnosť**
Yerva ([33])	Twitter	spoločnosť	0,84 *
Cremonesi ([9])	Twitter	film / televízna relácia	0,92
Gattani ([16])	Twitter	film	1,00
Gattani ([16])	Twitter	televízna relácia	0,50
Shen ([30])	Twitter	rôzne	0,85 *
Náš prístup	Facebook	film / televízna relácia	**0,94**

zameriavajú na sociálnu sieť Twitter a anglický jazyk. Pri návrhu nášho riešenia sme síce vychádzali z vlastností sociálnej siete Facebook, avšak toto riešenie nie je žiadnym spôsobom na ňu viazané. Zároveň sme navrhli metódu, ktorá je jazykovo nezávislá a nevyžaduje predspracovanie obsahu z rozsiahlych znalostných báz, ktoré často nie sú dostatočne rýchlo aktualizované v porovnaní s televíznymi programom.

8.5 Zhrnutie

Navrhnutú metódu mapovania sme kvantitatívne overili na anotovanej vzorke príspevkov zo sociálnej siete. Pomocou prezentovaných výsledkov sa nám podarilo preukázať skvalitnenie výsledkov mapovania, a to pre rôzne dátové množiny. Toto skvalitnenie sme pritom dosiahli pomocou nami zavedených vlastností (prefixový index, externé prepojenia), s ktorými sme sa v predchádzajúcich prácach nestretli. K skvalitneniu pritom nedošlo len z pohľadu skúmaných metrík, ale taktiež aj z pohľadu výpočtovej náročnosti mapovania. V neposlednom rade sa nám v prvotnej fáze návrhu podarilo overiť aj význam ko-učenia pri tvorbe dynamických televíznych metadát, pričom sme dosiahli ďalšie zlepšenie voči štandardnému prístupu k mapovaniu.

9. Záver

V súčasnosti sa klasické televízne vysielanie čoraz viac prepája s aktivitou používateľov na sociálnych sieťach. S neustále pribúdajúcim množstvom obsahu stúpa potreba jeho analyzovania. Medzi najdôležitejšie úlohy patrí aj identifikácia, k akým entitám z reálneho sveta môžeme sociálny obsah priradiť. Mnohí výskumníci sa preto snažia vytvoriť automatizované metódy, ktoré by vedeli odhaliť, k akej televíznej relácii môžeme zaradiť jednotlivé príspevky zo sociálnej siete. Využívajú pri tom rôzne techniky, akými sú napríklad extrakcia pomenovaných entít alebo identifikácia kľúčových slov.

V našej práci sme sa venovali doméne prepájania multimediálneho obsahu s mikroblogovou sieťou. Zamerali sme sa na špeciálny druh mikroblogovej siete, ktorý sa vyznačuje vyššou mierou dynamiky – sociálna sieť. Na problém sme sa pozreli z viacerých uhlov pohľadu – od spracovania textu, po vyhľadávanie pomenovaných entít či tém až po samotné prepájanie znalostných báz a obsahu zo sociálnych sietí. Pri skúmaní existujúcich riešení sme identifikovali nedostatky, ktoré sme sa pri návrhu našej vlastne metódy snažili odstrániť. Vykonali sme podrobnú analýzu správania tvorcov sociálneho obsahu, na základe ktorej sme identifikovali tri kľúčové prvky, ktoré sa vyskytujú v príspevkoch na sociálnej sieti:

- pomenované entity,
- hashtagy,
- externé prepojenia.

Hlavným výstupom našej práce je metóda, ktorá umožňuje využiť názvy televíznych relácií na mapovanie príspevkov zo sociálnej siete na televízny program. Existujúce prístupy sme obohatili o nový prvok, ktorý sa v predchádzajúcich prácach v danej doméne nevyskytoval – spracovanie externého prepojenia. Navrhli sme metódu mapovania, ktorá je jazykovo nezávislá, dokáže pracovať bez potreby anotovaných trénovacích dát a jej výsledky sú k dispozícii v reálnom čase.

Pre overenie nášho prístupu sme vykonali viacero experimentov, v ktorých sme sa snažili odhaliť, aký vplyv na výstup mapovania bude mať zavedenie našich vylepšení. Manuálne sme anotovali viac ako 700 príspevkov zo sociálnej siete z profilov 2 slovenských a 4 amerických televíznych staníc. Výsledkami experimentov sme preukázali, že ak pri spracovávaní príspevku zo sociálnej siete budeme okrem pomenovaných entít a hashtagov uvažovať aj externé prepojenie, zlepšíme celkovú presnosť mapovania a získame tak ďalšie relevantné výsledky. Zároveň sme vykonali overenie nad dátami zo sociálnej siete Facebook, s ktorou sme sa v existujúcich

prácach stretávali iba zriedka.

Otvoreným problémom naďalej zostávajú príspevky, ktoré neobsahujú žiadnu z nami identifikovaných čŕt a algoritmus štandardného mapovania tak nevie vyhodnotiť prepojenie na televízne relácie. Ako riešenie sme navrhli použiť prístup ko-učenia, ktorý sme už aj v prvej fáze overili. Výsledky naznačujú, že v ko-učení sa skrýva potenciál a v niektorých prípadoch dokáže značne zlepšiť presnosť výsledného mapovania.

V našej ďalšej práci sa budeme popri ko-učení zameriavať aj na širšie overenie nášho riešenia. Budeme sa snažiť overiť väčšie množstvo televíznych staníc a televíznych relácií. Taktiež by sme sa radi zamerali na kvalitnejšie určenie atraktívnosti jednotlivých príspevkov, kedy aj z pohľadu používateľa nemusia mať všetky príspevky danej televíznej relácie rovnaký prínos.

Popri samotnom mapovaní sme pri analýze existujúcich riešení načrtli aj ďalšiu zaujímavú oblasť – extrakcia metadát. Z namapovaných príspevkov sociálnej siete totiž môžeme získavať ďalšie zaujímavé informácie, akými sú napríklad vývoj sledovanosti, obľúbenosť televíznej relácie alebo sentiment, ktorý sa skrýva najmä v komentároch používateľov sociálnej siete. Ako sme diskutovali, tieto informácie môžu obohatiť klasické - statické metadáta z televízneho programu o nové - dynamické prvky, ktoré zvýšia jeho atraktívnosť.

Literatúra

[1] Aisopos, F., Papadakis, G., and Varvarigou, T. Sentiment analysis of social media content using n-gram graphs. In *Proceedings of the 3rd ACM SIGMM International Workshop on Social Media* (New York, NY, USA, 2011), WSM '11, ACM, pp. 9–14.

[2] Alexa. Top Sites in Slovakia. `http://www.alexa.com/topsites/countries/SK`. [Online; videné 11. 5. 2016].

[3] Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *the Journal of machine Learning research 3* (2003), 993–1022.

[4] Blink. Televízia chytá druhý dych. Vďaka internetu. `http://strategie.hnonline.sk/blogy/televizia-chyta-druhy-dych-vdaka-internetu`, 2016. [Online; videné 10. 5. 2016].

[5] Bravo-Marquez, F., Mendoza, M., and Poblete, B. Combining strengths, emotions and polarities for boosting twitter sentiment analysis. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining* (New York, NY, USA, 2013), WISDOM '13, ACM, pp. 2:1–2:9.

[6] Chapelle, O., Schlkopf, B., and Zien, A. *Semi-Supervised Learning*, 1st ed. The MIT Press, 2010.

[7] Cheng, Y.-H., Wu, C.-M., Ku, T., and Chen, G.-D. A predicting model of tv audience rating based on the facebook. In *Social Computing (SocialCom), 2013 International Conference on* (Sept 2013), pp. 1034–1037.

[8] Chung, E., Hwang, Y.-G., and Jang, M.-G. Korean named entity recognition using hmm and cotraining model. In *Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages - Volume 11* (Stroudsburg, PA, USA, 2003), AsianIR '03, Association for Computational Linguistics, pp. 161–167.

[9] Cremonesi, P., Pagano, R., Pasquali, S., and Turrin, R. Tv program detection in tweets. In *Proceedings of the 11th European Conference on Interactive TV and Video* (New York, NY, USA, 2013), EuroITV '13, ACM, pp. 45–54.

[10] Czwitkovics, T. Sledovanosť televízií cez peoplemetre bude naďalej merať TNS. `http://medialne.etrend.sk/televizia/sledovanost-televizii-cez-peoplemetre-bude-nadalej-merat-tns.`

`html`. [Online; videné 10. 5. 2015].

[11] Davidov, D., Tsur, O., and Rappoport, A. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (Stroudsburg, PA, USA, 2010), COLING '10, Association for Computational Linguistics, pp. 241–249.

[12] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE 41*, 6 (1990), 391–407.

[13] eBizMBA. Top 15 Most Popular Social Networking Sites — May 2015. `http://www.ebizmba.com/articles/social-networking-websites`, 2015. [Online; videné 10. 5. 2015].

[14] Ericsson ConsumerLab. TV and Media - Identifying the needs of tomorrow's video consumers.

[15] Garabík, R., Gianitsová, L., Horák, A., and Šimková, M. Tokenizácia, lematizácia a morfologická anotácia slovenského národného korpusu.

[16] Gattani, A., Lamba, D. S., Garera, N., Tiwari, M., Chai, X., Das, S., Subramaniam, S., Rajaraman, A., Harinarayan, V., and Doan, A. Entity extraction, linking, classification, and tagging for social media: A wikipedia-based approach. *Proc. VLDB Endow. 6*, 11 (Aug. 2013), 1126–1137.

[17] Giannakopoulos, G., Karkaletsis, V., Vouros, G., and Stamatopoulos, P. Summarization system evaluation revisited: N-gram graphs. *ACM Trans. Speech Lang. Process. Vol. 5*, 3 (Oct. 2008), 5:1–5:39.

[18] Hofmann, T. Probabilistic latent semantic indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 1999), SIGIR '99, ACM, pp. 50–57.

[19] Hua, W., Zheng, K., and Zhou, X. Microblog entity linking with social temporal context. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data* (New York, NY, USA, 2015), SIGMOD '15, ACM, pp. 1761–1775.

[20] Joachims, T. Text categorization with suport vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning* (London, UK, UK, 1998), ECML '98, Springer-Verlag, pp. 137–142.

[21] Kaššák, O., Kompan, M., and Bieliková, M. Extrakcia pomenovaných

entít pre slovenský jazyk. In *ZNALOSTI 2012: Sborník příspěvků 11 ročníku konference, Mikulov, hotel Eliška 14.-16. 10. 2012* (Praha, 2012), Matfyzpress, pp. 52–61.

[22] Kondrak, G. N-gram similarity and distance. In *String Processing and Information Retrieval*, M. Consens and G. Navarro, Eds., vol. 3772 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2005, pp. 115–126.

[23] Li, C., Weng, J., He, Q., Yao, Y., Datta, A., Sun, A., and Lee, B.-S. Twiner: Named entity recognition in targeted twitter stream. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2012), SIGIR '12, ACM, pp. 721–730.

[24] Li, Q. Searching for entities: When retrieval meets extraction, January 2012.

[25] Meng, X., Wei, F., Liu, X., Zhou, M., Li, S., and Wang, H. Entity-centric topic-oriented opinion summarization in twitter. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2012), KDD '12, ACM, pp. 379–387.

[26] Mohammad, S. M., and Turney, P. D. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence 29*, 3 (2013), 436–465.

[27] Plutchik, R. The nature of emotions. *American Scientist 89* (2001), 344.

[28] Ritter, A., Clark, S., Mausam, and Etzioni, O. Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Stroudsburg, PA, USA, 2011), EMNLP '11, Association for Computational Linguistics, pp. 1524–1534.

[29] Salton, G., and McGill, M. J. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.

[30] Shen, W., Wang, J., Luo, P., and Wang, M. Linking named entities in tweets with knowledge base via user interest modeling. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2013), KDD '13, ACM, pp. 68–76.

[31] Wakamiya, S., Lee, R., and Sumiya, K. Towards better tv viewing rates: Exploiting crowd's media life logs over twitter for tv rating. In *Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication* (New York, NY, USA, 2011), ICUIMC '11, ACM, pp. 39:1–39:10.

[32] Xu, Z., Jin, R., Huang, K., Lyu, M. R., and King, I. Semi-supervised text categorization by active search. In *Proceedings of the 17th ACM Conference on*

Information and Knowledge Management (New York, NY, USA, 2008), CIKM '08, ACM, pp. 1517–1518.

[33] Yerva, S. R., Miklós, Z., and Aberer, K. What have fruits to do with technology?: The case of orange, blackberry and apple. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics* (New York, NY, USA, 2011), WIMS '11, ACM, pp. 48:1–48:10.

[34] Zhang, C., Zhao, S., and Wang, H. Bootstrapping large-scale named entities using url-text hybrid patterns. In *IJCNLP* (2013), Citeseer, pp. 293–301.

A. Technická dokumentácia

V nasledujúcich riadkoch dokumentujeme vybrané vlastnosti našej metódy z pohľadu technickej dokumentácie. Zamerali sme sa pritom na dátový model, opis implementácie a overenie riešenia.

A.1 Štruktúra dát príspevkov sociálnej siete

Nasledujúci prehľad dokumentuje štruktúru dát, ktoré sme získavali zo sociálnej siete Facebook. Táto štruktúra vychádza z reprezentácie dát vo formáte JSON. Uvedenú štruktúru dát neposkytuje priamo služba Facebook, ale zostavili sme ju implementáciou modulu, ktorý predspracoval surové dáta získané z Facebooku cez Graph API.

A.1.1 Post

Predstavuje entitu príspevku na sociálnej sieti Facebook.

- *id* - jedinečný číselný identifikátor príspevku,
- *from* - jedinečný číselný identifikátor autora príspevku (stránka),
- *author_type* - určuje typ autor príspevku. Povolené hodnoty:
  - *page* - stránka, na ktorej sa príspevok nachádza,
  - *owner* - osoba alebo alebo iná stránka,
- *message* - textová správa príspevku,
- *picture* - odkaz na obrázok príspevku (ak sa nachádza v príspevku zdieľaný odkaz, ide o obrázok extrahovaný z tohto odkazu),
- *hashtags* - pole hashtagov extrahovaných z textovej správy príspevku,
- *link* - pole obsahujúce zoznam odkazov. Každý odkaz má nasledujúcu štruktúru:
  - *caption* - titulok zdieľaného odkazu,
  - *description* - popis zdieľaného odkazu,
  - *name* - názov zdieľaného odkazu,
  - *url* - adresa zdieľaného odkazu,
- *likes_count* - počet označení „Páči sa mi to“ na príspevok,

- *shares_count* - počet zdieľaní príspevku,
- *timestamp* - časová pečiatka vytvorenia príspevku,
- *type* - druh príspevku (možnosti: link, status, photo, video, offer),
- *source* - adresa na video pripojené k príspevku,
- *attachments* - zoznam priložených obrázkov, odkazov a videí. Každá príloha má nasledujúcu štruktúru:
  - *description* - popis prílohy,
  - *title* - titulok prílohy,
  - *type* - druh prílohy,
  - *media_image_url* - URL adresa obrázku (ak ide o obrázok),
  - *media_image_width* - šírka obrázka (ak ide o obrázok),
  - *media_image_height* - výška obrázka (ak ide o obrázok),
  - *subattachments* - podprílohy (ak ide o fotogalériu),
- *to* - zoznam profilov, ktoré sú označené v príspevku:
  - *id* - identifikátor profilu,
  - *category* - kategória profilu (ak ide o Page),
  - *name* - meno profilu,
- *comments* - zoznam komentárov (typ *Comment*, viď nižšie).

A.1.2 Comment

Predstavuje entitu komentára k príspevku na sociálnej sieti Facebook.

- *parent* - identifikátor rodičovského príspevku,
- *from* - identifikátor autora komentára,
- *message* - obsah komentára,
- *like_count* - počet označení „Páči sa mi to“ na komentár,
- *timestamp* - časová pečiatka vytvorenia komentára,
- *comment_count* - počet odpovedí (reakcií) na komentár,
- *link* - pole obsahujúce zoznam odkazov. Každý odkaz má nasledujúcu štruktúru:
  - *caption* - titulok zdieľaného odkazu,

  - *description* - popis zdieľaného odkazu,
  - *name* - názov zdieľaného odkazu,
  - *url* - adresa zdieľaného odkazu,
- *to* - zoznam profilov, ktoré sú označené v komentári. Každý profil má nasledujúcu štruktúru:
  - *id* - jedinečný číselný identifikátor profilu,
  - *category* - kategória profilu (ak ide o typ Page),
  - *name* - meno profilu,
  - *type* - typ profilu,
  - *length* - dĺžka označenia (v zmysle, fyzická dĺžka textu označenia v komentári),
  - *offset* - počet znakov v texte komentára nachádzajúcich sa pred prvým znakom označenia.

A.2 Štruktúra dát televízneho programu

Nasledujúci prehľad dokumentuje štruktúru dát, ktoré sme získavali z dostupného API na získavanie TV programu. Táto štruktúra vychádza z reprezentácie dát vo formáte XML.

Na rozdiel od dát získavaných z Facebooku, pri TV programe sme dáta žiadnym spôsobom nepredspracovávali.

A.2.1 Channel

Predstavuje entitu TV stanice.

- *id* - jedinečný číselný identifikátor TV stanice,
- *name* - názov TV stanice,
- *programmes* - zoznam relácií vysielaných danou TV stanicou (viď. entitu *Programme*)

A.2.2 Programme

Predstavuje entitu TV relácie vysielanej na konkrétnej TV stanici.

- *id* - jedinečný číselný identifikátor relácie[1],
- *name* - názov relácie,
- *start_time* - časová pečiatka začiatku vysielania,
- *end_time* - časová pečiatka konca vysielania,
- *images* - zoznam URL adries na externé obrázky,
- *type* - typ TV relácie (napr. film, seriál, a pod.),
- *short_description* - krátky opis,
- *long_description* - dlhý opis,
- *original_name* - originálny názov relácie (napr. cudzojazyčný),
- *year* - rok výroby,
- *actors* - zoznam mien hercov, ktorí v relácii vystupujú,
- *directors* - zoznam mien režisérov,
- *producers* - zoznam mien producentov,
- *scriptwriters* - zoznam mien scenáristov,
- *votes* - percentuálne vyjadrenie hodnotenia,
- *country* - kód krajiny pôvodu,
- *length* - dĺžka vysielania v minútach,
- *video* - URL na externé video (napr. *angl.* trailer),
- *motto* - uvádzacia veta charakterizujúca TV reláciu.

A.3 Detaily implementácie

Jadro našej metódy tvorí mapovač, ktorý sme z hľadiska implementácie členili do menných priestorov (*angl.* namespace) podľa funkcionality. Menné priestory reflektujú aj usporiadanie zdrojového kódu v priečinkoch (diskutované menné priestory nájdeme v zdrojovom kóde v priečinku */lib*).

- *ContentTools* - nástroje pracujúce s obsahom príspevku (generovanie hashtagov, extrahovanie pomenovaných entít, extrahovanie hashtagov, spracovanie externých odkazov),
- *DynamicKnowledgeBase* - vytváranie a práca s dynamickými televíznymi metadátami (pre účely ko-učenia),

[1]Pri reláciách, ktoré sa vo vysielaní opakujú (napr. reprízy alebo seriály) sa táto hodnota líši.

- *Evaluation* - implementácia štandardných metrík vyhodnocovania v oblasti vyhľadávania informácií (presnosť, presnosť@k, úplnosť, F1-skóre, priemerná presnosť, NDCG@k),
- *KnowledgeBase* - triedy umožňujúce vytvorenie, načítanie a aktualizáciu bázy televíznych metadát, načítanie a spracovanie TV programu slovenských televíznych staníc,
- *MachineLearning* - implementácia modelu TF-IDF spoločne s pomocnými metódami (výpočet kosínovej podobnosti),
- *Mapper* - implementácia štandardného mapovania (spolupracuje s triedami z iných menných priestorov),
- *MetadataHelper* - rôzne pomocné metódy pri práci s prirodzeným jazykom (identifikácia čísla série v názve programu, odstránenie čísla série, identifikácia rímskych čísiel v texte),
- *NLP* - (skratka z Natural Language Processing) rôzne metódy pracujúce s prirodzeným jazykom (zahŕňa lematizáciu, tokenizáciu, výpočet n-gramovej podobnosti, odstraňovanie rímskych čísiel z textu).

Súčasťou jadra implementácie je aj trieda *URLExpander*, ktorá slúži na expanziu skrátených URL adries. Na vytváranie skrátených URL adries slúžia rôzne komerčné riešenia, ktorá skracujú dlhé adresy (napr. http://www.fiit.stuba.sk) na ich skrátené ekvivalenty (napr. http://bit.ly/1WDf9VE). Takto skrátené adresy sa často vyskytujú v príspevkoch, čo je však pre účely nášho mapovača nežiadúce. V rámci implementácie komunikuje trieda *URLExpander* s externou službou Link Expander[2], ktorá zo skrátených URL získava originálne adresy.

Podrobnosti implementácie jednotlivých tried a príslušných metód dokumentujeme v priamo v zdrojovom kóde vo forme komentárov.

Poznámka: *V implementácii našej metódy sme nepoužili všetky uvedené funkcionality. Vo fáze návrhu sme ich však uvažovali, a preto sme ich implementovali.*

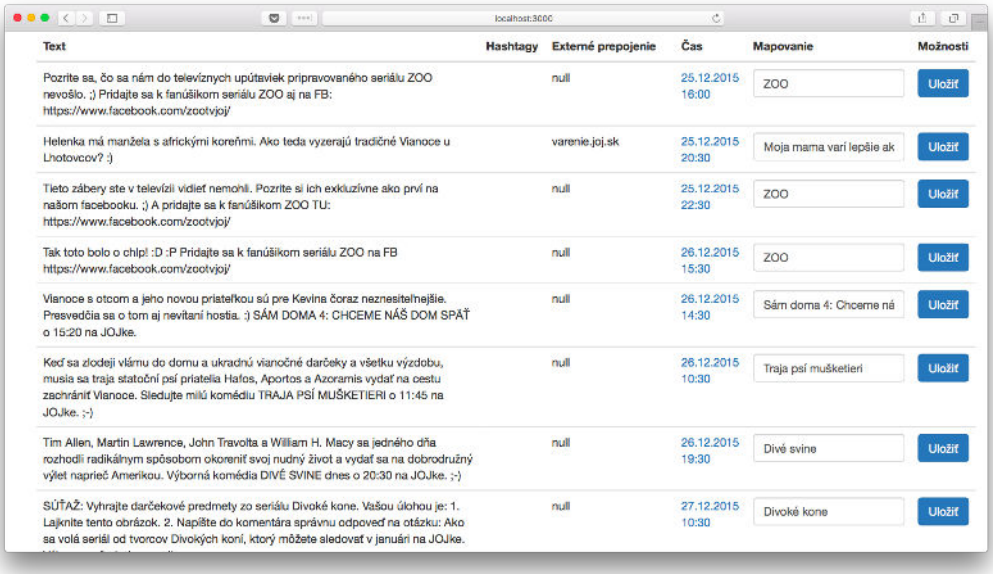
A.3.1 Realizácia mapovania

Nasledujúca ukážka zdrojového v jazyku Ruby (obrázok A.1) reflektuje implementáciu štandardného mapovania. Môžeme vidieť prechod cez všetkých kandidátov, výpočet jednotlivých podobností a záverečný výpočet priemernej podobnosti.

[2] www.linkexpander.com

```
results_similarities = {}

results.each do |result|

  result = result.join(' ') if result.is_a?(Array)
  next unless origin.has_key?(result[0])

  similarities = {  }
  origin[result[0]].each do |candidate, programmes|
    case mode
      when :ngrams

        similarities[candidate] = { programmes: programmes,
                                    similarity: calculate_similarity(candidate, result) }

      when :binary

        similarity = (Levenshtein.distance(candidate, result) <= (strict ? 0 : 1) ? 1.0 : 0.0)
        similarities[candidate] = { programmes: programmes, similarity: similarity }

      when :semibinary

        # Tu staci, aby bola aspon 50% podobnost, preto navysime umelo skore o rozdiel voci rel-thres.
        if (result.start_with?(candidate) || candidate.start_with?(result))
          similarity = calculate_similarity(candidate, result) + (@rel_thres - 0.50)
        else
          similarity = 0.0
        end

        similarities[candidate] = { programmes: programmes, similarity: similarity }

    end
  end

  # Pre kazdeho kandidata moze byt viac programov.
  similarities.each do |candidate, res|
    res[:programmes].each do |programme|

      # Vyberame iba toho kandidata, ktory ma rovnaku dlzku ako nazov TV relacie.
      if (!candidate_entities ||
          (candidate.split(/\W+/).size == get_best_length(programme) &&
              result.split(/\W+/).size == get_best_length(programme)))

        (results_similarities[programme] ||= []) << res[:similarity]

      end

    end
  end
```

Obr. A.1: Ukážka zdrojového kódu mapovania



Obr. A.2: Ukážka mapovača

A.3.2 Anotovanie dátovej množiny

Súčasťou nášho riešenie je aj Rails aplikácia, ktorá slúžila na vytvorenie zlatého štandardu. Prostredníctvom webového rozhrania bolo možné k jednotlivým príspevkom zo sociálnej siete priradiť súvisiace televízne relácie.

Rozhranie aplikácie bolo jednoduché a intuitívne a pre anotátora ponúkalo nasledujúce informácie o príspevku:

- text príspevku zo sociálnej siete,
- extrahované hashtagy,
- odkaz na zdieľaný obsah,
- odkaz na príspevok na sociálnej sieti.

Na obrázku č. A.2 vidíme ukážku z webového rozhrania anotovacej aplikácie.

B. Inštalácia a návod na použitie

B.1 Vytvorenie bázy televíznych metadát

Nami implementovaná metóda mapovania vychádza z bázy televíznych metadát. Bázu televíznych metadát vytvárame zo zoznamu televíznych relácií pre príslušné televízie.

V koreňovom adresári programu otvoríme súbor *programmes.yml*, ktorý obsahuje už pripravené vybrané televízne stanice a televízne programy. Tento súbor dodržiava pravidlá formátovacieho jazyka YAML[1]. Pre účely vytvorenia bázy televíznych metadát je potrebné dodržať nasledujúce pravidlá:

- *kľúč* je názov televíznej stanice (ak je názov viacslovný, je potrebné ho ohraničiť úvodzovkami),
- *hodnota* je zoznam (pole) televíznych relácií vysielaných na danej televíznej stanici uvedenej ako kľúč.

V ďalšom kroku sa pozrieme na to, ako nainštalovať mapovač a vygenerovať bázu televíznych metadát.

B.2 Inštalácia

Pred spustením programu si, prosím, overte, či sú na vašom zariadení nainštalované nasledujúce nástroje:

- Ruby verzie 2.3.0 a vyššej (viac informácií o inštalácii získate na oficiálnej stránke jazyka Ruby[2]),
- Bundler verzie 1.11 a vyššej (viac informácií o inštalácii získate na oficiálnej stránke nástroja Bundler[3]).

Po nainštalovaní všetkých potrebných nástrojov sa presunieme do adresára, v ktorom sa nachádza program. V konzole spustíme príkaz:

./install

Týmto príkazom sa nainštaluje všetko potrebné pre spustenie iniciálneho mapovania. Zároveň sa vygeneruje báza televíznych metadát z televízií, ktoré sme si pripravili v

[1] http://yaml.org/
[2] https://www.ruby-lang.org/
[3] http://bundler.io/

kapitole *Vytvorenie bázy televíznych metadát.*

V prípade, že zmeníme obsah súboru *programmes.yml*, je potrebné spustiť príkaz (v koreňovom adresári programu):

./update

B.3 Realizácia mapovania

Samotné mapovanie je možné realizovať prostredníctvom konzolových príkazov, pričom môžeme uviesť viaceré prepínače, pomocou ktorých zmeníme parametre mapovania.

Pred spustením mapovania si musíme pripraviť vzorové príspevky zo sociálnej siete, ktoré budeme analyzovať. V koreňovom adresári vytvoríme textový súbor, ktorého obsahom budú príspevky zo sociálnej siete. Každý príspevok sa musí nachádzať na novom riadku a začínať jedinečným identifikátorom príspevku nasledovaný znakom „#“ (viď súbor *example.txt* v koreňovom adresári).

Proces mapovania potom spustíme pomocou príkazu (v koreňovom adresári programu):

./map -c “nazov_televiznej_stanice” -f “cesta_k_suboru_so_vzorkou”

Pre príkaz *map* máme k dispozícii nasledujúce prepínače:

- *c* - názov televíznej stanice, pre ktorú hľadáme mapovanie (z bázy televíznych metadát),
- *f* - cesta k súboru so vzorkou, v ktorej sa nachádzajú príspevky zo sociálnej siete,
- *o* - cesta k súboru, do ktorého budú uložené výsledky mapovania (ak nie je tento parameter prítomný, výsledky sa vypíšu na štandardný výstup),
- *t* - prahová hodnota rel_{thres} (predvolene 0,70).

Okrem uvedených prepínačov je možné nastaviť ďalšie parametre v konfiguračnom súboru. Konfiguračný súbor *./config.yml* sa nachádza v koreňovom adresári. Pomocou tohto súboru je možné nastavovať nasledujúce parametre:

- *default_channel* - predvolená televízna stanica (použije sa v prípade, že nie je prítomný prepínač *c*),
- *default_file* - predvolená cesta k súboru so vzoroku (použije sa v prípade, že nie je prítomný prepínač *f*),
- *default_thres* - predvolená hodnota prahovej hodnoty rel_{thres} (použije sa v

prípade, že nie je prítomný prepínač *t*),

- *prefix_index_enabled* - ak je nastavené na true, počas mapovania bude povolené používať prefixový index,
- *named_entities_enabled* - ak je nastavené na true, počas mapovania bude povolené používať pomenované entity,
- *hashtags_enabled* - ak je nastavené na true, počas mapovania bude povolené používať hashtagy,
- *external_links_enabled* - ak je nastavené na *true*, počas mapovania bude povolené používať externé prepojenia.