Systém pre sumarizáciu dokumentov na webe

Michal Belica, Vladimír Bartík

University of Technology, Brno, Czech Republic,
{ibelica,bartik}@fit.vutbr.cz
WWW: http://www.fit.vutbr.cz/~{ibelica,bartik}/

Abstrakt Práca sa zaoberá sumarizáciou dokumentov vo formáte HTML. Ako jazyk webových dokumentov bola zvolená čeština. Najprv sú v krátkosti spomenuté predchádzajúce práce a súčasný stav sumarizačných metód. Spomenuté sú dva jednoduché sumarizačné algoritmy, pričom práca sa zameriava na pokročilý algoritmus využívajúci latentnú sémantickú analýzu. Výsledkom práce je návrh a implementácia sumarizačného modulu pre jazyk Python. Súhrny generované implementovanými sumarizačnými metódami sú v záverečnej časti práce porovnané pomocou evaluačných metód i z pohľadu subjektívneho hodnotenia autora práce.

Kľúčové slová: dolovanie z dát, sumarizácia textov, redukcia dát, extrakcia dát z webu, Python, NLP, spracovanie prirodzeného jazyka, latentná sémantická analýza, LSA, singulárna dekompozícia, SVD

1 Úvod

Pod pojmom automatická sumarizácia textu budem v tejto práci rozumieť proces vytvorenia súhrnu z textového dokumentu vo formáte HTML. Výsledný súhrn môže byť použitý napr. pre rýchle zoznámenie sa s niekoľkými dokumentami a pomôcť pri rozhodnutí o vhodnosti čítania ich celého obsahu (indikatívny extrakt). Dlhšie súhrny však môžu byť použité aj ako náhrada za pôvodný dokument (informatívny extrakt).

Nasledujúca sekcia v krátkosti popisuje hlavné míľniky v metódach pre automatickú sumarizáciu a existujúce nástroje. Sekcia 3 sa podrobnejšie zaoberá mnou použitým prístupom k automatickej sumarizácii webových dokumentov. Stručný popis vykonaných experimentov je možné nájsť v sekcii 4. V záverečnej sekcii 5 sú zhrnuté poznatky z tejto práce.

2 Súčasný stav

Medzi prvé pokusy o automatizované vytvorenie sumarizácie patrí Luhnova práca [5]. Ten využíval jednoduchú heuristiku, ktorá sa zakladala na tom, že najvhodnejšími vetami do súhrnu sú tie s najviac frekventovanými frázami. Ďalšou významnou prácou boli experimenty Edmundsona [1], ktorý prispel troma novými heuristikami a tým dosiahol zlepšenie kvality výsledných súhrnov. Tieto

metódy, hoci boli jednoduché, vykazovali veľmi dobré výsledky. Časom sa ale začali objavovať sofistikovanejšie prístupy ako napríklad metódy založené na Naivnej Bayesovskej klasifikácii [4] alebo rôzne metódy z oblasti soft-computingu [3].

Jednou z moderných metód používaných v súčastnosti, nie len pre potreby sumarizácie, je metóda založená na LSA[1] [2,3,6]. Jej veľkou výhodou pri spracovaní textu je to, že implicitne analyzuje skryté vzťahy medzi slovami, slovnými spojeniami a vetami nezávisle na použitom prirodzenom jazyku. Tým sa dokáže efektívne vysporiadať s problémami ako sú napr. nejednoznačnosť výrazov a problém synonymie.

Myšlienka automatickej sumarizácie nie je nová, a preto existuje niekoľko funkčných implementácií. Aplikácia *Open Text Summarizer*[2] je implementácia dostupná na platforme UNIX, ktorá využíva jednu zo základných sumarizačných techník. Modernými knižnicami sú *Musutelsa*[3] a *Almus*[4] využívajúce metódu LSA. Najkomplexnejším riešením je platforma *MEAD*[5], ktorá obsahuje aj nástroje určené pre vyhodnocovanie kvality sumarizácií. Každá z uvedených implementácií má však nejakú nevýhodu. Najčastejšou nevýhodou je nutnosť kompatibilného vstupného XML súboru, či podpora len niekoľkých jazykov.

3 Popis sumarizačného systému

Funkčnou implementáciou sumarizačného systému je modul *sumy* pre jazyk Python. Významnú časť sumarizácie tvorí predspracovanie, ktoré spočíva hlavne v prevedení dokumentu vo formáte HTML do modelu dokumentu vhodného pre ďalšie spracovanie. Automatickú extrakciu hlavného textu zo štruktúry HTML dokumentu zabezpečuje platforma *Readability*. Dokument je následne rozdelený na tokeny (slová, vety, ...), sú odstránené slová bez sémantického významu (tzv. stop-slová) a na zvyšné slová je aplikovaný stemming.

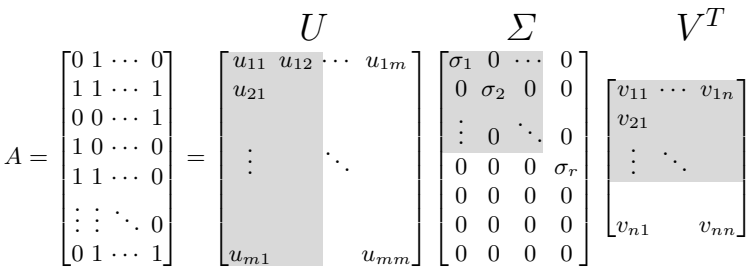
Predspracovaný dokument, reprezentovaný vektorovým modelom, je pred samotnou sumarizáciou upravený pre použitie metódou LSA. Pre jednotlivé zložky vektorového priestoru je použitá metrika uvedená v rovnici 1, kde $\mathrm{f}(t,d)$ vyjadruje počet termov t (slov) v dokumente d, $\mathrm{MaxFreq}(d)$ vyjadruje počet výskytov najfrekventovanejšieho termu v dokumente d a $0 \leq s < 1$ je parameter, ktorého úlohou je tlmiť príspevok druhého výrazu (zvyčajne $s = 0{,}5$).

$$\mathrm{TF}(t,d) = s + (1-s)\frac{\mathrm{f}(t,d)}{\mathrm{MaxFreq}(d)} \quad . \tag{1}$$

Celý dokument, ktorý obsahuje m slov a n viet je možno vyjadriť ako maticu $A = [A_1, A_2, \ldots, A_n]$ rozmeru $m \times n$, kde riadky predstavujú slová a stĺpce vety dokumentu. Stĺpcový vektor A_i reprezentuje frekvencie slov vo vete i pôvodného

[1] LSA - latentná sémantická analýza, angl. *Latent Semantic Analysis*
[2] Open Text Summarizer - `http://libots.sourceforge.net/`
[3] Musutelsa - `http://www.musutelsa.jamstudio.eu/`
[4] Almus - `http://textmining.zcu.cz/?section=download`
[5] MEAD - `http://www.summarization.com/mead/`

dokumentu spočítané podľa rovnice 1. Maticová reprezentácia dokumentu je pomocou metódy SVD[6] rozložená na 3 matice znázornené na obrázku 1.



Obr. 1: Singulárny rozklad matíc s naznačeným redukovaným priestorom

Na rozklad matice sa môžeme pozerať ako na rozklad pôvodného dokumentu do k lineárne nezávislých bázových vektorov reprezentujúcich hlavné témy textu. Výsledkom je, že podobná kombinácia slov sa bude vyskytovať pozdĺž rovnakého singulárneho vektoru. Potom matica A mapuje slová do jednotlivých viet a matice U a V mapujú slová resp. vety do k najvýznamnejších tém.

4 Vyhodnotenie sumarizácií

Pre vyhodnocovanie sumarizácií som použil tri metódy založené na podobnosti výberu viet, a to *precision*, *recall*, *F-score* a metódy *Cosine similarity* a *Unit overlap* založené na podobnosti obsahu. Ako dátové sady som použil HTML dokumenty s textom napísaným v českom jazyku. Prvá dátová sada je získaná zo serveru http://zdrojak.cz/ (ďalej len **ZDROJAK13**) a druhá je tvorená textami zo serveru http://cs.wikipedia.org/ (ďalej len **WIKI13**). Prvá dátová sada obsahuje dokumenty s kvalitným sémantickým značkovaním a členením textu. Články dátovej sady **ZDROJAK13** obsahovali aj krátke úvodníky, ktoré sa dali použiť ako abstrakty dokumentov pri evaluačných metódach založených na podobnosti obsahu. Dátová sada **WIKI13** neobsahuje sémantické značkovanie v takom množstve a kvalite ako sada **ZDROJAK13**, ale stále poskytuje dostatočné informácie o základnom členení textu.

Evaluácia sumarizácií bola uskutočnená pre Luhnovu, Edmundsonovu metódu a metódu LSA. Pre účely experimentovania bol dodatočne implementovaný sumarizátor, ktorý extrahuje vety do súhrnu náhodne. Náhodný sumarizátor slúžil ako spodná hranica ohodnotenia súrnu. Pretože takýto sumarizátor pracuje nedeterministicky, tak boli experimenty s ním vždy 100-krát opakované a výsledky spriemerované. Konfigurácia jednotlivých sumarizátorov bola ponechaná na východzie hodnoty modulu *sumy*.

[6] SVD - singulárna dekompozícia (angl. *Singular Value Decomposition*)

Ako prvé boli vyhodnotené sumarizácie pre indikatívne extrakty, ktoré predstavovali text v rozsahu približne 3 % rozsahu pôvodného dokumentu (1–3 vety). Ako ukazuje tabuľka 1, najlepšie výsledky pre indikatívny extrakt dosiahla v prevažnej väčšine metrík Edmundsonova metóda. To je spôsobené tým, že metóda dokáže využiť najviac metadát z HTML dokumentu. Slovníky *bonus slov* a *stigma slov* metóda získava na základe sémantických značiek dodaných priamo autorom textu, slová v nadpisoch sú korektne rozpoznané a je im priradená zodpovedajúca významnosť a metóda je tiež schopná využiť ohodnotenie viet na základe ich pozície. Preto je metóda schopná vybrať hodnotné vety už pri malej dĺžke súhrnu.

Hodnotiaca metrika/Typ sumarizátoru	*Random*	*Luhn*	*Edmundson*	*LSA*
Precision	0,153	0,000	**0,250**	0,083
Recall	0,018	0,000	**0,027**	0,007
F-score	0,032	0,000	**0,049**	0,013
Cosine similarity (abstract)	0,110	0,113	**0,196**	0,159
Cosine similarity (extract)	0,271	0,257	**0,381**	0,306
Cosine similarity (document)	0,321	**0,476**	0,465	0,435
Unit overlap (abstract)	0,062	0,050	**0,093**	0,077
Unit overlap (extract)	0,073	0,078	**0,118**	0,087
Unit overlap (document)	0,046	0,073	**0,073**	0,073

Tabuľka 1: Výsledky evaluácie indikatívneho súhrnu (3 % pôvodného textu) pre dátovú sadu **ZDROJAK13**

Druhá sada experimentov predstavovala vyhodnotenie sumarizácie pre informatívne extrakty, ktoré predstavovali text v rozsahu približne 10 % rozsahu pôvodného dokumentu (priemerne 8 viet). Pre HTML dokument, ktorý má bohatú sémantickú štruktúru dosahuje Edmundsonova metóda najlepšie výsledky i pre informatívne extrakty. Avšak, pri použití dokumentov so základným značkovanie už dosahuje lepšie výsledky metóda LSA (tabuľka 2). Tá si totiž vystačí so samotnou štruktúrou textu a sémantické značkovanie pre jej činnosť nepredstavuje významnú výhodu.

Pri hodnotení sumarizácií sa ukázalo, že dokonca aj náhodný sumarizátor dosahuje hodnôt blízkych k súhrnom, ktoré boli získané z algoritmických sumarizátorov. Je to spôsobené tým, že uvedené evaluačné metódy ohodnocujú súhrny veľmi povrchne. Kvôli tomu je potrebné hľadať nie len nové sumarizačné ale i evaluačné algoritmy pre automaticky tvorené súhrny.

Pre obe dátové sady boli subjektívne najhoršie, po obsahovej stránke i z hľadiska súvislosti textu, súhrny vytvorené Luhnovou sumarizačnou metódou. Súhrny sa líšili len málo od súhrnov vytvorených náhodným sumarizátorom a v niektorých prípadoch obsahovali aj vety bez informačnej hodnoty. Porovnanie metódy LSA s Edmundsonovou metódou nie je jednoznačné, pretože kvalita súhrnu závisí na použitej dátovej sade. Pre dátovú sadu **ZDROJAK13** produkuje jednoznačne najkvalitnejšie súhrny Edmundsonova metóda. Avšak pre dátovú sadu

Hodnotiaca metrika/Typ sumarizátoru	*Random*	*Luhn*	*Edmundson*	*LSA*
Precision	0,130	0,160	0,225	**0,264**
Recall	0,050	0,067	0,101	**0,117**
F-score	0,069	0,092	0,136	**0,157**
Cosine similarity (extract)	0,428	0,506	0,500	**0,537**
Cosine similarity (document)	0,554	**0,700**	0,572	0,639
Unit overlap (extract)	0,147	0,196	0,188	**0,226**
Unit overlap (document)	0,148	0,217	0,147	**0,223**

Tabuľka 2: Výsledky evaluácie informatívneho súhrnu (10 % pôvodného textu) pre dátovú sadu **WIKI13**

WIKI13 sa kvalita súhrnov vytvorených Edmundsonovou metódou značne zhoršila. Metóda LSA síce neprodukuje významne lepšie súhrny ako pre dátovú sadu **ZDROJAK13**, ale pretože Edmundsonova metóda nemá k dispozícii kvalitné metadáta o dokumente, sú pre dátovú sadu **WIKI13** výstupy sumarizátoru založeného na metóde LSA najkvalitnejšie z hodnotených.

5 Záver

V práci bol predstavený systém pre automatickú sumarizáciu webových dokumentov. Pretože celý proces sumarizácie vrátane predspracovania je plne automatizovaný, tak nie je potrebná žiadna interakcia človeka. Na samotnú sumarizáciu je využitá moderná metóda, ktorá je nezávislá na jazyku dokumentu a dokáže sa vyrovnať s problémami, ktorými trpia iné sumarizačné metódy. Oproti bežne rozšíreným sumarizátorom sa metóda snaží ťažiť z meta-informácií o dokumente, ktoré ponúka formát HTML aby sa dosiahlo vyššej kvality výsledného súhrnu.

Sumarizačný systém je vyvíjaný ako open-source[7] modul pre jazyk Python s využitím knižníc *nltk* a *scipy*. Súčasťou modulu sú aj nástroje využívané k vyhodnocovaniu výsledného súhrnu. Časť starajúca sa o extrahovanie hlavného textu z HTML štruktúry je implementovaná oddelene ako port platformy *Readability*.

V ďalšom vývoji sa počíta s niekoľkými vylepšeniami. Ako sa ukázalo, segmentácia textu na vety a slová nedokáže korektne spracovať niektoré časti webových dokumentov a vyžaduje použitie sofistikovanejších algoritmov. Aby bolo možné sumarizačný systém porovnávať s konkurenčnými systémami, bude nutné implementovať parser pre niekoľko nových vstupných formátov. Nakoniec, veľmi vhodným vylepšením bude zaradiť do procesu predspracovania dokumentu systém pre *rezolúciu anaforických vzťahov*.

Poďakovanie. Tento výskum je podporovaný výskumným zámerom MSM 0021630528 a projektom IT4Innovations Centre of Excellence CZ.1.05/1.1.00/02.0070.

[7] Zdrojové kódy modulu *sumy* - https://github.com/miso-belica/sumy

Literatúra

1. Edmundson, H. P. New Methods in Automatic Extracting. J. ACM. Apríl 1969, roč. 16, č. 2. S. 264–285. Dostupné na: <http://doi.acm.org/10.1145/321510.321519>. ISSN 0004-5411.
2. Gong, Y. a Liu, X. Generic text summarization using relevance measure and latent semantic analysis. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM, 2001. S. 19–25. SIGIR '01. Dostupné na: <http://doi.acm.org/10.1145/383952.383955>. ISBN 1-58113-331-6.
3. Ježek, K. a Steinberger, J. Sumarizace textů. In DATAKON. 2010.
4. Kupiec, J., Pedersen, J. a Chen, F. A trainable document summarizer. In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM, 1995. S. 68–73. SIGIR '95. Dostupné na: <http://doi.acm.org/10.1145/215206.215333>. ISBN 0-89791-714-6.
5. Luhn, H. P. The automatic creation of literature abstracts. IBM Journal Res. Dev. Apríl 1958, roč. 2, č. 2. S. 159–165. Dostupné na: <http://dx.doi.org/10.1147/rd.22.0159>. ISSN 0018-8646.
6. Steinberger, J. a Ježek, K. Using latent semantic analysis in text summarization and summary evaluation. In Proceedings ISIM '04. 2004. S. 93–100.