Porovnání technik předzpracování textu pro detekci plagiátů

Zdeněk Češka[1]

[1]Katedra informatiky a výpočetní techniky, Fakulta aplikovaných věd,
Západočeská univerzita v Plzni, Univerzitní 22, 306 14 Plzeň, Česká republika
zceska@kiv.zcu.cz

Abstrakt. Tento článek se zabývá technikami předzpracování textu a jejich vlivem na detekci plagiátů v psaném textu. V našich experimentech zkoumáme stop-slova, lemmatizaci, nahrazování synonym a jejich vzájemné kombinace. Dále navrhujeme pokročilou normalizaci slov s využitím hyperonym z WordNet tezauru. Testy jsme provedli na českém korpusu plagiátů čítajícím 950 dokumentů o politice, vytvořeném z ČTK korpusu. Pro experimenty používáme metodu postavenou na RFM, prostém srovnání N-gramů s Jaccard-Tanimoto koeficientem a metodu pracující na principu singulární dekompozice vztahů frází.

Klíčová slova: plagiátorství, stop-slova, lemmatizace, synonyma, hyperonyma.

1 Úvod

Pro detekci plagiátů v psaném textu existuje mnoho metod, nicméně ne všechny používají stejné předzpracování textu. V tomto článku ověříme vliv obvyklých technik pro předzpracování textu jako je odstraňování stop-slov a lemmatizaci, včetně méně obvyklých jako nahrazování synonym a čísel. Jednou z posledních technik je normalizace slov s využitím hyperonym, která dokáže zobecnit význam slova. Například slova „pes“ a „kočka“ mohou být převedena na slovo „zvíře“. Cílem těchto experimentů je odhalit závislosti mezi různými technikami předzpracování a doporučit nejvhodnější kombinaci. Tento článek volně navazuje na „Využití moderních přístupů pro detekci plagiátů“ [1], kde byly představeny základy metody pro detekci plagiátů s využitím singulární dekompozice vztahů frází.

Mimo jiné se zaměříme na vliv interpunkce na extrakci slovních N-gramů, představující fráze v textu. Mějme slovní N-gram, který se rozprostírá mezi dvěma větami oddělené tečkou, nebo kupříkladu mezi hlavní a vedlejší větou oddělené čárkou. Pak je nutné zvážit, zda je N-gram smysluplný a má být během výpočtu uvažován či nikoli. V experimentech se podíváme na množství extrahovaných N-gramů pod vlivem interpunkce a rovněž na kvalitu výsledků a možná rizika z toho plynoucí.

K experimentům používáme tři metody pro detekci plagiátů. Nejjednodušší metoda staví na relativních frekvencích slov (RFM), detailně popsaná Garciou-Molinou [5]. Pokročilejší metody pak využívají slovní N-gramy specifické délky. Při výpočtu podobnosti se nejčastěji setkáme s Jaccard-Tanimoto koeficientem. Příkladem může být systém Ferret [2], který počítá průnik společných slovních trigramů mezi dvěma dokumenty. Poslední použitá metoda využívá singulární dekompozici (SVD) pro analýzu vztahů frází v textu. Tato metoda byla popsána v článku [1], kde dosahuje lepších výsledků v porovnání s ostatními metodami.

2 Předzpracování textu

Odstraňování stop-slov (ST) je jednou z nejzákladnějších technik. V rámci tohoto procesu jsou z textu smazána veškerá nevýznamová slova, která mohou negativně ovlivnit podobnost mezi dokumenty. Pro experimenty jsme použili slovníkovou metodu. Seznam standardizovaných slov, včetně českých, je dostupný na [3].

Lemmatizace (LM) je proces určování základního tvaru slova. V českém jazyce, který má bohaté skloňování slov, se jedná o často užívanou metodu. V našich experimentech jsme použili lemmatizační slovník postavený na Ispellu a WordNetu [7]. Více informací o tvorbě slovníku a vyhledávání vhodných lemmat naleznete v [6].

Záměna synonym (SY) je často užívanou technikou plagiátorů pro zakrytí okopírovaného textu. K odhalení tohoto triku používáme WordNet tezaurus [7], kde slova se stejným významem jsou uskupena do synsetů.

Nahrazování čísel (NM) je méně známou technikou převádějící čísla na společný zástupný symbol. Zůstává pouze informace o existenci čísla, nikoli jeho hodnotě. Tato technika nalezne uplatnění u drobných modifikací textu podobně jako SY.

Normalizace slov s využitím hyperonym (HY) je pokročilejší variantou synonym, která zobecňuje význam slova. Příkladem je nahrazení slov „pes“ a „kočka“ za slovo „zvíře“. K tomuto účelu používáme již dříve zmíněný WordNet tezaurus, který mimo jiné obsahuje hyperonymické odkazy. Při zobecňování slov se na požadovanou úroveň dostaneme průchodem několika úrovní. Přitom hloubka hierarchie je rozličná pro různá slova. Námi navržený postup nahrazuje slova na předem danou úroveň od shora. Řekněme, že slovo „kočka“ se nachází na 7. úrovni a postupně k němu vedou slova „objekt“, „živoucí entita“, „zvíře“, „obratlovec“, „savec“ a „kočkovitá šelma“. Redukujeme-li hloubku stromu na 3, budou všechna slova na vyšších úrovních nahrazena slovem „zvíře“. V případě, že úroveň nahrazovaného slova je nižší, ponechá se původní slovo beze změn. Tudíž, slova „objekt“ a „živoucí entita“ se nemění. V experimentech používáme postup HY-3A, kde 3A je konfigurační parametr. Písmeno A označuje všechny slovní druhy a číslice 3 stanovenou úroveň pro nahrazení.

3 Experimenty

Srovnání jsme provedli na korpusu 950 článků o politice, který byl vytvořen ze standardního ČTK korpusu. Celkem je zastoupeno 150 plagiovaných článků, 300 originálních podkladových článků a 500 dalších náhodně vybraných článků o politice. Oproti korpusu v [1], bylo provedeno manuální čištění pro zvýšení kvality. K porovnání výsledků používáme klasickou míru přesnosti *p*, úplnosti *r* a míru F_1 dle [4].

V prvním experimentu se podíváme na množství extrahovaných N-gramů, viz Tabulka 1. Experiment jsme provedli na trigramech a 5-gramech, které dosáhli v předběžných testech s/bez užití interpunkce nejlepších výsledků. Rozdíl mezi trigramy a 5-gramy není bez interpunkce výrazný. Po aplikaci předzpracování počet N-gramů klesá, nicméně v případě SY a HY-3A je to skoro zanedbatelné číslo. Nejvyššího redukčního poměru okolo 23% dosahuje odstraňování stop-slov. Oproti ostatním technikám se redukční poměr zvyšuje s rostoucí délkou N-gramu. Při uvážení interpunkce se nárůst dále zvyšuje až na 38%. Lemmatizaci a nahrazování čísel lze rovněž použít s redukcí v rozsahu 1-2%. S rostoucí délkou N-gramu však klesá.

Tabulka 1. Množství trigramů/5-gramů s/bez uvážení interpunkce po aplikaci předzpracování.

Předzpracování	Bez užití interpunkce				S užitím interpunkce			
	trigramy		5-gramy		trigramy		5-gramy	
---	127970	redukce↴	137565	redukce↴	96788	redukce↴	76583	redukce↴
ST	99879	21,95%	104826	23,80%	68350	29,38%	47467	38,02%
LM	125325	2,07%	137106	0,33%	94535	2,33%	76228	0,46%
SY	127946	0,02%	137559	0,01%	96770	0,02%	76578	0,01%
NM	126669	1,02%	137322	0,18%	95552	1,28%	76374	0,27%
HY-3A	127672	0,23%	137536	0,02%	96563	0,23%	76565	0,02%

Nejlepších redukčních poměrů dosáhneme aplikací interpunkčních pravidel a volbou delších N-gramů. Kombinace 5-gramů a interpunkce však dosahuje pouhých 91,69% F_1 v případě SVD metody, viz Tabulka 2. Pokud se zaměříme na maximální přesnost a úplnost, jsou nejlepší variantou 5-gramy bez uvážení interpunkce. Za cenu vyššího množství N-gramů získáme 95,06% F_1. Jistým kompromisem může být kombinace trigramů a interpunkce, která dosahuje 93,00% F_1.

Tabulka 2. Vliv jednotlivých technik předzpracování na detekci plagiátů při užití 5-gramů.

Metoda	Předzpracování	Bez interpunkce			S interpunkcí		
		***p* [%]**	***r* [%]**	***F₁* [%]**	***p* [%]**	***r* [%]**	***F₁* [%]**
RFM	---	84,02	89,88	86,85	84,02	89,88	86,85
	ST	85,33	87,93	86,61	85,33	87,93	86,61
	LM	82,67	89,74	86,06	82,67	89,74	86,06
	SY	83,14	92,50	87,57	83,14	92,50	87,57
	NM	84,90	89,12	86,96	84,90	89,12	86,96
	HY-3A	85,36	90,02	87,63	85,36	90,02	87,63
Jaccard	---	88,18	98,98	93,27	85,91	96,43	90,87
	ST	90,00	94,29	92,09	84,55	90,29	87,32
	LM	88,18	97,00	92,38	86,36	95,96	90,91
	SY	88,18	98,98	93,27	85,91	96,43	90,87
	NM	88,18	98,48	93,05	86,36	96,45	91,13
	HY-3A	88,18	98,98	93,27	85,91	96,43	90,87
SVD	---	91,82	98,54	95,06	87,73	96,02	91,69
	ST	90,45	95,67	92,99	80,91	93,68	86,83
	LM	90,91	93,46	92,17	89,09	95,15	92,02
	SY	91,82	98,54	95,06	87,73	96,02	91,69
	NM	92,27	98,07	95,08	87,73	96,02	91,69
	HY-3A	91,82	98,54	95,06	87,73	96,02	91,69

Aplikací SY nebo HY-3A nedochází k žádnému výraznému zlepšení ani zhoršení míry F_1. Nahrazování čísel v případě RFM a SVD bez interpunkce nepatrně zlepšuje F_1. Zbylé techniky, jako odstraňování stop-slov a lemmatizace, výsledky zhoršují. Lemmatizace téměř pokaždé zhoršuje míru F_1. V případě SVD je pokles výrazný z 95,06% na 92,17%. Odstraňování stop-slov má podobné chování.

V závěru se podíváme na kombinace jednotlivých technik. Z důvodu jejich velkého množství uvedeme pouze slovní souhrn pro metodu SVD. Jako baseline nám poslouží technika bez předzpracování s 95,06% F_1. Z výsledků vyplývá, že dvě různé spolu nesouvisející techniky se mohou vzájemně doplňovat. Techniky ST a LM dávají výsledek 93,79% F_1 oproti samostatné aplikaci 92,99% a 92,17%. Často je použitím dvou a více technik výsledné skóre vyšší. Aplikací všech technik získáme skóre 94,04%. Vynecháním ST pak 94,44%, což není z pohledu redukční poměru výhodné.

4 Závěr

Na základě provedených experimentů není předzpracování schopné zlepšit výsledky detekce plagiátů. Pouze nahrazování čísel zanedbatelně zlepšuje výsledky. U SVD metody na 95,08% F_1 oproti baseline 95,06%. Preferujeme-li vyšší redukci N-gramů, při současně vysoké přesnosti, je vhodné aplikovat všechny techniky předzpracování. Výsledné F_1 pak klesá na 94,04% oproti baseline. Největší podíl na redukci příznaků zastávají stop-slova. Pokles míry F_1 však naznačuje narušení autorského stylu. Lemmatizace rovněž, jako v případě klasifikace [6], negativně ovlivňuje přesnost. Zabývat se synonymy či hyperonymy nemá příliš smysl, jejich výsledky jsou nepřesvědčivé.

Výše zmíněná baseline pracovala s 5-gramy bez interpunkce. Je-li nezbytné podstatně snížit objem příznaků, jsou interpunkční pravidla nejlepší volbou. Pro 5-gramy klesá počet příznaků z 137565 na 76583. Současně však, pro SVD, klesá míra F_1 z 94,06% na 91,69%. Aplikací všech technik dosáhneme dalšího snížení příznaků, ale i horších výsledků. Jistým kompromisem jsou trigramy s uvážením interpunkce, kdy bez předzpracování obdržíme 96788 příznaků a F_1 93,00%. Obecně se počet příznaků zvyšuje s rostoucí délkou N-gramu. Uvážením interpunkce naopak klesá vlivem výpadku krátkých vět s méně slovy než je délka N-gramu. Dopad na přesnost je radikální, dle experimentů nedoporučujeme volit N-gramy delší než 5 slov.

Tato práce byla částečně podporována z prostředků NPV II, projekt 2C06009 (COT-SEWing).

Reference

1. Z. Ceska, „Využití moderních přístupů pro detekci plagiátů", *Proceedings of the ITAT 2008*, pp. 23-26, Hrebienok, Slovakia, 2008. ISBN 978-80-969184-8-5.
2. P. Lane, C. Lyon, J. Malcolm, „Demonstration of the ferret plagiarism detektor", *Proceedings of the 2nd International Plagiarism Conference*, Newcastle, 2006.
3. RANKS.NL, „Czech stopwords". URL: http://www.ranks.nl/stopwords/czech.html
4. C. Rijsbergen, „Information Retrieval. Butterworth-Heinemann", 2nd rev. edition, 1979. ISBN 0-408-70929-4.
5. N. Shivakumar, H. Garcia-Molina, „SCAM: A copy detection mechanism for digital documents", *Proceedings of 2nd International Conference in Theory and Practice of Digital Libraries*, Austin, 1995.
6. M. Toman, R. Tesar, K. Jezek, „Influence of word normalization on text classification", *Proceedings of the 1st International Conference on Multidisciplinary Information Sciences & Technologies*, vol. 2, pp. 354-358, Merida, Spain, 2006. ISBN 84-611-3105-3.
7. P. Vossen, „Global WordNet Association: EuroWordNet". Last update 9/1/2001. URL: http://www.illc.uva.nl/EuroWordNet/

Annotation:

Comparison of Pre-processing Techniques for Plagiarism Detection

This paper deals with the comparison of stop-word removal, lemmatization, synonym replacement, and number replacement techniques for plagiarism detection. Further, we propose advanced word normalization with the use of hyperonyms. We examine the influence of different pre-processing on plagiarism detection methods and recommend the best one solution.