VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

AUTOMATICKÁ EXTRAKCE KLÍČOVÝCH SLOV V ČEŠTINĚ

AUTOMATIC KEYWORD EXTRACTION IN CZECH

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — **ĽUBOMÍR GALLOVIČ**
AUTHOR

VEDOUCÍ PRÁCE — **Doc. RNDr. PAVEL SMRŽ, Ph.D.**
SUPERVISOR

BRNO 2017

Zadání bakalářské práce/15374/2016/xgallo03

Vysoké učení technické v Brně - Fakulta informačních technologií

Ústav počítačové grafiky a multimédií Akademický rok 2016/2017

Zadání bakalářské práce

Řešitel: **Gallovič Ľubomír**
Obor: Informační technologie
Téma: **Automatická extrakce klíčových slov v češtině**
Automatic Keyword Extraction in Czech
Kategorie: Umělá inteligence

Pokyny:

1. Seznamte se s metodami automatické extrakce klíčových slov a existujícími nástroji, které umožňují zpracovávat český text.
2. Shromážděte rejstříky knih a dalších publikací dostupných v elektronické podobě a zpracujte je do tvaru vhodného pro vyhodnocování výsledků práce.
3. Navrhněte a implementujte systém pro porovnávání výsledků metod automatické extrakce klíčových slov z českých odborných knih.
4. Vyhodnoťte vytvořený systém srovnáním s daty skutečných rejstříků, diskutujte možnosti zvýšení výkonnosti a úspěšnosti.
5. Vytvořte stručný plakát prezentující práci, její cíle a výsledky.

Literatura:

- dle doporučení vedoucího

Pro udělení zápočtu za první semestr je požadováno:

- Prototyp systému

Podrobné závazné pokyny pro vypracování bakalářské práce naleznete na adrese http://www.fit.vutbr.cz/info/szz/

Technická zpráva bakalářské práce musí obsahovat formulaci cíle, charakteristiku současného stavu, teoretická a odborná východiska řešených problémů a specifikaci etap (20 až 30% celkového rozsahu technické zprávy).

Student odevzdá v jednom výtisku technickou zprávu a v elektronické podobě zdrojový text technické zprávy, úplnou programovou dokumentaci a zdrojové texty programů. Informace v elektronické podobě budou uloženy na standardním nepřepisovatelném paměťovém médiu (CD-R, DVD-R, apod.), které bude vloženo do písemné zprávy tak, aby nemohlo dojít k jeho ztrátě při běžné manipulaci.

Vedoucí: **Smrž Pavel, doc. RNDr., Ph.D.**, UPGM FIT VUT
Datum zadání: 1. listopadu 2016
Datum odevzdání: 17. května 2017

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
Fakulta informačních technologií
Ústav počítačové grafiky a multimédií
612 66 Brno, Božetěchova 2

L.S.

doc. Dr. Ing. Jan Černocký
vedoucí ústavu

Abstrakt

Táto práca sa zaoberá návrhom, implementáciou a testovaním aplikácie pre automatickú extrakciu kľúčových výrazov z odborných textov v českom jazyku. Sú implementované viaceré algoritmy výberu kandidátov a rôzne štatistické a linguistické metódy výpočtu skóre týchto kandidátov. Jednotlivé algoritmy boli analyzované a porovnávané, a tie, ktoré dosiahli v českom jazyku najlepšie výsledky, boli vybrané do finálnej verzie programu.

Abstract

This thesis describes design, implementation and testing of application for automatic keyterm extraction from technical texts in czech language. Multiple algorithms for candidate selection, as well as various statistical and linguistic methods for score calculation were implemented. All of these algorithms were analyzed and compared, and best performing ones were chosen to be included in the final version of the program.

Kľúčové slová

kľúčové slová, extrakcia, spracovanie prirodzeného jazyka, výrazy

Keywords

keywords, extraction, natural language processing, terms

Citácia

GALLOVIČ, Lubomír. *Automatická extrakce klíčových slov v češtině*. Brno, 2017. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Smrž Pavel.

Automatická extrakce klíčových slov v češtině

Prehlásenie

Prehlasujem, že som túto bakalársku prácu vypracoval samostatne pod vedením pána Doc. RNDr. Pavla Smrža Ph.D.. Uviedol som všetky literárne pramene a publikácie, z ktorých som čerpal.

.....
Ľubomír Gallovič
11. mája 2017

Poďakovanie

Rád by som sa poďakoval pánovi Doc. RNDr. Pavlovi Smržovi Ph.D. za odbornú pomoc a konzultácie pri tvorbe tejto práce.

Obsah

1 Úvod **4**

2 Metódy výberu kandidátov **6**
2.1 Vymedzenie pojmov . . . 6
2.2 Výber kandidátov na základe stop slov . . . 6
2.3 Výber kandidátov na základe POS tagov . . . 7

3 Štatistické metódy skórovania kolokácií **8**
3.1 Metódy výpočtu miery asociácie v bigramoch . . . 9
3.1.1 Z-score . . . 9
3.1.2 Z-score corrected . . . 9
3.1.3 T-score . . . 9
3.1.4 Log-likelihood ratio . . . 10
3.1.5 Logarithmic Odds Ratio . . . 10
3.1.6 Dice coefficient, Jaccard coefficient . . . 10
3.1.7 Pointwise Mutual Information . . . 10
3.1.8 Pearson's chi-squared test . . . 10
3.2 Metódy výpočtu miery asociácie v trigramoch . . . 11
3.2.1 Pointwise Mutual Information . . . 11
3.2.2 True Mutual Information . . . 11
3.2.3 Log-likelihood . . . 11
3.2.4 Poisson-Stirling . . . 11

4 Lingvistické metódy skórovania kandidátov **12**
4.1 Vymedzenie pojmov . . . 12
4.2 TF-IDF . . . 12
4.3 KP-Miner . . . 13
4.4 KX-FBK . . . 13
4.5 RAKE . . . 14

5 Metódy skórovania založené na grafoch **15**
5.1 TextRank . . . 15
5.2 TopicRank . . . 15
5.3 SingleRank . . . 16
5.4 ExpandRank . . . 16
5.5 SGRank . . . 16

6 Spôsoby vyhodnocovania výsledkov **17**
6.1 Presnosť . . . 17
6.2 Úplnosť . . . 17
6.3 Obmedzenia navrhnutého vyhodnocovania . . . 18

7 Príprava testovacích dát **20**
7.1 Proces výberu relevantných publikácií . . . 20
7.2 Analýza získaných textov . . . 20

8 Návrh a implementácia programu **22**
8.1 Návrh aplikácie . . . 22
8.1.1 Načítanie vstupných dát . . . 23
8.1.2 Predspracovanie textu . . . 23
8.1.3 Výber kandidátov . . . 23
8.1.4 Použitie štatistických prístupov k výberu najvhodnejších kandidátov 23
8.1.5 Výpočet skóre kandidátov pomocou lingvistických metód . . . 23
8.1.6 Vytvorenie správnej formy výstupných dát . . . 23
8.2 Morphodita . . . 24
8.3 Testované algoritmy . . . 24

9 Analýza a porovnanie prístupov výberu kandidátov **25**
9.1 Výber kandidátov na základe stop slov . . . 25
9.1.1 Zoznam stop slov . . . 25
9.1.2 Analýza výsledkov algoritmu . . . 25
9.2 Výber kandidátov na základe POS tagov . . . 30
9.2.1 Vytvorenie zoznamu vzorov POS tagov . . . 30
9.2.2 Analýza výsledkov algoritmu . . . 31
9.3 Porovnanie výsledkov . . . 34

10 Analýza jednotlivých prístupov skórovania kandidátov **36**
10.1 Spôsob porovnávania jednotlivých algoritmov . . . 36
10.2 Štatistické metódy skórovania dvojslovných výrazov . . . 36
10.2.1 Lift grafy . . . 36
10.2.2 Grafy presnosti a úplnosti . . . 37
10.3 Štatistické metódy skórovania trojslovných výrazov . . . 39
10.3.1 Lift grafy . . . 39
10.3.2 Grafy presnosti a úplnosti . . . 40
10.4 Zhodnotenie výsledkov štatistických metód . . . 42
10.5 Analýza lingvistických prístupov skórovania kandidátov . . . 43
10.6 Lingvistické metódy skórovania jednoslovných výrazov . . . 43
10.6.1 Lift grafy . . . 43
10.6.2 Grafy presnosti a úplnosti . . . 44
10.7 Lingvistické metódy skórovania dvojslovných výrazov . . . 46
10.7.1 Lift grafy . . . 47
10.7.2 Grafy presnosti a úplnosti . . . 47
10.7.3 Analýza kombinácie algoritmov Ch-Squared a RAKE . . . 49
10.8 Lingvistické metódy skórovania trojslovných výrazov . . . 50
10.8.1 Lift grafy . . . 50
10.8.2 Grafy presnosti a úplnosti . . . 51

10.8.3 Analýza kombinácie algoritmov Poisson-Stirling a RAKE 53

11 Dosiahnuté výsledky **55**

12 Záver **56**
12.1 Možné vylepšenia 56

Literatúra **57**

Prílohy **59**

A Manuál **60**

B Obsah CD **61**

Kapitola 1

Úvod

Identifikácia kľúčových výrazov patrí medzi historicky dôležitú súčasť práce s textom. Kľúčové výrazy sa môžu skladať z jedného alebo viacerých slov a slúžia na sumarizáciu dokumentu a reprezentáciu informácií opísaných v dokumente. Správne vybrané kľúčové výrazy by mali čitateľovi ukázať, aké témy a body boli v dokumente rozoberané. V minulosti sa kľúčové výrazy vyskytovali v publikáciách na konci textu v podobe registru, ktorého účelom bolo zlepšenie orientácie v texte vyhľadaním požadovaných kľúčových výrazov, a strán na ktorých sa tieto výrazy vyskytujú. Avšak v súčasnosti existuje väčšina textov v digitálnej forme, v ktorej je vyhľadávanie a orientácia v texte omnoho jednoduchšia. Dnes teda majú kľúčové výrazy iné využitie, ako je klasifikácia a kategorizácia textu, vyhľadávanie relevantných článkov a filtrovanie podľa zvolených kritérií. Tieto funkcie sú obzvlášť dôležité pri používaní internetu, ktorý poskytuje prístup k obrovskému množstvu dát, v ktorých je potrebné sa orientovať.

Manuálny výber kľúčových výrazov je však časovo náročná činnosť a rýchlo narastajúci objem existujúcich dokumentov si vyžaduje určitú formu automatizácie vo vyhľadávaní týchto výrazov. Nie je preto žiadnym prekvapením, že touto problematikou sa zaoberá veľké množstvo vedeckých publikácií a boli vytvorené rôzne algoritmy na vykonávanie tejto činnosti. Je dôležité poznamenať, že žiadny z týchto algoritmov sa nesnaží textu sémanticky porozumieť a z kontextu vybrať správne kľúčové výrazy, využívajú len určité predvídateľné informácie spojené s výskytom hľadaných výrazov. Väčšina týchto algoritmov bola vytvorená pre a bola testovaná na anglickom jazyku a niektoré z nich pracujú s javmi, ktoré nie sú jazykovo univerzálne. Preto je otázna jednoduchá prenositeľnosť týchto algoritmov z jedného jazyka do druhého. Zatiaľ neexistuje veľa výskumu o automatickej extrakcii kľúčových výrazov z českých textov, preto hlavným zameraním tejto práce je testovanie a porovnávanie rôznych prístupov výberu kľúčových slov v českom jazyku.

Vyhľadávanie kľúčových výrazov sa zväčša rozdeľuje na dve základné časti: výber kandidátov z textu a ich zoradenie podľa skóre vypočítané daným skórovacím algoritmom. Cieľom práce je porovnávanie jednotlivých prístupov výberu kľúčových výrazov, preto je potrebné tiež zostaviť korpus a porovnať výsledky jednotlivých algoritmov. Zostavovaný korpus sa musí skladať z českých odborných textov, mal by byť dostatočne veľký a musí obsahovať manuálne zostavený zoznam kľúčových výrazov pre zisťovanie úspešnosti jednotlivých algoritmov. Po zhromaždení dostatočne veľkého korpusu nasleduje vyhľadanie, implementácia a vyhodnotenie rôznych spôsobov vyhľadania kandidátov kľúčových výrazov. Pôjde o metódy využívajúce rôzne gramatické vlastnosti jazyka a ich cieľom je vybrať výrazy, ktoré majú najväčšiu šancu byť kľúčovými. Nasleduje zoradenie kandidátov podľa skóre, ktoré bude vypočítané na základe existujúcich a prípadne aj modifikovaných algoritmov. Nie-

ktoré algoritmy pracujú len s textom, iné na výpočet skóre potrebujú podporné korpusové dáta. Po výpočte skóre nasleduje vyhodnotenie, ktoré porovná automaticky vyhľadané kľúčové výrazy so zoznamom skutočných kľúčových výrazov a zistí ich jednotlivú úspešnosť. Na záver sú zhrnuté dosiahnuté výsledky a skutočnosti vyvodené z testovania jednotlivých algoritmov.

Kapitola 2

Metódy výberu kandidátov

2.1 Vymedzenie pojmov

Prvou časťou automatickej generácie kľúčových výrazov je výber kandidátov z textu. Na určenie vhodnosti kandidátov sa používa pojem "unithood", ktorý je definovaný ako "stupeň stability syntagmatických kombinácií a kolokácií" [12]. Kandidátmi môžu byť slová aj viacslovné výrazy, viacslovné výrazy sú však často preferované, pretože práve tie bývajú najčastejšie termínmi v odbornej literatúre. Navyše majú jednoslovné výrazy častokrát príliš všeobecný význam a existuje predpoklad, že viacslovné výrazy majú užšiu sémantickú definíciu [7]. Líši sa aj spôsob ohodnocovania jednoslovných a viacslovných výrazov, kde jednoslovné výrazy častokrát potrebujú dáta o početnosti vo veľkom korpuse, avšak viacslovné výrazy je možné ohodnocovať na základe štatistickej analýzy v samotnom dokumente.

V tejto kapitole budú opísané rôzne možnosti výberu kandidátnych výrazov, ktorých vhodnosť na českých textoch budeme neskôr analyzovať a porovnávať. Popísané prístupy využívajú morfo-syntaktické vzory, ktorých dodržanie je považované za dôkaz vhodnosti potenciálneho kľúčového výrazu. Tieto metódy najviac využívajú predvídateľné vlastnosti jazyka, preto môže byť prenesenie niektorých z týchto metód z anglického jazyka do českého problematické a môže si vyžadovať určité modifikácie.

2.2 Výber kandidátov na základe stop slov

Nasledujúci spôsob výberu kandidátov bol vytvorený pre algoritmus RAKE [15], ktorý dosahuje veľmi dobré výsledky vo výbere viacslovných výrazov v odborných textoch. Táto metóda využíva skutočnosti, že v kľúčových výrazoch sa len veľmi zriedka vyskytujú určité pomocné slová, ako sú napríklad predložky a spojky. Pre prácu algoritmu je teda zhotovený zoznam týchto slov - nazývané stop slová, pomocou ktorých sú rozdeľované vety na jednotlivé výrazy. Algoritmus pracuje nasledovne:

- 1. Vstupný text je rozdelený na zoznam neprerušovaných častí textu. Znamienka ohraničujúce dané časti textu sú bodka, čiarka, dvojbodka, bodkočiarka, zátvorky a úvodzovky. Týmto získame neprerušené časti textu, v ktorých sa potenciálne môžu nachádzať kľúčové výrazy.
- 2. Jednotlivé časti sa ďalej rozdeľujú podľa stop slov, ktoré boli algoritmu poskytnuté. Vytvorené výrazy medzi stop slovami môžeme považovať za potenciálnych kandidátov.

- 3. Odstránenie príliš dlhých výrazov. V algoritme je vybraný maximálny počet slov, ktorý môžu kľúčové výrazy obsahovať. V prípade presiahnutia tohto počtu je výraz odstránený.
- 3. Filtrácia kandidátov na základe početnosti. Podobne ako v predošlom bode je v algoritme vybraný minimálny počet výskytov, pri dosiahnutí ktorého sa z výrazu stáva kandidát kľúčového výrazu.

Algoritmus sa spolieha na predvídateľné vlastnosti anglického jazyka, pričom vetné konštrukcie v iných jazykoch môžu obmedziť prenositeľnosť algoritmu. Je tiež zrejmé, že vhodnosť nájdených kandidátov je tiež závislá od kvality a robustnosti zoznamu stop slov, teda v prípade zhromaždenia príliš malého počtu stop slov budú vzniknuté konštrukcie pridlhé a nepoužiteľné ako kľúčové výrazy.

2.3 Výber kandidátov na základe POS tagov

Druhou metódou výberu kandidátov kľúčových výrazov je selekcia výrazov. ktoré spĺňajú kritéria predom vytvorených vzorov part-of-speech (POS) tagov. Prvým krokom je teda použitie taggera na zistenie POS tagov jednotlivých slov. Tieto tagy nám dávajú informácie o slove a kontexte, v ktorom je slovo použité. Ide o informácie ako slovný druh, pád, rod, číslo, a ďalšie špecifické vlastnosti daného slova, z ktorých niektoré sa môžu líšiť v závislosti na používanom jazyku. Pri výbere kandidátov nám stačia len niektoré z týchto vlastností, pričom najčastejšie ide o slovný druh a prípadne pád, ak napríklad chceme nájsť podstatné mená v genitívnej väzbe. Nasledujúci zoznam vzorov POS tagov bol navrhnutý na vyhľadávanie kľúčových výrazov v matematiky zameraných textoch v anglickom jazyku [8].

Adjective Noun	linear function
Noun Noun	regression coefficients
Adj. Adj. Noun	Gaussian random variable
Adj. Noun Noun	cumulative distribution function
Noun Adj. Noun	mean squared error
Noun Noun Noun	class probability function
Noun Preposition Noun	degrees of freedom

Tabuľka 2.1: Anglické POS tagy

Je možné predpokladať, že dané vzory POS tagov nemusia fungovať po prenesení do iného jazyka, preto bude pravdepodobne nutné zhotoviť vlastné vzory podľa analýzy skutočných kľúčových výrazov. Takisto je zrejmé, že použiteľnosť niektorých POS tagov sa líši podľa témy danej vedeckej publikácie: napríklad tag podstatné meno - prídavné meno má v češtine vyššie využitie v oblasti biológie, kde označuje názvy rastlín a živočíchov (napr. Tiger indický). Vhodnosť jednotlivých POS tagov je teda možné posudzovať podľa dvoch kritérií: koľko výrazov dokázali identifikovať, a aká je ich percentuálna úspešnosť identifikovania správnych výrazov. Rozhodnutie využiť, alebo nevyužiť jednotlivé tagy teda závisí od preferencie užívateľa medzi získaním čo najväčšieho počtu výrazov a nájdením výrazov, ktoré majú väčšiu pravdepodobnosť, že sú správne. Súčasťou testovania algoritmu bude teda aj porovnávanie dosiahnutých výsledkov jednotlivých tagov.

Kapitola 3

Štatistické metódy skórovania kolokácií

Pomocou jedného z algoritmov popísaných v predošlej kapitole je získaný zoznam kandidátov, ktorých vhodnosť je však potrebné kvantifikovať. V tejto kapitole budú predstavené štatistické metódy ohodnocovania dvoj a trojslovných výrazov využívajúce dáta získané z textu. Výška skóre určuje mieru unithood popísanú v predošlej kapitole, teda cieľom týchto metód je zistiť, či sú výrazy viacslovnými pomenovaniami. Popísané prístupy fungujú na predpoklade, že frekvencia spoluvýskytu slov tvoriacich kľúčové výrazy je príliš vysoká na to, aby mohla byť považovaná za náhodný jav. Prvým krokom práce algoritmu je získanie informácií o počte výskytov výrazu a slov tvoriacich tento výraz v texte. Po získaní všetkých potrebných informácií z textu je zhotovená kontingenčná tabuľka pre daný výraz [5].

	b	**not b**	
a	o_{11}	o_{12}	o_{1p} **(R1)**
not a	o_{21}	o_{22}	o_{2p} **(R2)**
	o_{p1} **(C1)**	o_{p2} **(C2)**	o_{pp} **(N)**

Tabuľka 3.1: kontingenčná tabuľka pre bigramy

		c	**not c**
a	**b**	o_{111}	o_{112}
a	**not b**	o_{121}	o_{112}
not a	**b**	o_{211}	o_{212}
not a	**not b**	o_{221}	o_{222}

Tabuľka 3.2: kontingenčná tabuľka pre trigramy

Táto tabuľka obsahuje informácie o spoluvýskyte slov testovaného výrazu, výskyte daných slov s inými výrazmi a celkovom počte n-gramov. Vypočítajú sa očakávané hodnoty počtov jednotlivých výskytov a rozdiely medzi očakávanými a skutočnými hodnotami sú použité ako základ pre výpočet štatistickej významovosti spoluvýskytu testovaných výrazov. Cieľom týchto výpočtov je určiť, či dôvodom častého spoluvýskytu slov testovaného výrazu je, že sa jedná o viacslovný výraz, alebo či ide len o štatisticky náhodný jav. Z toho

vyplýva, že tieto metódy nie je možné aplikovať na jednoslovné výrazy, najčastejšie sa aplikujú na bigramy a trigramy. Výsledkom výpočtu týchto metód je skóre, pričom vyššie skóre indikuje väčšiu pravdepodobnosť, že ide o združený výraz. Cieľom tejto kapitoly je predstavenie si rôznych spôsobov výpočtu skóre v bigramoch a trigramoch, ktoré budú neskôr porovnávané.

3.1 Metódy výpočtu miery asociácie v bigramoch

Existuje veľké množstvo metód výpočtu miery asociácie v bigramoch, ktorých princíp bude v tejto sekcii predstavený. Tieto algoritmy pracujú s nameranými hodnotami výskytov bigramov (viz tabuľka 3.1) a ich vypočítanými očakávanými hodnotami. Metóda spôsobu výpočtu je ilustrovaná nižšie [9].

	b	**not b**
a	$e_{11} = \frac{R1*C1}{N}$	$e_{12} = \frac{R1*C2}{N}$
not a	$e_{21} = \frac{R2*C1}{N}$	$e_{22} = \frac{R2*C2}{N}$

Tabuľka 3.3: Výpočty očakávaných hodnôt v bigramoch

3.1.1 Z-score

Z-score je algoritmus, ktorý napriek svojej jednoduchosti dosahuje veľmi dobré výsledky vo vyhľadávaní dvojslovných pomenovaní. K výpočtu potrebuje skutočný počet spoluvýskytov slov a očakávaný počet vypočítaný spôsobom popísaným v tabuľke 3.3. Algoritmus vypočíta skóre odvíjajúce sa od rozdielu medzi očakávanou a skutočne nameranou hodnotou výskytov.

$$z - score = \frac{o_{11} - e_{11}}{\sqrt{e_{11}}} \tag{3.1}$$

Z-score je možné vypočítať pre všetky bunky tabuľky, pre účely práce je však potrebná iba bunka o_{11}.

3.1.2 Z-score corrected

Z-score používa spojité normálne rozdelenie na aproximáciu diskrétneho binomického rozdelenia, čo môže viesť k jej nepresnosti. Na zvýšenie presnosti aproximácie bola navrhnutá úprava hodnoty skutočného počtu výskytov podľa nasledujúceho vzorca:

$$z - score_{corrected} = \frac{o_{11} - e_{11}}{\sqrt{e_{11}}} \begin{cases} o_{ij} - 0.5 & \text{if} \quad o_{ij} > e_{ij} \\ o_{ij} + 0.5 & \text{if} \quad o_{ij} < e_{ij} \end{cases} \tag{3.2}$$

3.1.3 T-score

T-score je alternatívou k Z-score a jeho cieľom je určenie, či asociácia medzi dvoma slovami je náhodná pomocou výpočtu kvocientu očakávaných a skutočných výskytov. Narozdiel od Z-score je menovateľom hodnota získaná z dát, nie očakávaná hodnota.

$$z - score = \frac{o_{11} - e_{11}}{\sqrt{o_{11}}} \tag{3.3}$$

3.1.4 Log-likelihood ratio

Log-likelihood meria rozdiel medzi očakávanými a dosiahnutými hodnotami. Je vypočítaná ako súčet pomerov medzi týmito hodnotami. Vzorec výpočtu je nasledovný:

$$Log - likelihood\ ratio = 2 * \sum_{ij} *log(\frac{o_{ij}}{e_{ij}}) \tag{3.4}$$

3.1.5 Logarithmic Odds Ratio

Táto metóda vráti pomer medzi počtom spoluvýskytov slov bigramu a počtom jednotlivých výskytov daných slov. Keďže v prípade nulovej početnosti samostatných výskytov slov bigramu nastane delenie nulou, bola navrhnutá modifikácia algoritmu, kde ku každej skutočnej hodnote výskytu je pripočítané 0.5.

$$Odds\ ratio_{disc} = log\frac{(o_{11} + 0.5)(o_{22} + 0.5)}{(o_{12} + 0.5)(o_{21} + 0.5)} \tag{3.5}$$

3.1.6 Dice coefficient, Jaccard coefficient

Tieto dve metódy sa často využívajú v technológiách získavania informácií a sú vypočítané nasledujúcimi vzoracmi:

$$Dice = \frac{2o_{11}}{R1 + C1} \tag{3.6}$$

$$Jaccard = \frac{o_{11}}{o_{11} + o_{12} + o_{21}} \tag{3.7}$$

3.1.7 Pointwise Mutual Information

Princíp výpočtu skóre tejto metódy spočíva v porovnávaní frekvencie výskytu skórovaného kandidáta a frekvencií výskytov slov z ktorých sa tento kandidát skladá. Môžeme to vidieť vo vzorci, kde o_{11} reprezentuje nameraný počet výskytov výrazu a e_{11} očakávaný počet získaný z počtu samostatných výskytov slov. Táto metóda má tendenciu zvýhodňovať kandidátov s malým počtom výskytov.

$$Pointwise\ MI = log(\frac{o_{11}}{e_{11}}) \tag{3.8}$$

3.1.8 Pearson's chi-squared test

Chi-squared test sa často využíva v štatistike a slúži k určení štatistickej významovosti nerovnomerného rozdelenia početností meraných hodnôt. V oblasti extrakcie výrazov táto metóda dosahuje najlepšie výsledky pre výber kľúčových výrazov. Metóda porovnáva rozdiely medzi nameranými hodnotami, a hodnotami, ktoré by boli dosiahnuté, keby nebola medzi slovami výrazu žiadna väzba. V teste musí byť definovaná nulová hypotéza, teda hypotéza ktorú sa snažíme testom vyvrátiť. V našom prípade je nulovou hypotézou "Medzi slovami daného bigramu neexistuje žiadny vzťah". Použitý bude vzorec "chi-squared homogeneity corrected", ktorý sa často využíva v aplikáciách.

$$\chi^2_{h,c} = \frac{N(|o_{11}o_{22} - o_{12}o_{21}| - \frac{N}{2})}{R1R2C1C2} \tag{3.9}$$

3.2 Metódy výpočtu miery asociácie v trigramoch

Keďže trigramy pracujú s väčším počtom kombinácií dát, je na ich výpočet vhodné menšie množstvo metód. Tabuľka frekvencií a takisto aj spôsob výpočtu očakávaných hodnôt sú zložitejšie, pretože pracujú s troma dimenziami. Formát tabuľky aj potrebné vzorce sú znázornené nižšie. V tejto sekcii budú predstavené štyri metódy výpočtu miery asociácie v trigramoch. [9].

$e_{111} = \frac{o_{1pp} * o_{p1p} * o_{pp1}}{(o_{ppp})^2}$	$e_{222} = \frac{o_{2pp} * o_{p2p} * o_{pp2}}{(o_{ppp})^2}$	
$e_{112} = \frac{o_{1pp} * o_{p1p} * o_{pp2}}{(o_{ppp})^2}$	$e_{121} = \frac{o_{1pp} * o_{p2p} * o_{pp1}}{(o_{ppp})^2}$	$e_{211} = \frac{o_{2pp} * o_{p1p} * o_{pp1}}{(o_{ppp})^2}$
$e_{122} = \frac{o_{1pp} * o_{p2p} * o_{pp2}}{(o_{ppp})^2}$	$e_{212} = \frac{o_{2pp} * o_{p1p} * o_{pp2}}{(o_{ppp})^2}$	$e_{221} = \frac{o_{2pp} * o_{p2p} * o_{pp1}}{(o_{ppp})^2}$

Tabuľka 3.4: Výpočty očakávaných hodnôt v trigramoch

3.2.1 Pointwise Mutual Information

Táto metóda pracuje na rovnakom princípe ako pri bigramoch, teda porovnáva frekvencie výskytov trigramu s frekvenciami jeho komponentov. PMI dosiahol pri testovaní v nórskom jazyku výrazne najlepšie výsledky.

$$Pointwise\ Mutual\ Information = log_2(\frac{o_{111}}{e_{111}}) \tag{3.10}$$

3.2.2 True Mutual Information

TMI určuje veľkosť rozdielu medzi očakávanými a skutočnými hodnotami výpočom váženého priemeru PMI všetkých očakávaných a dosiahnutých hodnôt.

$$Pointwise\ Mutual\ Information = \sum_{ijk}(\frac{o_{ijk}}{N})(log_2\frac{o_{ijk}}{e_{ijk}}) \tag{3.11}$$

3.2.3 Log-likelihood

Podobne ako v dvojslovných výrazoch táto metóda meria rozdiel medzi očakávanými a skutočnými hodnotami.

$$Log - likelihood = 2\sum_{ijk} o_{ijk} * log\frac{o_{ijk}}{e_{ijk}} \tag{3.12}$$

3.2.4 Poisson-Stirling

Posledným testovaným algoritmom je Poisson-Stirling. Táto metóda využíva Stirlingov vzorec na aproximáciu binomického rozdelenia pomocou Poissonovho rozdelenia. Vzorec je popísaný nižšie.

$$Poisson - Stirling = o_{111}(log(o_{111}) - log(e_{111}) - 1) \tag{3.13}$$

Kapitola 4

Lingvistické metódy skórovania kandidátov

4.1 Vymedzenie pojmov

Predošlá kapitola sa zaoberala štatistickými spôsobmi skórovania kandidátov, pričom hlavnou filozofiou týchto metód bolo určiť, či je daný výraz združeným pomenovaním. Nebola teda zisťovaná relevantnosť výrazov v rámci kontextu textu, boli len vybrané výrazy, ktorých vlastnosti môžu naznačovať, že ide o kľúčové výrazy. Druhou fázou extrakcie kľúčových slov je teda skúmanie kandidátov z hľadiska relevantnosti v rámci textu, prípadne väčšieho korpusu. Používa sa pojem "termhood", ktorý je definovaný ako "stupeň, ktorým je stabilná lexikálna jednotka súvisiaca s doménovo špecifickými konceptmi" [12]. Unithood a termhood sú rozdielnými vlastnosťami výrazov, a vysoké skóre jedného nemusí implikovať vysoké skóre druhého. Napríklad výrazy ako "brať ohľad" majú vysokú mieru unithood, pretože slová tohto výrazu sa často vyskytujú spoločne, nemôžeme ich však považovať za kľúčové výrazy.

Cieľom tejto kapitoly bude skúmanie a priblíženie jednotlivých lingvistických metód výpočtu skóre kandidátov. Tieto metódy využívajú vlastnosti kľúčových slov vyplývajúce z ich kontextu, konkrétne vyššia frekvencia výskytov kľúčových výrazov v texte, výskyt jednotlivých slov výrazu v iných kľúčových výrazoch a nižší počet výskytov kľúčových slov v obecnom korpuse v porovnaní s bežnejšími slovami.

4.2 TF-IDF

Jeden z najznámejších a najpoužívanejších algoritmov pre selekciu kľúčových výrazov je miera TF-IDF [14]. Celý názov algoritmu je Term frequency-Inverse document frequency, ktorý označuje názvy dvoch metrík výpočtu skóre výrazu. Skóre sa počíta v algoritme pre jednotlivé slová testovaného výrazu, preto je potrebné na konci výsledky jednotlivých slov pripočítať. Prvým zohľadňovaným faktorom pri výpočte skóre je Term frequency, teda frekvencia slova. Táto hodnota určuje počet výskytov daného slova v texte, väčší rozsah textu bude mať za následok vyššiu hodnotu TF, preto je táto hodnota normalizovaná podľa celkového počtu slov v texte. Vzorec pre výpočet TF:

$$TF(t) = \frac{\text{Počet výskytov výrazu t v dokumente}}{\text{Celkový počet výrazov v dokumente}} \tag{4.1}$$

Druhou metrikou pri výpočte skóre je Inverse document frequency, ku ktorej výpočtu potrebujeme mať vytvorené štatistiky z veľkého počtu článkov a publikácií. Konkrétne potrebujeme vedieť, v koľkých textoch nachádzajúcich sa v korpuse sa vyskytuje dané slovo. Hlavnou myšlienkou tejto metriky je, že kľúčové slovo je využívané len v špecifickej tematickej oblasti, preto počet publikácií, v ktorých sa toto slovo nachádza bude relatívne malý. Z toho vyplýva, že kvalita algoritmu bude ovplyvnená počtom zhromaždených článkov a rôznorodosťou ich tém. Potrebné dáta o všetkých slovách z korpusu sú vypočítané ešte pred použitím TF-IDF algoritmu a obsahujú štatistiky pre každé slovo o počte dokumentov s výskytom daného slova. Vzorec pre výpočet IDF je nasledovný:

$$IDF(t) = log_e \frac{\text{Celkový počet dokumentov}}{\text{Počet dokumentov obsahujúcich výraz t}} \tag{4.2}$$

Cieľom algoritmu je teda nájsť slová, ktoré sa mnoho krát vyskytujú v danom texte, a zároveň ide o málo používané slová, ktoré sa mimo daného textu využívajú len zriedka. Teda slová, ktoré sa v texte vyskytujú častokrát, ako napríklad predložky a spojky, dostanú nízke skóre kvôli ich všeobecnej vysokej frekvencii, avšak slová, ktoré nie sú v texte najfrekventovanejšie, no ich použitie je v obecnom korpuse zriedkavé budú mať skóre vyššie. Celkový vzorec pre výpočet je teda nasledovný:

$$TF - IDF = TF(t) * IDF(t) \tag{4.3}$$

4.3 KP-Miner

KP-Miner je metóda založená na algoritme TF-IDF, rozširuje ho však o možnosť extrakie viacslovných kľúčových výrazov [4]. Prvou fázou algoritmu je výber kandidátnych výrazov z textu, pričom výrazy sú oddelené na základe interpunkčných znamienok a stop slov. Pre elimináciu prebytočných výrazov sú aplikované dve ďalšie kritéria, z ktorých prvým je minimálny počet výskytov v texte, ktorý je stanovený na dva v anglických textoch. Druhým kritériom je povinnosť výskytu daného výrazu v úvodných častiach textu, pričom dĺžka tejto sekcie je určená experimentálne. Táto podmienka však nemusí byť vhodná pre všetky typy textov. Druhou fázou algoritmu je výpočet skóre, ktorý je s istými obmenami založený na algoritme TF-IDF. Prvou zmenou oproti tomuto algoritmu je maximálna hodnota IDF u viacslovných výrazoch je limitovaná na číslo jedna. Teda v prípade hodnôt jedna a vyšších bude braná ako by sa vyskytovala len jeden krát. Dôvodom je neexistencia korelácie medzi nízkym počtom výskytov v obecnom korpuse a kľúčovosťou výrazu. Keďže výrazy iných dĺžok dosahujú častokrat iné rozsazy hodnôt TF-IDF, bola pridaná konštanta pre dosiahnutie porovnateľnoti týchto hodnôt. Posledným nepovinným krokom algoritmu je prepočítanie výskytových štatistík ktoré berie do úvahy prekrývanie výrazov, pričom tento krok nemusí mať vždy za dôsledok lepšie dosiahnuté výsledky.

4.4 KX-FBK

Ďalším algoritmom pre výber kľúčových výrazov je KX-FBK, ktorý na skórovanie využíva korpusové data [13]. Prvým krokom algoritmu je výber všetkých n-gramov z korpusu spolu s ich počtom výskytov. N-gramy vyskytujúce sa primálo krát sú odstránené pretože nie sú považované za dôležité. V druhom kroku sú selektované len n-gramy spĺňajúce predom určené lexikálne vzory. Následne je vypočítaná hodnota IDF, je však počítaná len pre výrazy spĺňujúce podmienky mnimálneho počtu výskytov v korpuse a vo vstupnom dokumente.

V poslednom štvrtom kroku sú z dokumentu extrahované viacslovné výrazy spĺňajúce podmienku minimálneho počtu výskytov a ich skóre je vypočítané na základe rôznych heurestík. Tieto heurestiky sú hodnota IDF, dĺžka výrazu, kde sú preferované dlhšie výrazy kvôli vyššej špecifickosti, pozícia prvého výskytu keďže dôležitejšie výrazy bývajú často spomenuté skôr. Poslednou heurestikou je znižovanie skóre kratších a zvyšovanie skóre dlhších konceptov, čo znamená, že tie selektované výrazy, ktoré sú súčasťou dlhších selektovaných výrazov majú znížené skóre, a dlhšie výrazy majú skóre zvýšené.

4.5 RAKE

Algoritmus RAKE dosahuje veľmi dobré výsledky pri extrakcii viacslovných kľúčových výrazov. V predošlej kapitole bol popísaný spôsob výberu kandidátov tohto algoritmu, a v tejto bude priblížený spôsob skórovania. RAKE pracuje s jednotlivými slovami skórovaného výrazu, preto prvým krokom býva rozdelenie výrazu na slová. Skórovací proces počíta tematickú podobnosť jednotlivých kandidátov a na výpočet potrebuje dve štatistiky skórovaného slova: frekvencia slova a stupeň slova. Frekvencia slova, zapisovaná ako freq(w), kde w je testované slovo označuje celkový počet výskytov daného slova v zozname kandidátov. Druhou štatistikou je stupeň slova, zapisovaný ako deg(w), ktorý označuje súčet počtu slov vo výrazoch, v ktorých sa slovo w vyskytuje. Teda napríklad, ak testované slovo je "diskriminant" a toto slovo sa nachádza v kandidáte "diskriminant kvadratickej rovnice", bude k stupňu daného slova pripočítané číslo tri, pretože má daný kandidát dĺžku tri slová. Po získaní hodnôt frekvencie a stupňa testovaného slova je možný výpočet samotného skóre podľa nasledujúceho vzorca [15]:

$$word\ score = deg(w)/freq(w) \tag{4.4}$$

Po vypočítaní skóre všetkých slov, ktoré sú súčasťou kandidátov spočítame súčet jednotlivých slov skórovaného kandidáta. Ako bolo vyššie popísané, je možné vidieť, že algoritmus uprednostňuje výrazy skladajúce sa z väčšieho počtu slov. Nakoniec sú kandidáti zoradení podľa skóre od najvyššieho po najnižší, a prípadne je vybrané len požadované množstvo kandidátov.

Kapitola 5

Metódy skórovania založené na grafoch

Hlavnou myšlienkou algoritmov pre výber kľúčových výrazov založených na grafoch je nájdenie a reprezentácia vzťahov medzi jazykovými entitami [11]. Tieto entity môžu byť slová, kolokácie, alebo celé vety, v závislosti od požadovanej funkcie algoritmu. Jazykové entity sú uzlami v grafe, a hrany reprezentujú vzťah medzi jednotlivými entitami. Väčší počet uzlov vedúcich k danej entite spravidla znamená jej väčšiu mieru dôležitosti v grafe. Hrany môžu byť orientované aj neorientované, pričom tradičnejšie sa v grafovo založených skórovacích algoritmoch využívajú orientované grafy. Hrany môžu byť takisto vážené aj neválené, pričom váha slúži na reprezetáciu sily vzťahu medzi uzlami, a prejaví sa pri výpočte skóre. Skóre je vypočítané pre každý uzol a indikuje dôležitosť uzlu v rámci grafu. Začínajúc so základných hodnôt pridelených jednotlivým uzlom sa skóre počíta vo viacerých iteráciách, kým dôjde k jeho konvergencii. Existuje veľké množstvo grafovo zložených skórovacích algoritmov, pričom niektoré z nich budú v tejto časti predstavené.

5.1 TextRank

Prvou metódou ktorá použila grafickú reprezentáciu textu pre výber kľúčových výrazov je TextRank [11]. Uzlami v grafe sú slová z textu, ktoré boli otagované a prešli syntaktickými filtrami. Pre prevenciu prílišného zvyšovania veľkosti grafu sú uzlami len slová, n-gramy sú rekonštruované po ukončení výpočtu skóre. Jedná sa o neorientovaný graf, pričom hrany medzi slovami sú pridané na základe spoluvýskytu daných slov v texte. Maximálna vzdialenosť spoluvýskytu je nastaviteľná od dvoch do desiatich slov. Po vytvorení grafu je všetkým uzlom pridelené skóre jedna a vo viacerých iteráciách je spustený skórovací algoritmus PageRank, vytvorený spoločnosťou Google, kým nedôjde ku konvergencii skóre, pričom obvykle ide o 20-30 iterácií. Nakoniec je vybraná tretina najlepšie skórujúcich slov, a sekvencie po sebe nasledujúcich slov sú spojené do viacslovných výrazov.

5.2 TopicRank

Druhou popísanou metódou je TopicRank, ktorá je považovaná za vylepšenie metódy TextRank [2]. Prvou zásadnou zmenou oproti algoritmu TextRank je uzlami vo vytvorenom grafe nie sú slová, ale témy. Hlavnou myšlienkou tejto zmeny je, že väčšina tém popísaných

v skórovanom dokumente je vyjadriteľná viac ako jedným kľúčovým výrazom, čo má za dôsledok redundanciu niektorých kľúčových výrazov. V tejto metóde je teda skupina výrazov reprezentovaná ako jedná entita, teda téma, z ktorej bude na konci vybraný najvhodnejší výraz, ktorý túto tému reprezentuje. Výrazy sú považované za podobné ak majú aspoň 25 percent spoločných slov. Na automatické triedenie výrazov bol použitý algoritmus HAC. Druhom zmenou oproti algoritmu TextRank je, že graf obsahuje vážené hrany, pričom váha reprezentuje silu sémantického vzťahu medzi uzlami vyplývajúcu z ich vzdialenosti výskytu. Následne je vypočítané skóre rovnakým spôsobom ako v algoritme TextRank a z vybraných tém je selektovaný najvhodnejší výraz podľa počtu výskytov alebo miery podobnosti kandidátov.

5.3 SingleRank

Ďalšou modifikáciou algoritmu TextRank je metóda SingleRank [19]. Prvou zmenou oproti TextRanku je, že jednotlivé hrany grafu sú vážené, pričom ich váha reprezentuje počet spoluvýskytov daných slov. SingleRank takisto nevyberá len najlepšie skórujúce slová, ale skóre je vypočítané ako súčet čiastkových skóre jednotlivých slov daného výrazu. Okno spoluvýskytu je vždy nastavené na hodnotu 10.

5.4 ExpandRank

ExpandRank je vylepšenie metódy SingleRank, ktoré k výpočtu skóre využíva informácie z iných dokumentov [19]. Prvým krokom je teda zhotovenie skupiny dokumentov na základe zisťovania podobnosti medzi ohodnocovaným dokumentom a inými dokumentami z korpusu. Dokumenty sú reprezentované ako vektory výrazov, hodnoty jednotlivých dimenzií sú vypočítané pomocou TF-IDF a ich vhodnosť ohodnotená podľa kosínovej podobnosti. Táto hodnota podobnosti je použitá ako miera vierohodnosti slov získaných z daného dokumentu. Na slová získané z vytvorenej kolekcie dokumentov sú použité syntaktické filtre, a z vyhovujúcich slov je vytvorený graf. Zvyšok priebehu je identický s algoritmom SingleRank, pričom na výpočte váhy sa tiež podieľa miera vierohodnosti.

5.5 SGRank

SGRank je hybridný algoitmus, ktorý okrem grafovej reprezentácie textu využíva linguistické a iné metódy výpočtu skóre [3]. Prvým krokom je extrakcia n-gramov a filtrácia tých, ktorí majú nizku pravdepodobnosť byť kľúčovými výrazmi, podľa stop slov, interpunkcie a POS tagov. Ostávajúcim výrazom je vypočítané skóre pomocou upravenej verzie TF-IDF použitej v KP-Miner. Modifikácia spočíva v tom, že pri viacslovných výrazoch môže IDF dosiahnúť maximálne hodnotu jedna, pretože pri týchto výrazoch nie je nízky počet výskytov v korpuse indikátorom kľúčovosti. Je vybraný určitý počet najlepšie skórovaných výrazov, ktorých skóre je prepočítané na základe ďalších heurestík, konkrétne pozícia prvého výskytu, dĺžka výrazu a počet vybraných výrazov, ktoré obsahujú skórovaný výraz. Je vypočítaná váha jednotlivých výrazov, pričom v prípade jej pozitívej hodnoty je výraz testovaný v poslednej, štvrtej fáze algoritmu. Zo zvyšných výrazov je vytvorený graf s váženými hranami, pričom váha je počítaná na základe vzdialenosti všetkých spoluvýskytov v kombinácii s váhou vypočítanou v predošlom kroku. Konvergujúce skóre grafu označuje kľúčovosť jednotlivých výrazov a je finálnym výstupom algoritmu.

Kapitola 6

Spôsoby vyhodnocovania výsledkov

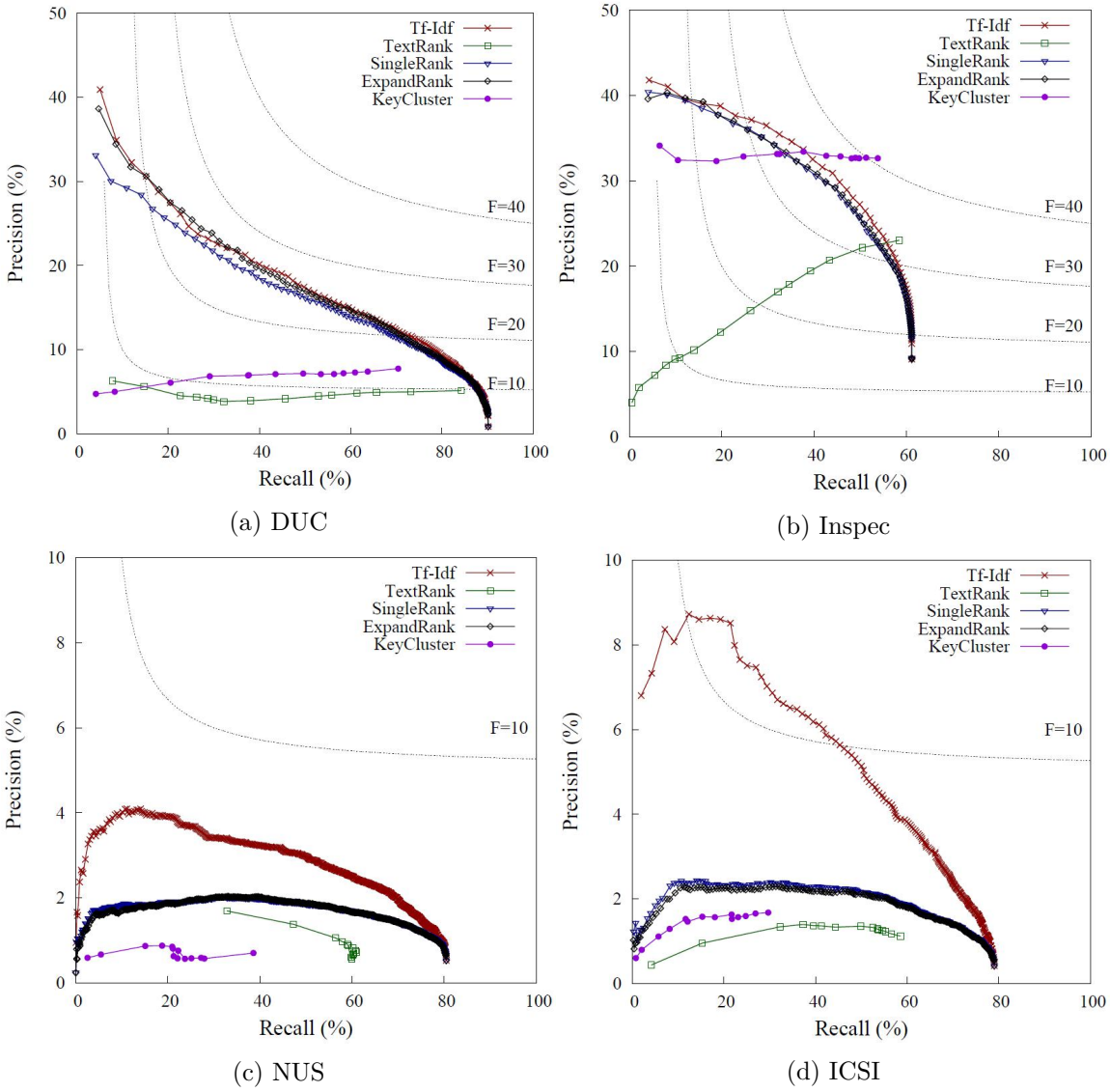
V predošlých kapitolách boli predstavené algoritmy a metódy pre výber kľúčových výrazov z odborných textov. Po ich implementácii bude posledným krokom tejto práce zistiť ich použiteľnosť a porovnať ich výsledky v českom jazyku. Kvalitu kľúčového výrazu nie je možné objektívne posúdiť, ani kvantifikovať, jeho vhodnosť je možné zistiť na základe názoru experta, ktorý je s textom oboznámený, a rozumie popisovaným súvislostiam. Názory expertov sa však častokrát v mnohých ohľadoch budú líšiť, preto pravdepodobne najlepšou metrikou vhodnosti slov je pomer súhlasov k nesúhlasom expertov k danému výrazu. Táto metóda je ale nepraktická a málo využívaná, častejšie sa používajú metódy, ktoré porovnávajú manuálne vybrané a automaticky generované kľúčové výrazy [17]. Z toho vyplýva, že zhromaždené texty, ktoré budú použité pri testovaní, budú musieť mať manuálne vytvorený zoznam kľúčových výrazov. Nasledujúca časť predstaví jednotlivé metódy skórovania algoritmov [10].

6.1 Presnosť

Prvá miera vyhodnocovania sa označuje pojmom “precision”, po slovensky presnosť. Presnosť je vyjadrená percentuálne a počíta sa ako pomer správnych extrahovaných výrazov ku všetkým extrahovaným výrazom. Vyššia hodnota presnosti je dosiahnutá prísnejšími podmienkami pre extrakciu a menším počtom algoritmom vybraných výrazov, pretože algoritmus má v tomto prípade tendenciu vybrať tie výrazy, u ktorých je väčšia istota, že ide k kľúčové slová. Vysoká hodnota presnosti je preferovaná v prípadoch, v ktorých je pre užívateľa dôležitejší nízky počet falošne pozitívnych výsledkov, a jeho cieľom nie je získať čo najväčší počet kľúčových výrazov.

6.2 Úplnosť

Druhá miera vyhodnocovania sa nazýva “recall”, po slovensky úplnosť, a reprezentuje pomer správnych extrahovaných výrazov ku všetkým manuálne vybraným výrazom. Táto metrika poukazuje na úroveň pokrytia pri extrahovaní kľúčových výrazov. Lepšie výsledky sú dosiahnuté v prípade zvoľnenia požiadavkov, ktoré musia kľúčové slová spĺňať, čo má za následok vyšší počet extrahovaných kľúčových slov. Na druhú stranu sa to však negatívne prejavuje na počte falošne pozitívnych vybraných kľúčových výrazov. Typický vzťah medzi presnosťou a úplnosťou je reprezentovaný v nasledujúcich grafoch:



(a) DUC (b) Inspec

(c) NUS (d) ICSI

Obr. 6.1: Vzťah medzi hodnotami presnosti a úplnosťou pri rôznych metódach extrakcie, prevzané z [6]

Tieto grafy reprezentujú skutočný vzťah presnosti a úplnosti rôznych algoritmov na štyroch korpusoch [6]. Vidíme, že vo väčšine prípadov platí medzi hodnotami úplnosti a presnosti nepriama úmernosť, a teda nie je možné dosiahnuť dobré výsledky v obidvoch kritériách. Vysoká presnosť znižuje počet nájdených výrazov, čo má negatívny dopad na úplnosť, a veľký počet nájdených správnych výrazov znižuje hodnotu presnosti z dôvodu priveľkého počtu falošne pozitívnych výrazov. Funkcia algoritmu teda bude závisieť od rozhodnutie užívateľa, ktorá metrika je preňho dôležitejšia, pričom v praxi je častejšie preferovaná presnosť.

6.3 Obmedzenia navrhnutého vyhodnocovania

Táto práca, aj väčšina súvisiacich vedeckých publikácií vyhodnocuje výsledky pomocou postupov popísaných vyššie. Základom týchto postupov je porovnávanie automaticky vy-

generovaných kľúčových výrazov s výrazmi manuálne vybranými, čo však so sebou prináša určité obmedzenia. Algoritmom vybraný výraz je považovaný za správny len v prípade, že sa nachádza len v zozname manuálne vybraných výrazov. Tieto výrazy sú však selektované zo subjektívneho hľadiska a absencia výrazu v tomto zozname nemusí znamenať, že nejde o kľúčový výraz. Tiež je možné, že sú výrazy napísané iným slovosledom alebo pomocou synoným (napr. Pacifik - Tichý oceán). Preto je pravdepodobné, že bude kvalita výsledkov testovaných algoritmov podceňovaná a výsledky budú horšie, ako keby boli manuálne vyhodnocované ľuďmi.

Kapitola 7

Príprava testovacích dát

Cieľom tejto práce je testovať a porovnať rôzne spôsoby výberu kľúčových slov, teda prvým krokom je zhromaždenie dostatočného množstva článkov a publikácií, na ktorých tieto algoritmy budú testované. Aby boli texty použiteľné musia spĺňať dve podmienky: musí ísť o vedecké texty, ktoré využívajú odbornú terminológiu a musia mať manuálne vytvorený zoznam kľúčových slov. V tejto kapitole bude popísaný proces zhromažďovania a prípray testovacích dát.

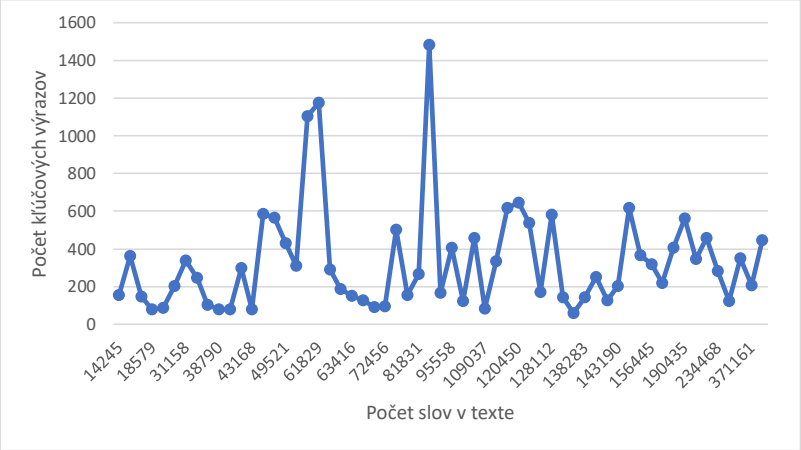
7.1 Proces výberu relevantných publikácií

Ako zdroj publikácií na testovanie programu bol použitý korpus fakulty informačných technológii VUT. Ide o preskenované fyzické kópie kníh, ktorých znaky boli rozoznané pomocou OCR technológie, a výsledné pdf súbory boli prekonvertované do podoby txt. Jedná sa o knihy rôznych žánrov, teda nie všetky z týchto publikácií sú vhodné pre účel projektu, čo znamená potrebu ich automatickej filtrácie a neskôr manuálneho výberu.

Na základe metadát boli vybrané publikácie, ktoré sú odborného žánru, z ktorých boli vybrané knihy obsahujúce register pomocou vyhľadania slova "rejstřík" v posledných desiatich percentách textu. Toto bol limit automatického filtrovania, a teda bolo potrebné manuálne zhodnotiť použiteľnosť jednotlivých kníh. Pri manuálnom výbere boli vybrané tie publikácie, ktoré boli skutočne odborného žánru a obsahovali register kľúčových výrazov. Takisto boli vyradené publikácie obsahujúce priveľký počet nečitateľných znakov alebo rozhádzanú štruktúru textu. Zo zhromaždených publikácií bol extrahovaný register, ktorý bol následne upravený do požadovanej podoby. Z textu boli takisto odstránené nepotrebné súčasti ako obsah a zoznam autorov a referencií.

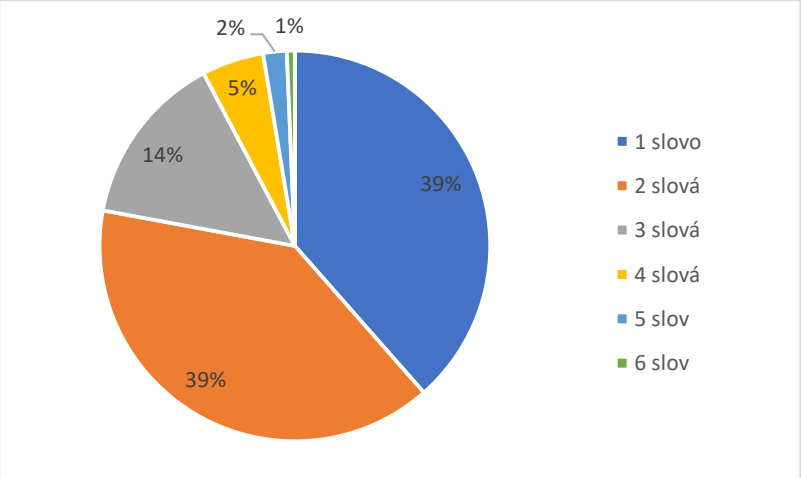
7.2 Analýza získaných textov

Zhromaždené publikácie majú veľkú variabilitu v dĺžke textu, kde najkratší text má 14245 a najdlhší 387063 slov, pričom mediánová dĺžka textu je 91969 slov. Texty sa tiež výrazne líšia v počte kľúčových výrazov, pričom najmenej je 59 a najviac 1482 výrazov, a medián je 265 výrazov. Nasledujúci graf reprezentuje vzťah medzi dĺžkou textu a počtom manuálne vybraných kľúčových slov.



Obr. 7.1: Závislosť dĺžky textu od počtu kľúčových výrazov

Cieľom grafu je zistiť existenciu korelácie medzi dĺžkou textu a počtom kľúčových slov, ktoré text obsahuje. Tieto dáta budú neskôr využiteľné pri určovaní aký počet kľúčových výrazov je najvhodnejší pre danú dĺžku textu. Na grafe však vidíme, že sa hypotéza nepotvrdila, čo znamená, že variabilita počtu kľúčových výrazov nesúvisí s dĺžkou textu. Túto variabilitu je možné vysvetliť inými, ťažšie klasifikovanými vlastnosťami, ako sú vedecké odvetvie publikácie a osobný štýl tvorcu registru. Metódy extrakcie kľúčových výrazov sú častokrát viazané k požadovanej dĺžke daných výrazov, preto je dôležité poznať najčastejšie početnosti slov kľúčových výrazov, čo je ukázané v nasledujúcom grafe.



Obr. 7.2: Najčastejšie počty slov v kľúčových výrazoch

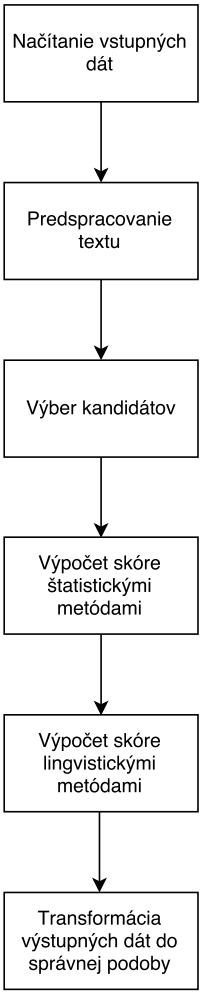
Vidíme, že v grafe výrazne prevažujú jedno, dvoj a trojslovné výrazy, ktoré dokopy zahŕňajú viac ako 90 percent výrazov. Nazbierané dáta sú odzrkadlené aj v popísaných spôsoboch automatickej extrakcie kľúčových výrazov, ktoré sa zameriavajú práve na jedno, dvoj a trojslovné výrazy.

Kapitola 8

Návrh a implementácia programu

8.1 Návrh aplikácie

Aplikácia je rozdelená na nasledujúce logické celky, pričom tieto celky budú implementované všetkými spôsobmi popísanými v kapitolách 3 a 4. Vo finálnej verzii aplikácie sa však budú anchádzať len najlepšie fungujúce algoritmy.



Obr. 8.1: Návrh programu

8.1.1 Načítanie vstupných dát

Program prijíma ako vstupné dáta textový súbor obsahujúci český text, z ktorého budú neskôr extrahované kľúčové výrazy.

8.1.2 Predspracovanie textu

Pred samotnou analýzou je potrebné text previesť do vhodnejšej podoby, čo bude dosiahnuté použitím modulu Morphodita. Text bude lematizovaný pre lepšie vyhľadávanie frekvencií výrazov a tagovaný pre výber kandidátov. Taktiež prebehne tokenizácia pre vytvorenie párov lematizovanej a nelematizovanej formy jednotlivých slov.

8.1.3 Výber kandidátov

Ďalším krokom je využitie jedného z vyššie popísaných algoritmov na identifikáciu a extrahovanie kandidátov z predspracovaného textu. Budú testované obidva spôsoby popísané v kapitole 2, pričom lepší z nich bude aplikovaný vo finálnej verzii programu. Výsledkom bude zoznam vybraných kandidátnych výrazov, pričom jednotliví kandidáti budú obsahovať informáciu o počte výskytov, čo bude potrebné v nasledujúcich častiach programu.

8.1.4 Použitie štatistických prístupov k výberu najvhodnejších kandidátov

Zoznam kandidátov získaný v predošlom kroku bude ďalej filtrovaný pomocou jednej z metód popísaných v kapitole 3. Touto metódou bude daným kandidátom vypočítané skóre, podľa ktorého budú neskôr zoradení. Na základe testovania a analýzy bude vybraná najvhodnejšia metóda, ktorej vypočítaná výška skóre najviac koreluje so správnosťou výrazu. Skóre bude vypočítané pre výrazy získané v predošlom kroku, a na tento výpočet sú potrebné dáta o početnosti bigramov a trigramov v danom texte.

8.1.5 Výpočet skóre kandidátov pomocou lingvistických metód

Skóre jednotlivých kandidátov bude takisto počítané jednou z lingvistických metód popísaných v kapitole 4. Po vykonaní algoritmu budú kandidáti zoradení podľa skóre, a určitý počet najlepšie skórujúcich z nich bude vybraný ako kľúčové výrazy pre daný text. Pomocou analýzy a vzájomného porovnávania bude vybraný najvhodnejší algoritmus, pričom finálne skóre môže byť získané kombináciou štatistickej a linguistickej metódy. Požadovaný počet zoradených výrazov bude vybraný ako zoznam kľúčových výrazov, ktoré budú následne prekonvertované do správnej formy.

8.1.6 Vytvorenie správnej formy výstupných dát

Po vykonaní algoritmov v predošlých krokov bol získaný zoznam automaticky vygenerovaných kľúčových výrazov, ktorý je očakávanými výstupovými dátami tohto programu. Tieto dáta však zatiaľ nemajú správnu formu pre výstup, pretože následkom činnosti extrahovacích algoritmov sú lematizované a obsahujú len malé písmená. Proces odstránenia lematizácie bude vykonaný knižnicou Morphodita, konkrétne jej metódami pre morfologickú generáciu. Tieto metódy však potrebujú POS tag popisujúci dané slovo v požadovanom tvare. Niektoré z týchto informácií je možné získať zo základných pravidiel vyplývajúcich z kombinácií POS tagov, napríklad u dvojice prídavné meno - podstatné meno je tvar prídavného

mena odvoditeľný z rodu podstatného mena. U iných výrazoch došlo k strate kontextových informácií, kvôli čomu nie je možné deterministicky určiť správnu formu výrazov. Príkladom je dvojica podstatné meno - podstatné meno, u ktorých môže ísť o genitívnu väzbu, alebo obidve slová môžu byť v nominatíve. Riešením je nájdenie nelematizovanej formy týchto výrazov v zdrojovom texte, čo je možné pomocou tokenizácie a vytvorenia korešpondujúcich dvojíc lematizovaných a nelematizovaných foriem slov. Tento výraz nemusí byť v základnom tvare, avšak kontextový vzťah jeho slov bude zachovaný.

8.2 Morphodita

Morphodita je modul, ktorý bol vyvinutý Karlovou Univerzitou pre prácu s jazykom v českých textoch [16]. Jeho funkcionalita zahŕňa morfologickú analýzu a generáciu, tagovanie, tokenizáciu a lematizáciu. Prvým z využití Morphodity je predspracovanie textu pomocou lematizácie, na dosiahnutie efektívnejšieho výberu kandidátov na základe ich frekvencie. Jedna z vyššie popísaných metód využíva na extrahovanie kanditátov POS tagy, na čo bude využitá tagovacia funkcionalita Morphodity. Pri lematizácii budú pre jednotlivé slová stratené niektoré kontextové informácie, ako sú pád a číslo, ktoré sú potrebné pri morfologickej generácii. Z toho dôvodu je potrebné získať nelematizovanú formu slov tak, ako sa nachádzala v texte, čo je možné s využitím tokenizácie textu a výberu slova nachádzajúceho sa na rovnakej pozícii. Výsledné kľúčové slová bude podľa vytvorených pravidiel potrebné previesť do ich nelematizovanej formy, k čomu bude využitá morfologická generácia slov Morphodity.

8.3 Testované algoritmy

V predošlých kapitolách bolo predstavené veľké množstvo algoritmov na vyhľadávanie kľúčových výrazov, z ktorých niektoré budú v nasledujúcich častiach testované a porovnávané. Pri výbere kandidátov budú testované obidve metódy, pretože k svojej funkcii využívajú gramatické vlastnosti jazyka a je teda potrebné porovnať ich úspešnosť na českom jazyku. Bolo predstavené veľké množstvo štatistických metód, pričom testované budú tie, ktoré dosiahli v zdrojovej publikácii najlepšie výsledky [9]. Konkrétne ide u bigramoch o metódy Z-score, Z-score corrected, Logarithmic Odds Ratio, Pointwise Mutual Information a Pearson's Chi-squared test, a u trigramoch Pointwise Mutual Information a Poisson-Stirling. Z lingvistických metód budú testované TF-IDF a RAKE, keďže ide o univerzálnejšie metódy vhodné pre rôzne typy textov. Naopak algoritmy KP-Miner a KX-FBK sú málo flexibilné a ich cieľom je dosiahnúť čo najlepšie výsledky u krátkych abstraktov, preto nie sú pre túto prácu vhodné. Metódy založené na grafoch dosahujú vo väčšine prípadov výsledky horšie ako TF-IDF [6], a mnohé z nich obsahujú časti, ktoré zťažšujú implementáciu a testovanie v českom jazyku. Konkrétne ide o potrebu mať korpus obsahujúci veľký počet dokumentov, nutnosť podporných modulov (na určovanie podobnosti výrazov), horšie možnosti skórovania viacslovných výrazov, a motivácia extrakcie prevažne z kratších textov. Z týchto dôvodov nie sú grafovo založené algoritmy testované.

Kapitola 9

Analýza a porovnanie prístupov výberu kandidátov

9.1 Výber kandidátov na základe stop slov

Prvou testovanou metódou je algoritmus, ktorý rozdeľuje vety na menšie časti na základe stop slov a z týchto častí vyberá kandidátov podľa povoleného počtu slov a počtu výskytov. Táto metóda pochádza z algoritmu RAKE a je bližšie popísaná v kapitole 2. K svojej funkcii potrebuje mať prístup k zoznamu českých stop slov. Zdrojový kód bol prevzatý z [1] a upravený pre vlastnú potrebu.

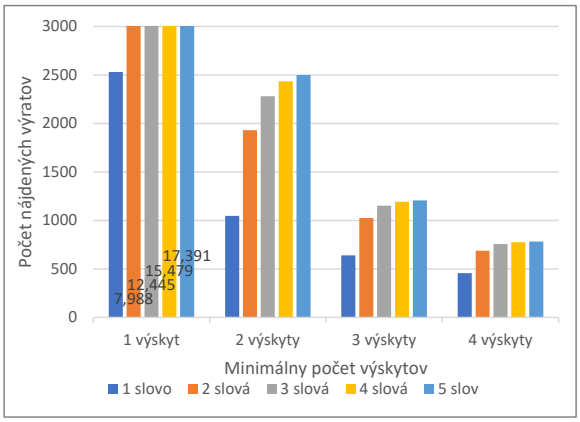
9.1.1 Zoznam stop slov

Stop slová je možné charakterizovať ako slová, ktoré samy o sebe majú veľmi malú výpovednú hodnotu, ide však o slová, ktoré slúžia na vyjadrenie gramatického vzťahu medzi jednotlivými vetnými konštrukciami, a ich výskyt v texte je relatívne častý. Stop slová majú široké využitie v oblasti spracovania prirodzeného jazyka (natural language proccessing), vďaka čomu sú verejne prístupné zoznamy českých stop slov. Zo snahy získať čo najväčší počet stop slov pre prácu algoritmu boli spojené viaceré verejne prístupné zoznamy týchto slov. Zoznam nakoniec obsahoval 405 stop slov.

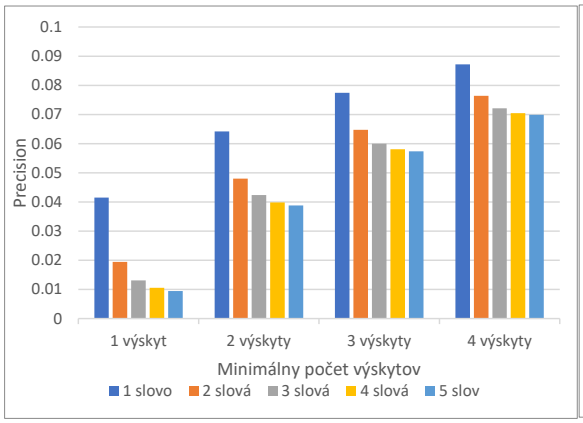
9.1.2 Analýza výsledkov algoritmu

Vplyv parametrov na dosiahnuté výsledky algoritmu

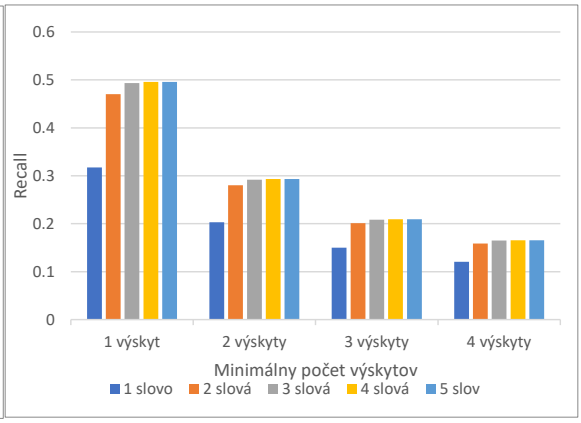
Táto časť obsahuje výsledky analýzy fungovania algoritmu, pričom oblasti záujmu sú počet nájdených výrazov, počet správne identifikovaných výrazov a počet falošne pozitívnych výrazov. Je dôležité podotknúť, že algoritmus pracuje s parametrami maximálny počet slov vo výraze a minimálny počet výskytov výrazu, ktorých hodnoty s najvyššou pravdepodobnosťou ovplyvnia dosiahnuté výsledky, ktoré je potrebné porovnať. Nasledujúce grafy ukazujú počty extrahovaných výrazov a hodnoty úplnosti a presnosti pre jednotlivé kombinácie parametrov.



(a) Počet nájdených výrazov



(b) Presnosť



(c) Úplnosť

Obr. 9.1: Dosiahnuté výsledky s rôznymi hodnotami parametrov

V prvom grafe môžeme vidieť priemerné počty extrahovaných výrazov z jednotlivých textov vzhľadom na parametre algoritmu. Zvyšujúci sa minimálny požadovaný počet výskytov výrazu sa podľa očakávaní odzrkadľuje na znižujúcom sa počte nájdených výrazov, pričom najvýraznejšia zmena prebehla medzi jedným a dvoma výskytmi, kde sa počet znížil o viac ako 75 percent. Takisto je v grafe zaznamenaná závislosť zmeny maximálneho povoleného počtu slov vo výraze od počtu nájdených výrazov. Graf ukazuje, že najvýraznejšie rozdiely v získaných výrazoch sa nachádzajú medzi jedno, dvoj a trojslovnou maximálnou dĺžkou výrazov a pri väčších hodnotách daného parametra sú tieto rozdiely výrazne nižšie. Môže to byť vysvetlené tým, že výrazy nachádzajúce sa v texte sú väčšinou jedno, dvoj alebo trojslovné, a výrazy iných dĺžok sa v texte vyskytujú zriedkavejšie.

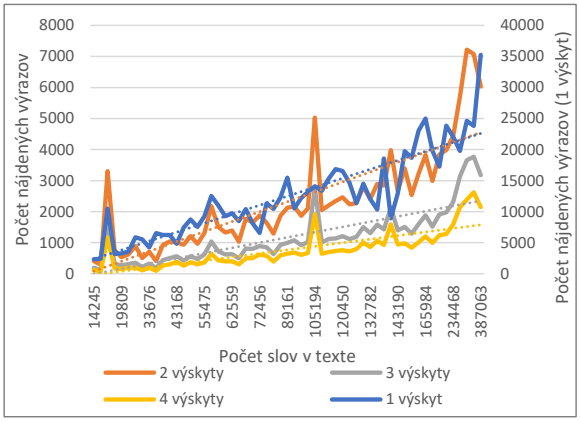
V zvyšných dvoch grafoch vidíme, ako sú ovplyvnené hodnoty presnosti a úplnosti v závislosti na rôznych kombináciách parametrov. Graf presnosti ukazuje, že zvyšovanie minimálneho počtu výskytov má za následok zlepšenie hodnôt presnosti. Z toho môžeme usúdiť, že medzi vyšším počtom výskytov a kľúčovosťou výrazu existuje určitá miera korelácie. Presnosť však celkovo nedosahuje veľmi dobré výskedky, v najlepšom prípade ide o hodnotu tesne pod 10 percent, avšak v tejto fáze algoritmu to nemusí byť problém, pretože dodatočné vylučovanie výrazov nastane v nadchádzajúcich fázach. Vidíme tiež, že hodnoty presnosti mierne klesajú pri zvýšení minimálneho počtu slov vo výraze, z čoho vyplýva, že u väčšiny týchto štyri a päť slovných extrahovaných výrazov sa nejedná o kľúčové výrazy.

V grafe úplnosti vidíme, že zvýšenie počtu extrahovaných výrazov znížením minimálneho počtu výskytov má podľa očakávaní za následok zvýšenie počtu extrahovaných správnych výrazov. Algoritmus dokáže pri najmenej prísnom výbere extrahovať až 50 percent správnych výrazov. Najväčší pokles úplnosti je medzi jedným a dvoma výskytmi, čo môže indikovať, že kľúčové výrazy nemusia mať v texte viacnásobný výskyt. Takisto to môže znamenať, že algoritmus tieto viacnásobné výskyty nedokáže detekovať. Rozdiely v úplnosti medzi maximálnym povoleným počtom slov sú takmer nedetekovateľné, z čoho vyplýva, že štvor a päť členné výrazy sú len málokedy kľúčové. Naopak najväčšie rozdiely v úplnosti sú viditeľné medzi jedno, dvoj a trojslovnou početnosťou výrazu, čo znamená, že výrazy práve týchto dĺžok najčastejšie bývajú kľúčovými.

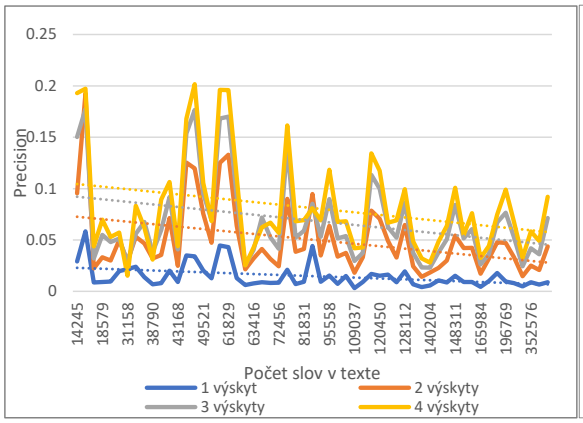
Analýza jednotlivých grafov ukázala, že najvhodnejším parametrom maximálneho počtu slov vo výraze je číslo tri, pretože vyššie čísla negatívne ovplyvňujú presnosť, a na hodnoty úplnosti nemajú žiaden vplyv. Nižšie hodnoty naopak zlepšujú presnosť, majú však väčší dopad na hodnoty úplnosti, ktorý je v tejto fáze programu dôležitejší. Je však o čosi náročnejšie určiť vhodnú minimálnu hodnotu výskytov, pretože rôzne hodnoty môžu fungovať lepšie pre rôzne rozsahy kníh. Navyše grafy ukázali, že hodnoty úplnosti a presnosti sú nepriamo úmerné, je teda na používateľovi, aby sa rozhodol či uprednostňuje vyššiu hodnotu úplnosti (nízky minimálny počet výskytov), presnosti (vysoký minimálny počet výskytov) alebo ich kompromis.

Vzťah minimálneho počtu výskytov výrazu a rozsahu publikácie

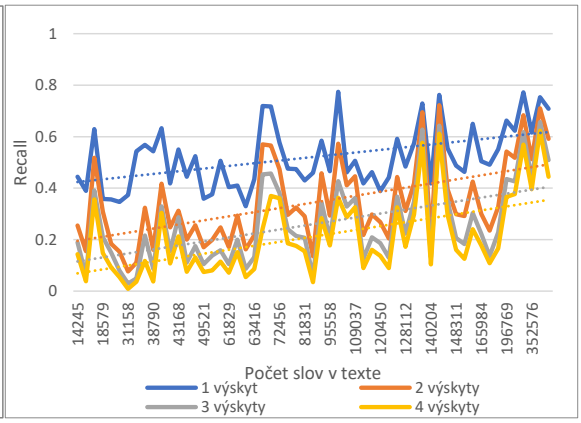
Ako bolo popísané v predchádzajúcej časti, filtrovanie kandidátov na základe minimálneho počtu výskytov v texte môže mať variabilné výsledky vzhľadom na rozsah textu. Hypotéza je, že kratšie texty celkovo obsahujú menší počet slov a teda prísnejšia filtrácia podľa minimálneho počtu výskytov bude mať negatívnejší dopad na nájdené množstvo kandidátov ako u dlhších textov. Cieľom testov je zistiť, či by sa hodnota daného parametru mala dynamicky odvíjať od rozsahu textu. Výsledky analýzy sú popísané v nasledujúcich troch grafoch.



(a) Počet nájdených výrazov



(b) Presnosť



(c) Úplnosť

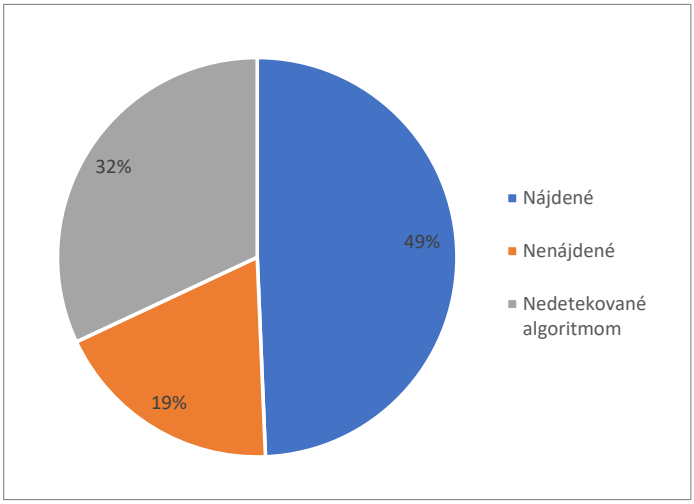
Obr. 9.2: Vplyv dĺžky publikácií a hodnôt parametrov na dosiahnuté výsledky

V prvom grafe je znázornený extrahovaný počet kandidátov pri rôznych minimálnych počtov výskytov jednotlivých publikácií zoradených podľa počtu slov. Pre lepšiu čitateľnosť bola pre 1 výskyt použitá sekundárna vertikálna os. Nie je prekvapujúce, že s väčším počtom slov v knihe stúpa aj počet extrahovaných výrazov, a takisto že tento počet výrazov je nižší pri vyššom počte výskytov. Nás zaujíma, či je rozdiel v počte nájdených výrazov výrazne vyšší pri kratších publikáciách ako pri dlhších. Z grafu vidíme, že sa hypotéza nepotvrdila a pomalšie rastúce trendové čiary u troch a štyroch výskytov indikujú, že filtrácia na základe počtu výskytov má výraznejší efekt v eliminácii výskytov v dlhších textoch.

Zvyšné dva grafy ukazujú hodnoty úplnosti a presnosti v rôznych počtoch výskytov zoradené podľa dĺžky textov. Rovnobežnosť trendových čiar u presnosti aj úplnosti opäť ukazuje, že iné hodnoty parametrov nemajú rozdielne vplyvy na dosiahnuté hodnoty. Mierne klesajúce trendové čiary v grafe presnosti indikujú, že dosiahnutie vyššej hodnoty presnosti je jednoduchšie v kratších textoch a na druhú stranu rastúce trendové čiary v grafe úplnosti indikujú vyššiu priemernú dosiahnutú hodnotu úplnosti v dlhších textoch. Analýza teda ukázala, že hodnoty parametra určujúceho minimálny počet výskytov výrazov nie je nutné meniť v závislosti od rozsahu textu.

Schopnosť algoritmu detekovať kľúčové výrazy

Keďže úlohou algoritmov v prvej fáze programu je detekcia a výber možných kľúčových slov, sú dosiahnuté hodnoty úplnosti dôležitejšie ako hodnoty presnosti, pretože niektoré z vybraných výrazov budú podľa skóre v neskorších fázach vyradzované. Táto časť sa teda pozrie na schopnosti detekcie kľúčových slov. Nasledujúci graf ukazuje úspešnosť algoritmu pri najmenej prísne nastavených parametroch:

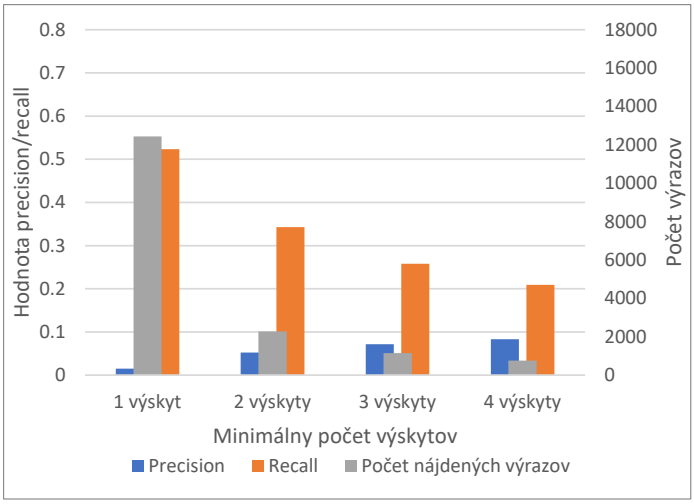


Obr. 9.3: Štatistika skutočných výrazov

Vidíme, že algoritmus dokáže v najlepšom prípade detekovať takmer 50 percent všetkých výrazov. 19 percent všetkých výrazov sa v texte vôbec nenachádza, nie je možné teda očakávať, aby ich algoritmus detekoval. Absencia výskytov môže mať viacero príčin, ako sú chyby detekcie znakov OCR technológiou, chyby pri prevode z formátu pdf do txt, chyby pri vytváraní zoznamov kľúčových výrazov, rozdielné poradie slov vo výrazoch, alebo používanie synoným. Každopádne sú tieto výrazy pre analýzu nepodstatné a dôležitejších je 32 percent výrazov, ktoré sa v texte nachádzajú, ale neboli algoritmom selektované.

Celkové dosiahnuté výsledky

Nasledujúci graf ukazuje celkové dosiahnuté výsledky presnosti, úplnosti a počtu nájdených slov.



Obr. 9.4: Výsledky dosiahnuté algoritmom RAKE

Jednotlivé hodnoty v tomto grafe budú neskôr porovnávané s výsledkami dosiahnutými algoritmom na základe POS tagov za účelom zistenia vhodnejšieho algoritmu získavania kandidátov.

9.2 Výber kandidátov na základe POS tagov

Druhá metóda je založená na tagovaní jednotlivých slov a výbere tých výrazov, ktorých tagy sa zhodujú s jedným s predo predpísaných vzorov. Takisto je možná filtrácia len výrazov, ktoré majú v texte aspoň určité množstvo výskytov. Algoritmus na svoju funkciu potrebuje mať vytvorený zoznam vzorov POS tagov, podľa ktorých bude daných kandidátov vyberať.

9.2.1 Vytvorenie zoznamu vzorov POS tagov

Prvým krokom je získanie POS tagov, pomocou ktorých budú vyberaní kandidáti. Snahou je vybrať také vzory, ktoré sa najčastejšie vyskytujú v skutočných kľúčových výrazoch a teda majú najväčšiu pravdepodobnosť správne identifikovať kľúčové výrazy v texte. Zo zoznamu všetkých skutočných kľúčových výrazov bol vytvorený zoznam POS tagov, ktoré sú v týchto výrazoch najčastejšie. Z tohto zoznamu boli vybrané najvhodnejšie tagy, podľa ktorých budú vyhľadávané výrazy v texte. Jednotlivé tagy sú popísané v nasledujúcej tabuľke.

Tag	Slovný druh	Príklad
NN	podstatné meno (všeobecné)	metafora
AA NN	prídavné meno (všeobecné)-podstatné meno	čierna diera
NN NN	podstatné meno-podstatné meno	hranica súhvezdia
NN AA	podstatné meno-prídavné meno	tiger indický
NN AA NN	podstatné meno-prídavné meno-podstatné meno	fáza životného cyklu
AA AA NN	prídavné meno-prídavné meno-podstatné meno	vegetatívny nervový systém
AA NN NN	prídavné meno-podstatné meno-podstatné meno	rovníkové súradnice hviezd
NN NN NN	podstatné meno-podstatné meno-podstatné meno	cyklus vznikania prvkov
AU NN	prídavné meno (privlastňovacie)-podstatné meno	Pytagorova veta

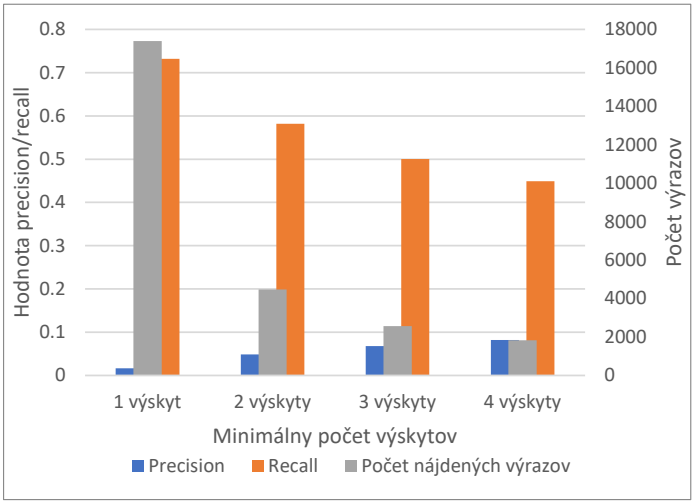
Tabuľka 9.1: Použité vzory POS tagov

Kľúčové výrazy bývajú najčastejšie jedno, dvoj alebo trojslovné, čo je reflektované v dĺžke týchto tagov. Cieľom analýzy algoritmu bude aj zisťovanie vhodnosti jednotlivých tagov.

9.2.2 Analýza výsledkov algoritmu

Výsledky algoritmu pri rozdielnych minimálnych počtoch výskytov

Jediným parametrom tejto metódy je výber len tých kandidátov, ktorí dosiahli minimálny počet výskytov v texte. Táto časť porovná rozdiely v počte extrahovaných kandidátov a hodnotách úplnosti a presnosti pre jednotlivé hodnoty výskytov. Namerané dáta sú reprezentované v nasledujúcom grafe.

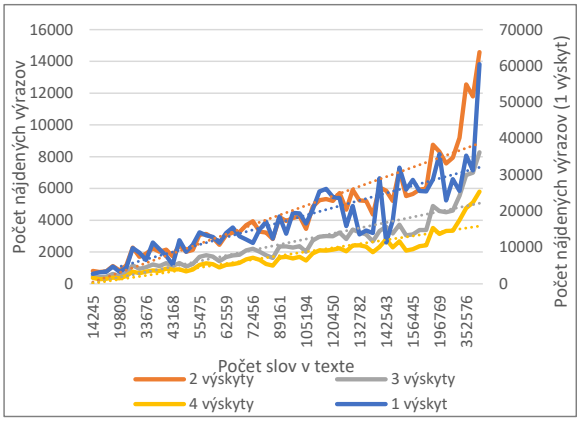


Obr. 9.5: Výsledky dosiahnuté extrakciou na základe POS tagov

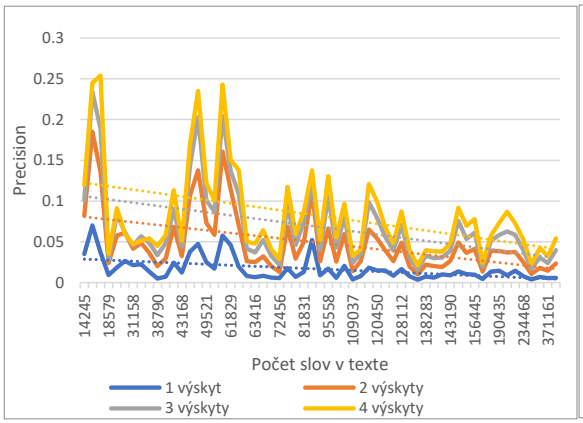
Na sekundárnej vertikálnej osi vidíme počty extrahovaných kľúčových výrazov. Pri jednom výskyte bol vybraný najvyšší počet výrazov, avšak pri podmienke opakovaných výskytov sa tieto čísla výrazne znižujú. Primárna vertikálna os ukazuje jednotlivé hodnoty presnosti a úplnosti, pričom je tu znovu videný jav nepriamej úmernosti týchto metrík. Hodnota úplnosti dosahuje najlepší výsledok až 70 percent a postupne sa znižuje so sprísňujúcou sa podmienkou minimálneho počtu výskytov. Naopak presnosť sa postupne zvyšuje, nepomerne vyššia je však miera znižovania hodnoty úplnosti, čo indikuje, že hoci je vyšší počet výskytov ukázateľom vyššej pravdepodobnosti kľúčovosti výrazu, nejde však o najvhodnejšiu metódu výberu, pretože eliminuje priveľké množstvo správnych výrazov.

Rozdiely vplyvu rôznych počtov výskytov na počty vybraných kandidátov vzhľadom na dĺžku textu

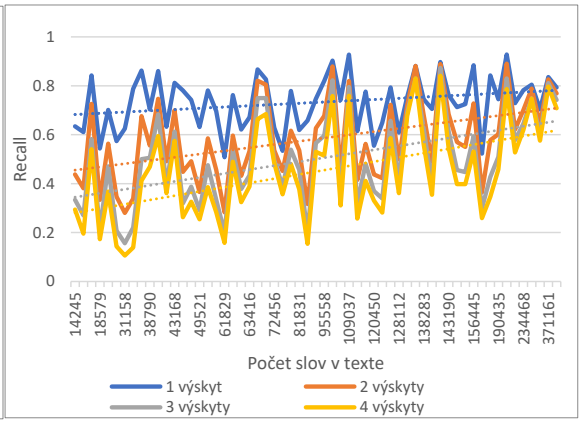
Táto analýza bola robená v predchádzajúcej časti pre výber kandidátov podľa stop slov, avšak tam bolo potrebné počítať so skutočnosťou, že viacnásobné výskyty výrazov neboli detekované. To by sa nemalo stať pri použití metódy POS tagov, preto je potrebné zistiť aký je rozdiel pri extrakcii výrazov v dlhých a krátkych publikáciách v rôznych počtoch výskytov. Namerané dáta sú prezentované v nasledujúcich troch grafoch:



(a) Počet nájdených výrazov



(b) Presnosť



(c) Úplnosť

Obr. 9.6: Vplyv dĺžky publikácií a hodnôt parametrov na dosiahnuté výsledky

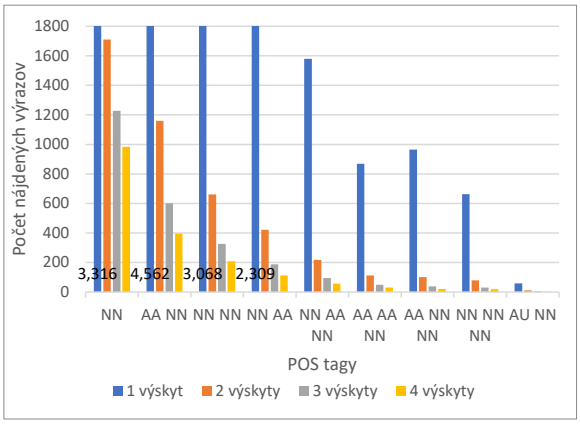
Prvý graf zobrazuje počty extrahovaných kľúčových slov pri rôznych počtoch výskytoch. Pre lepšiu viditeľnosť je počet výrazov pri jednom výskyte mapovaný na sekundárnej vertikálnej osi. Trendové čiary sú rastúce, čo ukazuje, že z dlhších textov je podľa očakávaní extrahované väčšie množstvo výrazov. Miera rastu sa však pri vyšších hodnotách parametru znižuje, čo znamená, že z dlhších textov je vyfiltrovaný väčší počet výrazov.

Grafy presnosti potvrdzujú určitú mieru korelácie medzi počtom výskytov v texte a pravdepodobnosťou, že je výraz kľúčový, keďže pri vyšších hodnotách parametra sú hodnoty presnosti vyššie. Trendové čiary sú klesajúce, pričom miera klesania sa zvyšuje pri zväčšovaní hodnoty parametra, čo na rozdiel od predošlého algoritmu potvrdzuje hypotézu, že rovnaký parameter počtu výskytov má výraznejší efekt na texty kratších dĺžok. Nemožnosť túto hypotézu dokázať v predošlom algoritme je možné pripísať neschopnosti daného algoritmu detekovať všetky výskyty daných výrazov.

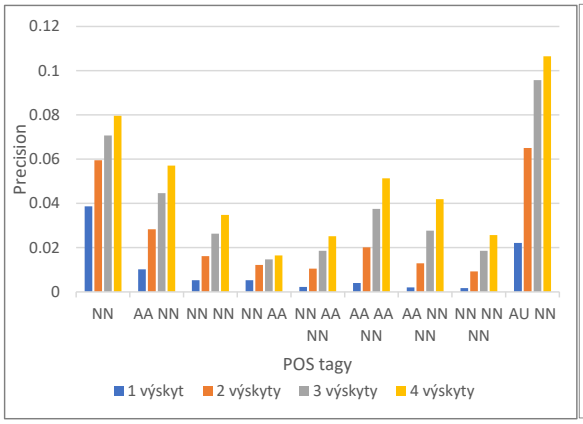
Toto potvrdzuje aj graf úplnosti, v ktorom je viditeľná vyššia miera rastu trendových čiar pri zvyšujúcich sa počtoch výskytov výrazov. Znamená to, že pre dosiahnutie rovnakých výsledkov vo všetkých textoch bude potrebné prispôsobiť hodnotu parametru k rozsahu textu.

Úspešnosť jednotlivých tagov pri detekcii výrazov

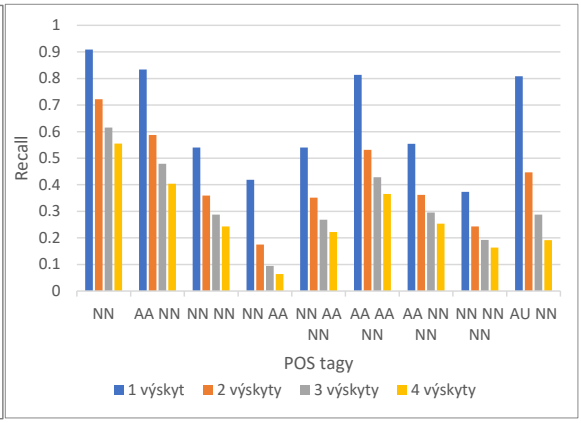
Táto časť sa zaoberá analýzou úspešnosti vyhľadávania výrazov jednotlivých tagov využitých v algoritme. Nasledujúce grafy znázorňujú dosiahnuté výsledky daných tagov pri rôznych hodnotách parametra počtu výskytov.



(a) Počet nájdených výrazov



(b) Presnosť



(c) Úplnosť

Obr. 9.7: Výsledky dosiahnuté jednotlivými vzormi tagov

Tagy v jednotlivých grafoch sú zoradené od najčastejšie sa vyskytujúcich, po najmenej často sa vyskytujúce v zozname skutočných kľúčových výrazoch. Prvý graf znázorňuje priemerný počet extrahovaných výrazov pre jednotlivé tagy, pričom výsledky pri hodnote parametru 1 sú reprezentované na sekundárnej vertikálnej osi. Môžeme vidieť, že počty vybraných výrazov pre jednotlivé tagy postupne klesajú, čo znamená, že pomery frekvencií tagov v texte a v zozname kľúčových výrazov sú podobné.

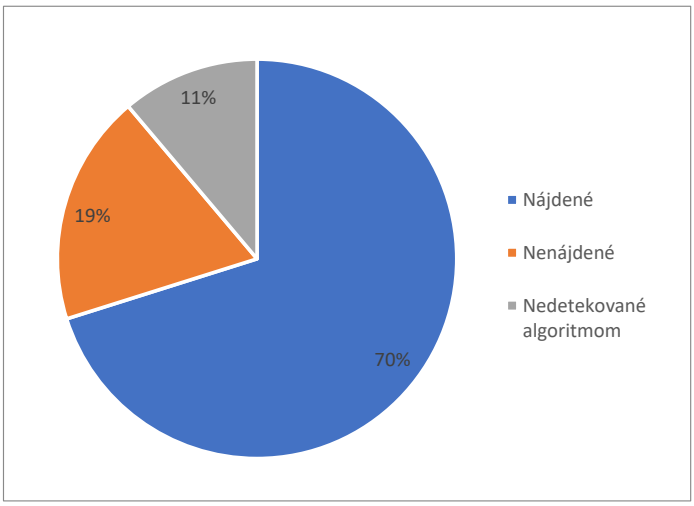
Druhý graf ukazuje hodnoty úplnosti pre jednotlivé tagy a môžeme v ňom vidieť určitý nepomer týchto hodnôt. Hodnota úplnosti pri jednom výskyte výrazu by mala teoreticky dosahovať hodnotu 100 percent, avšak ako ukázal graf v 9.3, až 20 percent kľúčových výrazov sa v texte vôbec nenachádza. Toto číslo nie je medzi jednotlivými tagmi rovnomerne distribuované, a hodnota úplnosti pri jednom výskyte skôr znamená ako veľmi bol daný tag touto nedokonalosťou postihnutý a nemusí nutne vyjadrovať jeho vhodnosť. Vhodnosť tagov je zisťovaná mierou znižovania hodnôt úplnosti pri vyšších počtoch výskytov, čo indikuje ťažšiu detekovateľnosť výrazov daných tagov. Výrazne najväčší prepad je zaznamenaný

v tagu NN AA, ktorý pri jeho relatívne vysokej početnosti môže znamenať nepomerne vyššie množstvo falošne pozitívnych výsledkov. Druhý vysoký prepad nastal v tagu AU NN, čo však nemusí byť problémom kvôli malému počtu výskytov výrazov s týmto tagom.

Potvrdzuje sa to aj v grafe presnosti, kde tag s najnižšou mierou rastu je opäť NN AA. Tieto poznatky znamenajú, že správne výrazy s týmto tagom sa nemusia nutne viac krát vyskytovať v texte. Tento tag bude teda z finálneho zoznamu tagov odstránený. Graf tiež ukazuje, že hodnota presnosti je nepomerne nižšia pri viacslovných výrazoch, a jej zlepšenie bude predmetom nasledujúcej kapitoly.

Schopnosť algoritmu detekovať kľúčové výrazy

Podobne, ako aj v algoritme fungujúcom na základe stop slov je aj tu potrebné zistiť celkovú schopnosť detekcie kľúčových výrazov. Výsledky dosiahnuté týmto algoritmom sú prezentované v nasledujúcom grafe:

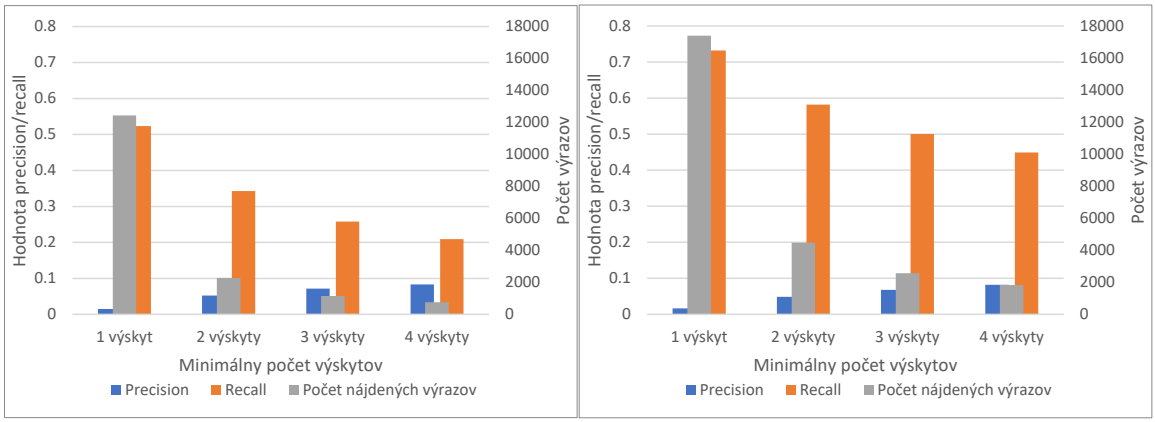


Obr. 9.8: Výsledky dosiahnuté extrakciou na základe POS tagov

Algoritmus má úspešnosť detekcie až 70 percent všetkých výrazov. 20 percent výrazov sa však v texte nenachádza vôbec, čo znamená, že algoritmus nezachytil len 10 percent všetkých výrazov. Toto číslo je potenciálne možné znížiť pridaním ďalších, menej konvenčných tagov, avšak týchto tagov bude musieť pre výrazné zvýšenie hodnoty úplnosti byť pridané relatívne veľké množstvo z dôvodu malého počtu existujúcich kľúčových výrazov pre jednotlivé tagy. Navyše napriek malým návratovým hodnotám môžu tieto tagy detekovať nepomerne veľké množstvo falošne pozitívnych výrazov. Z toho dôvodu sú dosiahnuté výsledky s existujúcimi tagmi považované za dostačujúce.

9.3 Porovnanie výsledkov

Poslednou úlohou analýzy je porovnanie jednotlivých algoritmov a výber toho, ktorý dosahuje pre potreby programu lepšie výsledky. Prvým bodom bude porovnávanie algoritmov na základe počtu nájdených slov a dosiahnutých hodnôt presnosti a úplnosti pri rôznych hodnotách parametru počtu výskytov. Algoritmus výberu na základe stop slov má tiež parameter o maximálnom počte slov vo výraze, experimenty však ukázali, že za každých podmienok je najvhodnejšia hodnota tohto parametru tri.



(a) RAKE

(b) POS tagy

Obr. 9.9: Porovnanie výsledkov jednotlivých algoritmov

Vidíme, že druhý algoritmus dosiahol výrazne vyššie hodnoty úplnosti pre všetky parametre, pričom miera klesania úplnosti je u oboch algoritmov relatívne rovnaká. Obe algoritmy dosahujú pri všetkých hodnotách parametrov veľmi podobné hodnoty, čo dokazuje, že vyššia hodnota úplnosti druhého algoritmu nie je zapríčinená extrakciou vyššieho počtu slov (čo je v grafe tiež naznačené), ale je dôsledkom lepšej schopnosti druhého algoritmu identifikovať kľúčové výrazy.

Algoritmus výberu kandidátov na základe POS tagov však má aj ďalšie výhody, ako sú lepšia modifikovateľnosť a flexibilita. Umožňuje jednoducho pridávať a odoberať požadované POS tagy pre iné druhy textov a takisto je pre jednotlivé tagy možné nastaviť rôzne hodnoty minimálneho počtu výskytov. Z týchto dôvodov bude v algoritme využívaný spôsob výberu kandidátov na základe POS tagov.

Kapitola 10

Analýza jednotlivých prístupov skórovania kandidátov

V predošlej kapitole boli analyzované a porovnávané jednotlivé prístupy výberu kandidátov, pričom bolo dokázané, že lepšie výsledky dosahuje prístup výberu kandidátov na základe POS tagov. Takisto bol ukázaný vplyv parametrov na namerané hodnoty presnosti a úplnosti, a nepriamu úmernosť medzi týmito dvoma metrikami. Táto kapitola sa bude zaoberať rôznymi skórovacími algoritmami výrazov získaných na základ POS tagov. Cieľom kapitoly bude porovnanie výsledkov jednotlivých metód a určenie, ako budú tieto výsledky ovplyvnené zmenou parametra minimálneho počtu výskytov výrazu.

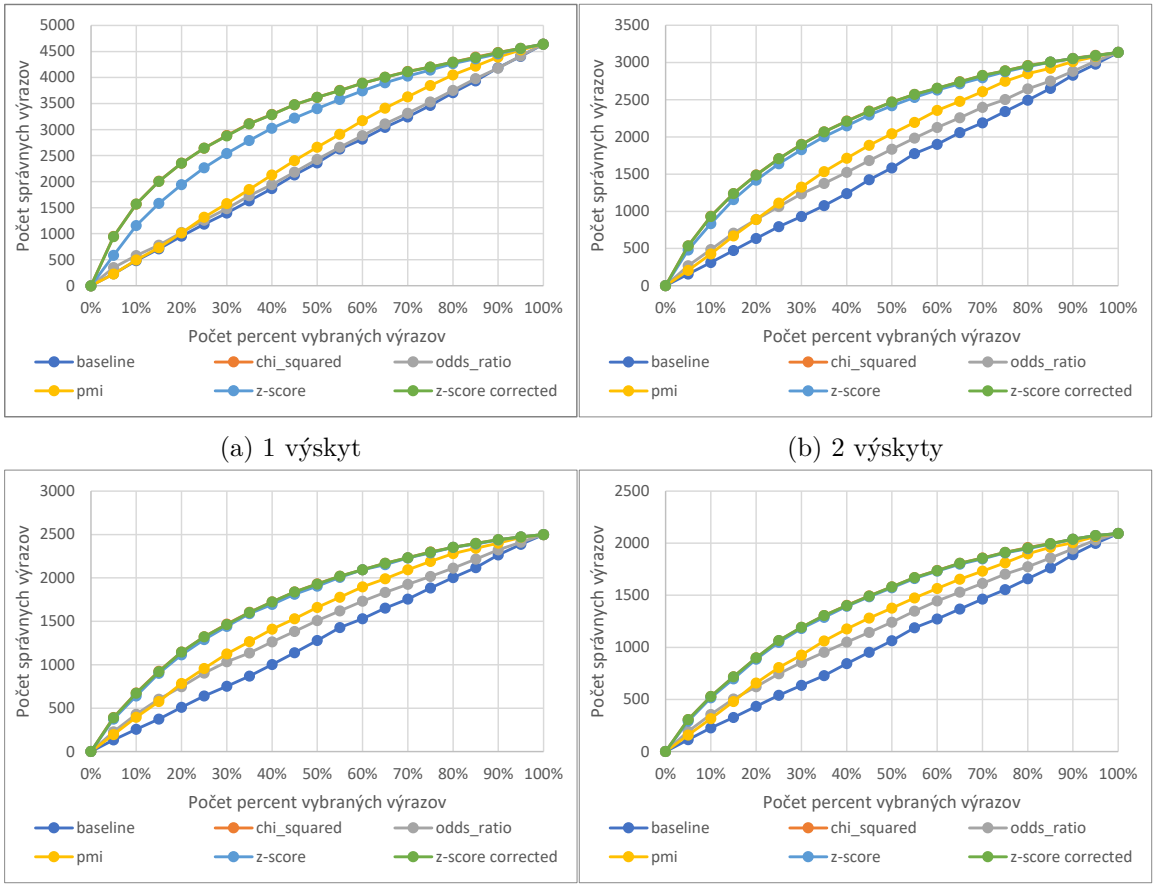
10.1 Spôsob porovnávania jednotlivých algoritmov

Spôsoby výpočtu jednotlivých algoritmov sú popísané v kapitolách 3 a 4, pričom výsledkom týchto algoritmov je skóre týchto výrazov. Vyššie skóre indikuje väčšiu pravdepodobnosť, že sa jedná o kľúčový výraz, teda ak by boli jednotlivé výrazy zoradené podľa skóre, mala by byť pri vysokých skóre detekovateľná vyššia hustota správnych výrazov. Na zistenie efektívnosti takýchto prediktívnych modelov sa využívajú Lift grafy [18]. Zoznam výrazov zoradených podľa skóre bude rozdelený na 10 častí, z ktorých každá reprezentuje určitú percentuálnu časť daného zoznamu, teda prvá časť zahŕňa najlepších 10 percent výrazov, druhá najlepších 20 percent a tak ďalej. V každej z týchto častí bude zistený počet nájdených správnych výrazov, pričom bez použitia akéhokoľvek algoritmu očakávame lineárny rast, teda prvá časť obsahuje 10 percent správnych nájdených výrazov, druhá 20 percent atď. Po použití algoritmov by však koncentrácia správnych výrazov mala byť na začiatku vyššia, čo sa v grafe prejaví strmším úvodným rastom hodnôt nájdených správnych výrazov. Miera tohto rastu je reprezentovaná lift krivkou, pričom väčšia výška tejto krivky znamená lepšie dosiahnuté výsledky algoritmu.

10.2 Štatistické metódy skórovania dvojslovných výrazov

10.2.1 Lift grafy

V sekcii 3.1 bolo predstavených päť skórovacích algoritmov, ktorých úspešnosť bude v tejto časti porovnávaná. Nasledujúce grafy zaznamenávajú lift krivky týchto metód pri rôznych hodnotách parametra minimálneho počtu kľúčových výrazov:



(a) 1 výskyt

(b) 2 výskyty

(c) 3 výskyty

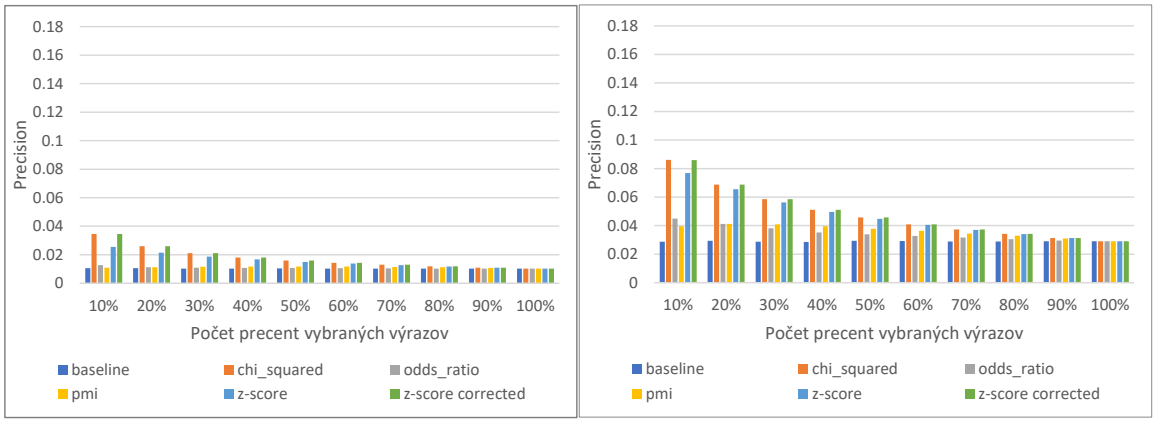
(d) 4 výskyty

Obr. 10.1: Lift grafy štatistických metód dvojslovných výrazov

Úspešnosti jednotlivých metód budú porovnávané voči hodnote baseline, ktorá má podľa očakávaní lineárny rast. Vidíme, že všetky algoritmy sú aspoň čiastočne efektívne, pretože vo všetkých prípadoch je krivka lift vyššia ako hodnota baseline. Najlepšie si počínajú algoritmy Chi squared a z-score corrected, pričom ich dosiahnuté výsledky sú vo všetkých prípadoch takmer identické. Ich úspešnosť je najvyššia pri hodnote parametra 1, avšak prvenstvo si zachovávajú vo všetkých prípadoch. Vidíme, že efektivita algoritmov spravidla klesá so zvyšujúcim sa minimálnym počtom výskytov, čo však nemusí znamenať, že sú preferované nižšie hodnoty parametra, pretože stále nepoznáme, ako budú pri jednotlivých algoritmoch ovplyvnené hodnoty úplnosti a presnosti.

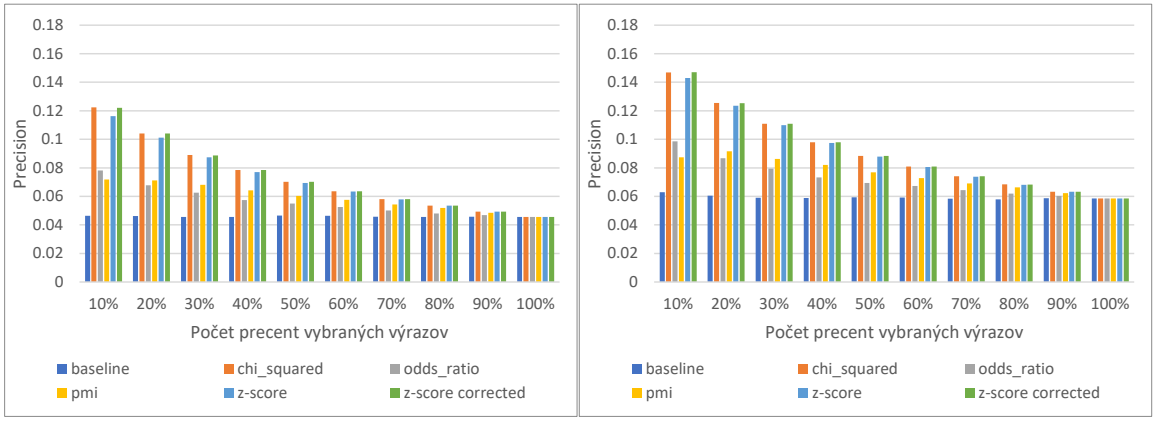
10.2.2 Grafy presnosti a úplnosti

Zistili sme, že jednotlivé algoritmy dosahujú zlepšenie výsledkov, predošlé grafy však neukazujú mieru tohto zlepšenia vzhľadom na hodnotu parametra. Budú teda vyhotovené grafy dosiahnutých hodnôt úplnosti a presnosti týchto algoritmov pre jednotlivé hodnoty parametrov. Hodnoty úplnosti a presnosti budú zisťované pre zvyšujúce sa hodnoty percent najvyššie hodnotených výrazov jednotlivých algoritmov. Ako prvé sú zobrazené grafy hodnôt presnosti:



(a) 1 výskyt

(b) 2 výskyty



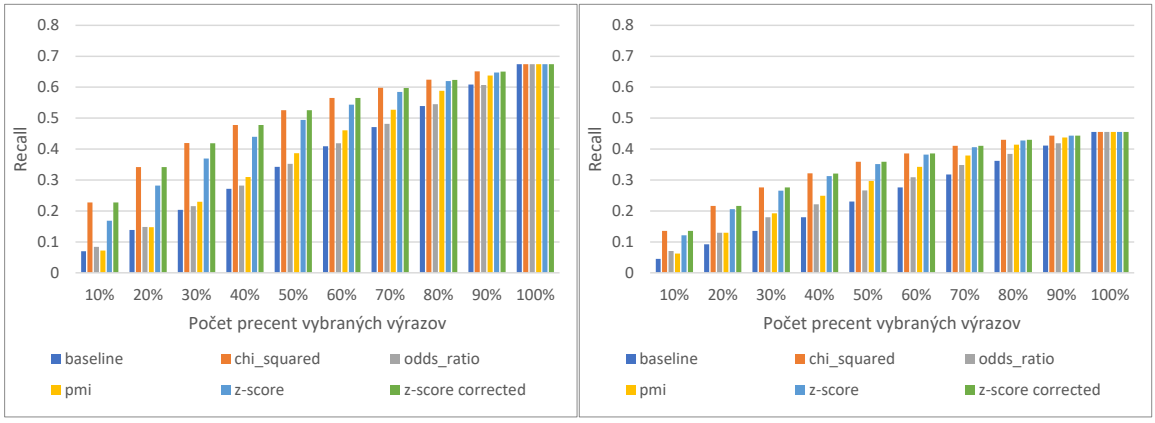
(c) 3 výskyty

(d) 4 výskyty

Obr. 10.2: Hodnoty presnosti štatistických metód dvojslovných výrazov

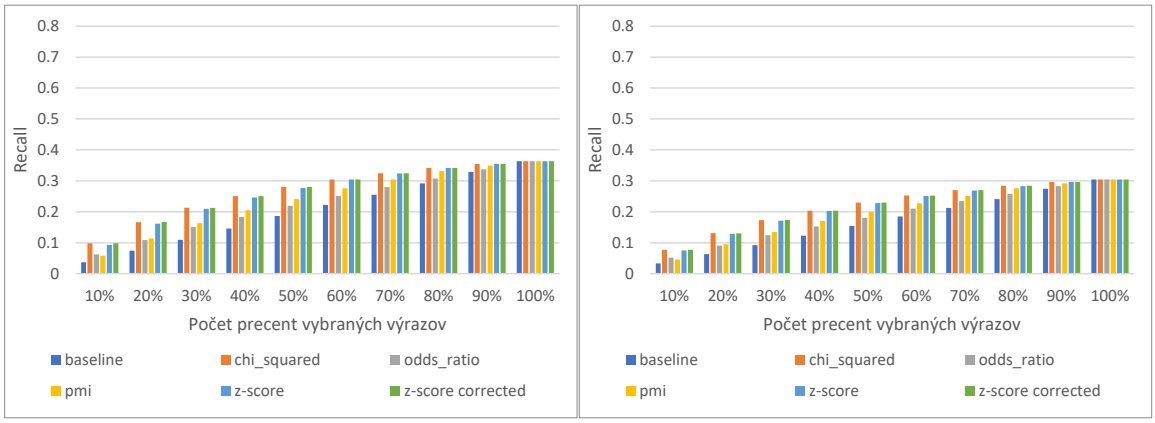
V grafoch vidíme, že hodnoty presnosti sú najvyššie pri výberoch malého počtu najlepšie hodnotených kľúčových výrazov, čo potvrdzuje, že jednotlivé algoritmy sú efektívne v identifikácii kľúčových výrazov. Vidíme tiež, že hodnoty minimálneho počtu výskytov majú stále veľký vplyv na dosiahnuté výsledky, čo sa ukazuje výrazným zvyšovaním hodnôt presnosti pri zvyšujúcej sa hodnote parametra.

Nasledujú grafy úplnosti, ktoré ukazujú ako výrazný je vplyv zmenšovania počtu vybraných výrazov na dosiahnuté hodnoty úplnosti. Tieto hodnoty pre jednotlivé algoritmy pri rôznych hodnotách parametru sú ukázané v nasledujúcich grafoch:



(a) 1 výskyt

(b) 2 výskyty



(c) 3 výskyty

(d) 4 výskyty

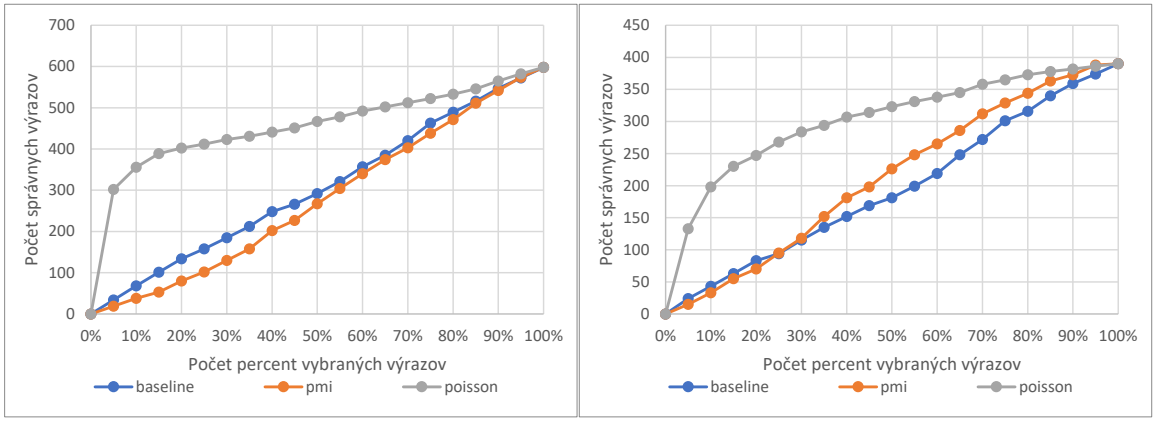
Obr. 10.3: Hodnoty úplnosti štatistických metód dvojslovných výrazov

Vidíme, že jednotlivé algoritmy zlepšujú hodnoty úplnosti oproti hodnotám dosiahnutých pri baseline, pričom tento efekt sa stráca so zvyšujúcim sa počtom vybraných výrazov. Napriek zlepšeniu výsledkov pri použití algoritmu sú však dosiahnuté hodnoty príliš malé na to, aby mohli byť súčasne uspokojené metriky presnosti aj úplnosti. Vyplýva z toho, že užívateľ sa stále musí rozhodnúť, ktorá z týchto metrík je preňho dôležitejšia. Napriek tejto skutočnosti je v grafe viditeľné, že pri odstránení 30-40 percent výrazov s najmenším skóre by straty hodnoty úplnosti boli relatívne malé.

10.3 Štatistické metódy skórovania trojslovných výrazov

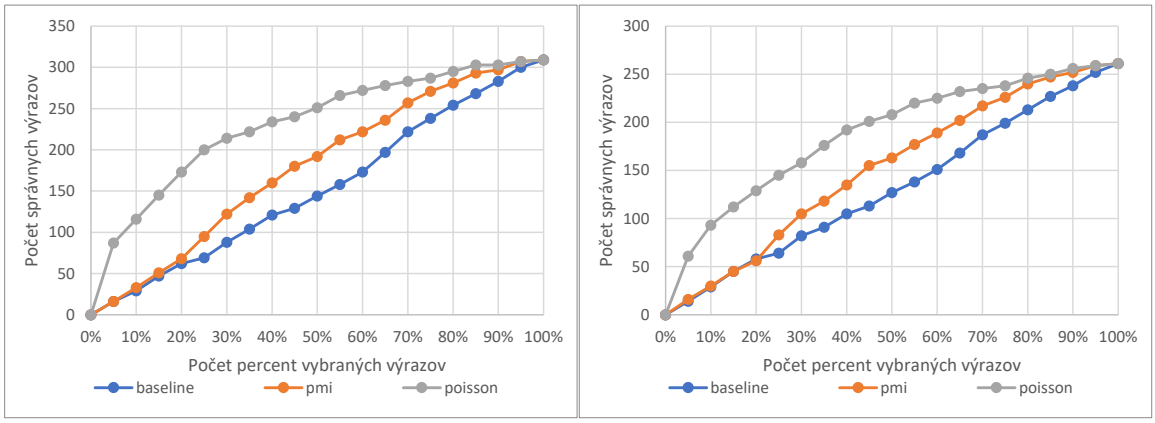
10.3.1 Lift grafy

V sekcii 3.2 boli predstavené dve štatistické metódy pre trojslovné výrazy, ktoré budú v tejto časti analyzované. Podobne ako v predchádzajúcej časti popisujú nasledujúce lift grafy úspešnosť jednotlivých algoritmov:



(a) 1 výskyt

(b) 2 výskyty



(c) 3 výskyty

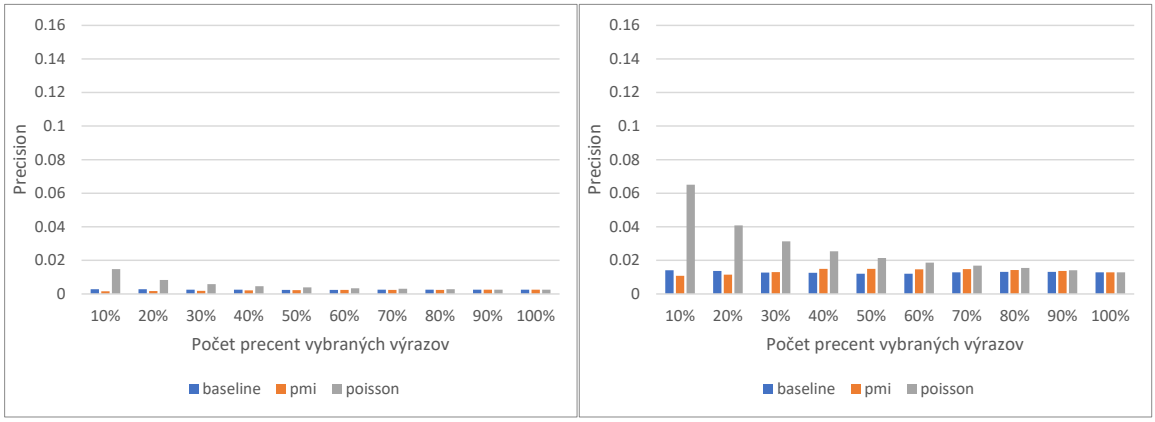
(d) 4 výskyty

Obr. 10.4: Lift grafy štatistických metód trojslovných výrazov

V grafe vidíme, že najvyššia úspešnosť bola dosiahnutá algoritmom Poisson-Stirling. Táto úspešnosť klesá so zvyšujúcom sa hodnotou parametru, čo naznačuje, že týmto parametrom boli odstránené niektoré správne výrazy. Vyššie hodnoty parametra však tiež odstraňujú aj mnoho nesprávnych výrazov, teda vhodnosť hľadania výrazov s nižšou hodnotou parametru sa ukáže až v grafoch presnosti.

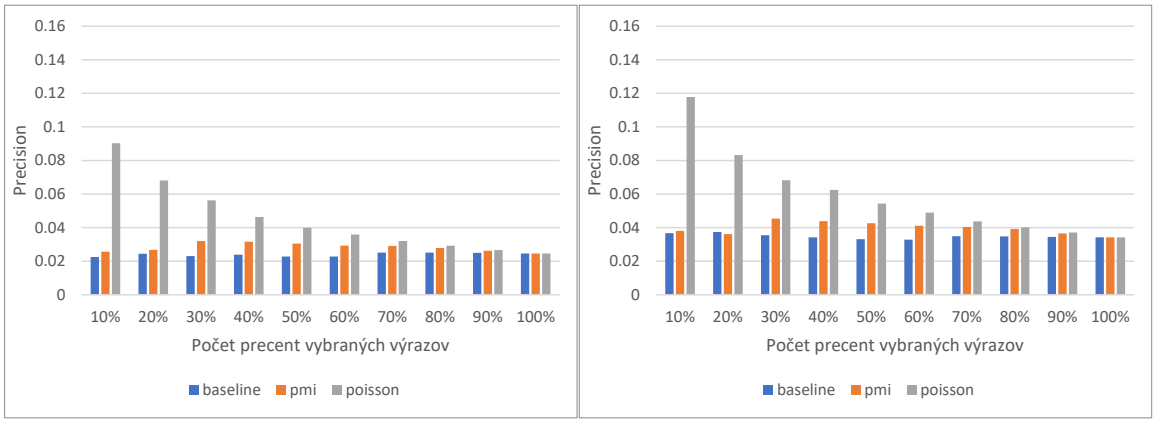
10.3.2 Grafy presnosti a úplnosti

V tejto časti sa nachádzajú grafy dosiahnutých hodnôt presnosti a úplnosti testovaných algoritmov s rôznymi hodnotami parametru. Ako prvé sú grafy presnosti:



(a) 1 výskyt

(b) 2 výskyty



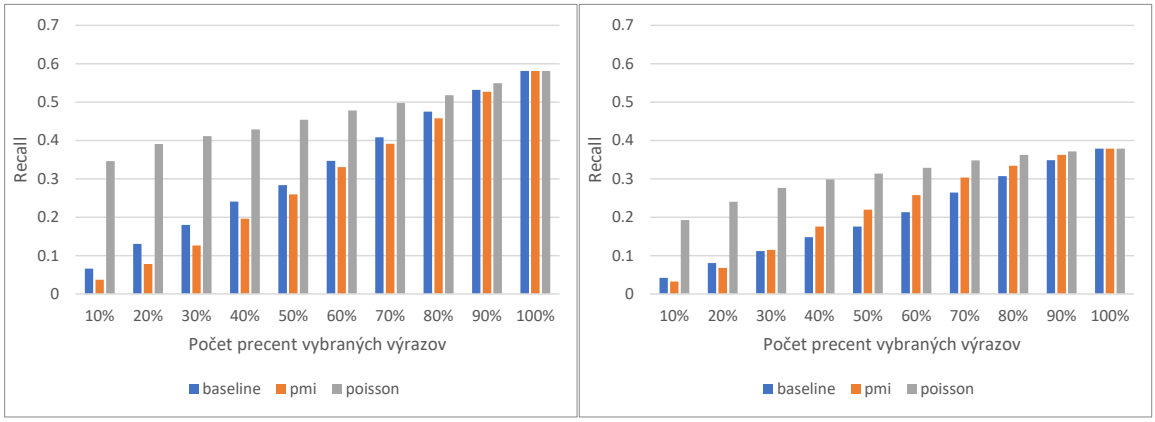
(c) 3 výskyty

(d) 4 výskyty

Obr. 10.5: Hodnoty presnosti štatistických metód trojslovných výrazov

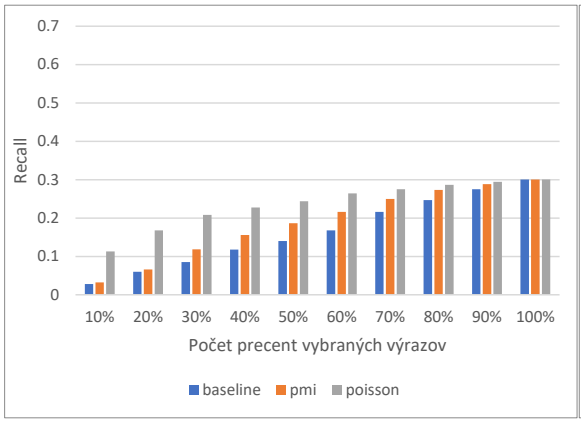
Hodnoty sú opäť najvyššie pri výbere malého množstva výrazov, pričom táto hodnota sa ďalej zvyšuje so zvýšením minimálneho počtu výskytov výrazu, čo znamená, že algoritmus nie je dostatočne presný na to, aby odvrátil nízke hodnoty presnosti zapríčinené nízkou hodnotou parametru.

Druhá séria grafov znázorňuje hodnoty úplnosti:

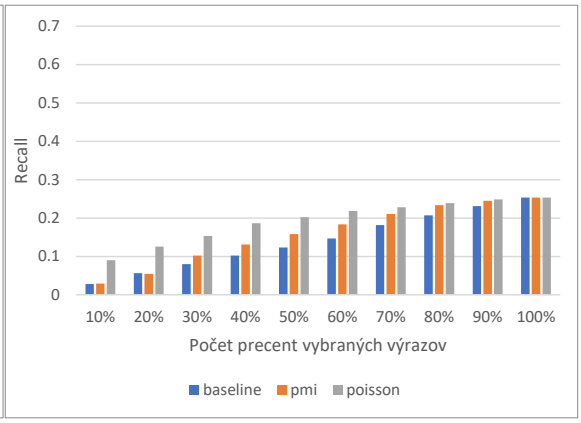


(a) 1 výskyt

(b) 2 výskyty



(c) 3 výskyty



(d) 4 výskyty

Obr. 10.6: Hodnoty úplnosti štatistických metód trojslovných výrazov

Vysoká úspešnosť vyhľadávania správnych výrazov algoritmu Poisson-Stirling má za následok relatívne vysoké hodnoty úplnosti aj pri menších počtoch vybraných výrazoch. Tento efekt je menej intenzívny pri vyšších hodnotách parametru, napriek tomu by však bolo možné dosiahnuť relatívne vysoké hodnoty úplnosti pri odstránení vysokého množstva najnižšie skórovaných výrazov. Pri veľmi nízkych množstvách vybraných výrazov sú však hodnoty úplnosti stále príliš malé, teda užívateľ sa musí rozhodnúť, či je preňho dôležitejšia hodnota presnosti, alebo úplnosti.

10.4 Zhodnotenie výsledkov štatistických metód

Z nameraných výsledkov pri dvoj aj trojslovných výrazoch vidíme, že štatistickémetódy použité na predikciu kľúčových výrazov sú efektívne, pričom pri bigramoch si konzistentne udržiaval najlepšie výsledky algoritmus Chi-squared a pri trigramoch to bol algoritmus Poisson-Stirling. Hodnoty presnosti a úplnosti sú stále výrazne ovplyvňované parametrom minimálneho počtu výskytov, teda výber hodnoty tohto parametra naďalej odzrkadľuje cieľ použitia tejto aplikácie. Napriek efektívnosti testovaných algoritmov nebol plne potlačený vzťah nepriamej úmernosti medzi úplnosťou a presnosťou, z čoho vyplýva, že neexistuje objektívne najlepšia percentuálna hodnota výberu najlepšie skórovaných kandidátov, táto voľba musí byť robená na základe preferovanej metriky pri výbere kľúčových výrazov.

10.5 Analýza lingvistických prístupov skórovania kandidátov

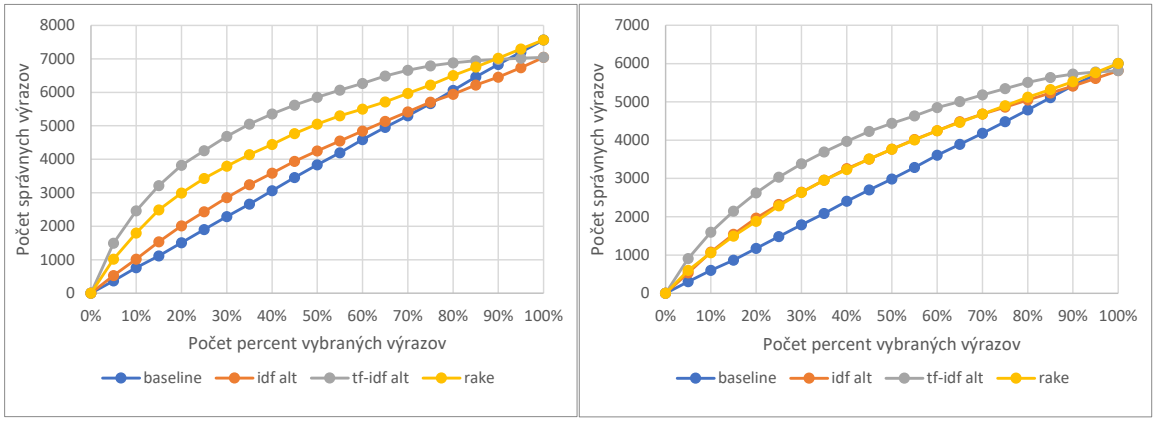
Táto kapitola sa zaoberá analýzou a porovnávaním lingvistických metód skórovania kandidátov popísaných v kapitole 4. Testované a porovnávané budú algoritmy TF-IDF a RAKE. V predošlých kapitolách bolo dokázané, že frekvencia výrazu koreluje s vyššou pravdepodobnosťou jeho kľúčovosti, avšak je potrebné tiež ukázať, že nižšia miera výskytov v iných dokumentoch je tiež prediktorom kľúčových výrazov. Z toho dôvodu budú testované okrem hodnôt tf-idf aj hodnoty idf. Ako bolo v predošlej časti spomenuté, algoritmus TF-IDF potrebuje na svoje fungovanie korpus dokumentov, z ktorého je získaná početnosť jednotlivých slov. Na túto funkciu bude použitá štatistika početosti výrazov spoločnosti Google, ktorá okrem samotných slov ponúka početnosti dvoj a trojslovných výrazov. Korelácia kľúčovosti a nízkej výskytovosti samostatných slov je v iných vedeckých publikáciách podrobne dokázaná, existujú však pochybnosti, či to platí aj pre viacslovné výrazy. Z toho dôvodu budú jednotlivé algoritmy testované samostatne pre rôzne počty slov vo výrazoch. Viacslovné výrazy budú tiež testované s alternatívnym spôsobom výpočtu IDF, ktoré sa bude rovnať priemeru početností jednotlivých slov výrazu. Taktiež bude sledovaný vplyv parametra počtu výskytov na dosiahnuté výsledky jednotlivých algoritmov. Zdrojový kód algoritmu RAKE bol prevzatý z [1] a upravený pre vlastné potreby.

10.6 Lingvistické metódy skórovania jednoslovných výrazov

V prvej časti sa budeme zaoberať dosiahnutými výsledkami pri jednoslovných výrazov. Keďže štatistické metódy nie sú na tieto výrazy aplikovateľné, lingvistické metódy sú jediným spôsobom ohodnocovania týchto výrazov. Navyše sú jednoslovné výrazy najpočetnejším typov kľúčových výrazov, preto je dôležité nájsť vhodný spôsob získavania týchto výrazov. Na rozdiel od nasledujúcich častí nebudú u jednoslovných výrazoch vypočítané alternatívne hodnoty IDF a TF-IDF, ktoré boli popísané vyššie, keďže by dosiahli rovnaké výsledky.

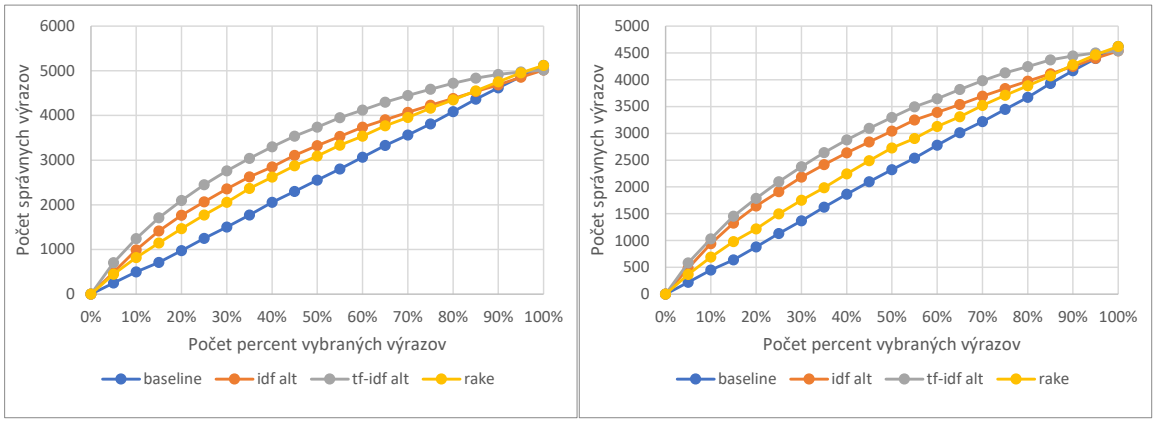
10.6.1 Lift grafy

Podobne ako v predošlých kapitolách bude na porovnávanie jednotlivých algoritmov použitý lift graf, ktorého výsledky pre jednotlivé hodnoty parametra sú ukázané nižšie:



(a) 1 výskyt

(b) 2 výskyty



(c) 3 výskyty

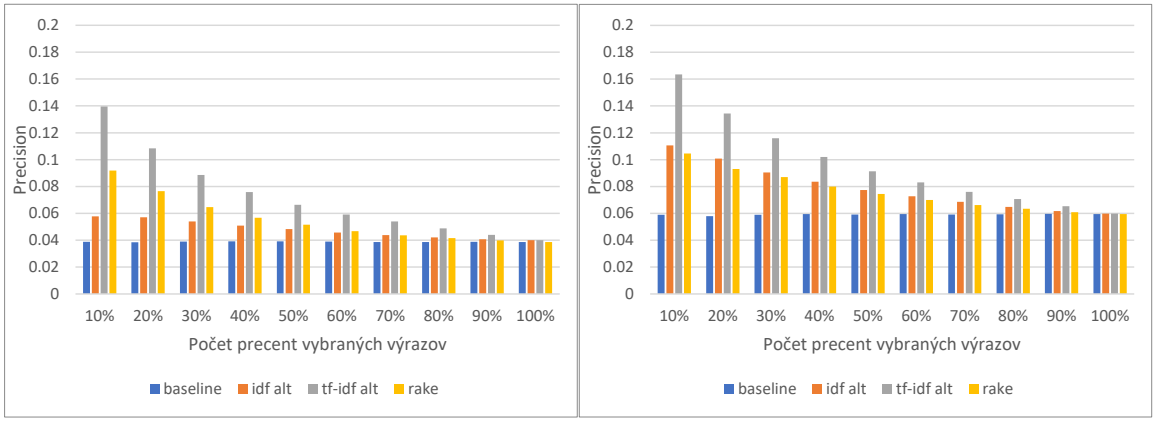
(d) 4 výskyty

Obr. 10.7: Lift grafy lingvistických metód jednoslovných výrazov

Vidíme, že najlepšie výsledky vo všetkých grafoch dosahuje algoritmus TF-IDF, napriek celkovej klesajúcej účinnosti algoritmov pri vyšších hodnotách parametra. Algoritmus RAKE dosahuje tiež detekovateľný lift, avšak jeho účinnosť výrazne klesá so zvyšujúcou sa hodnotou parametru. Výsledky IDF sú nižšie ako TF-IDF, avšak elevácia lift grafu dokazuje, že hodnota IDF koreluje s kľúčovosťou výrazu. Lift grafy sú unikátne v tom, že jednotlivé algoritmy na konci nedosahujú rovnaké výsledky, ale hodnoty baseline a RAKE majú na konci nájdený vyšší počet výrazov ako IDF a TF-IDF. To je spôsobené nutnosťou týchto algoritmov hľadať jednotlivé slová v zozname, pričom sa môže stať, že niektoré z týchto slov nebudú nájdené, čo znemožňuje výpočet skóre týchto algoritmov. To však nemusí byť nutne negatívom, pretože sa takisto potenciálne zníži počet falošne pozitívnych výrazov, čo prispeje k vyšším hodnotám presnosti.

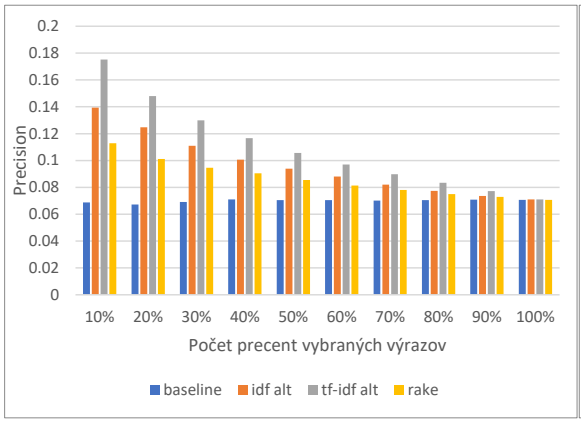
10.6.2 Grafy presnosti a úplnosti

Nasledujúce grafy ukazujú hodnoty presnosti dosiahnuté jednotlivými algoritmami:

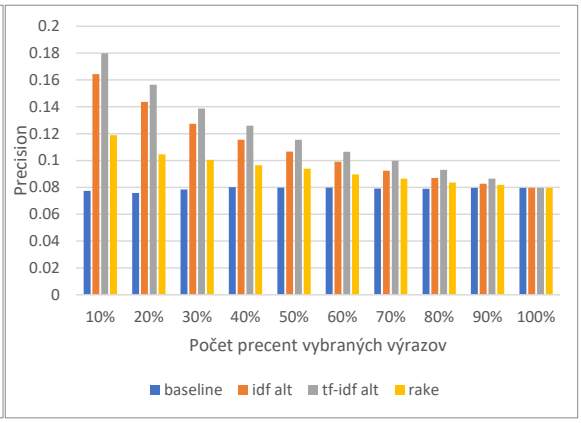


(a) 1 výskyt

(b) 2 výskyty



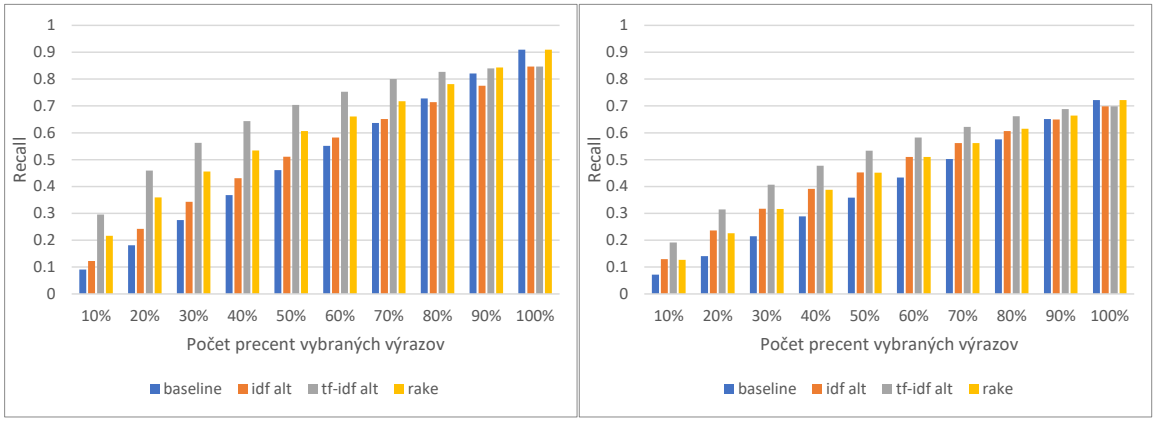
(c) 3 výskyty



(d) 4 výskyty

Obr. 10.8: Hodnoty presnosti lingvistických metód jednoslovných výrazov

Vidíme, že v najlepšom prípade dosahujú hodnoty presnosti až 20 percent, pričom podobne ako v predošlých častiach sú tieto hodnoty vyššie pri väčších hodnotách minimálneho počtu výskytov. Najlepšie si počína algoritmus TF-IDF, k čomu mohla prispieť aj eliminácia niektorých nesprávnych výrazov. Táto vlastnosť však mohla dopomôcť k nižším hodnotám úplnosti, ktoré sú popísané nižšie:



(a) 1 výskyt

(b) 2 výskyty

(c) 3 výskyty

(d) 4 výskyty

Obr. 10.9: Hodnoty úplnosti lingvistických metód jednoslovných výrazov

V grafoch vidíme, že vo väčšine prípadov dosahuje najlepšie výsledky metóda TF-IDF a jej dôsledok spôsobu skórovania sa negatívne odzrkadľuje až pri najvyšších percentách vybraných výrazov. Podobne ako v predošlých algoritmoch, aj tu vidíme určitý prepad hodnôt úplnosti pri vyšších hodnotách minimálneho počtu výskytov. Opäť je možné z jednotlivých grafov usúdiť, že nie je možné súčasne dosiahnuť vysoké hodnoty úplnosti a presnosti a užívateľ si musí vybrať, ktorú z nich preferuje.

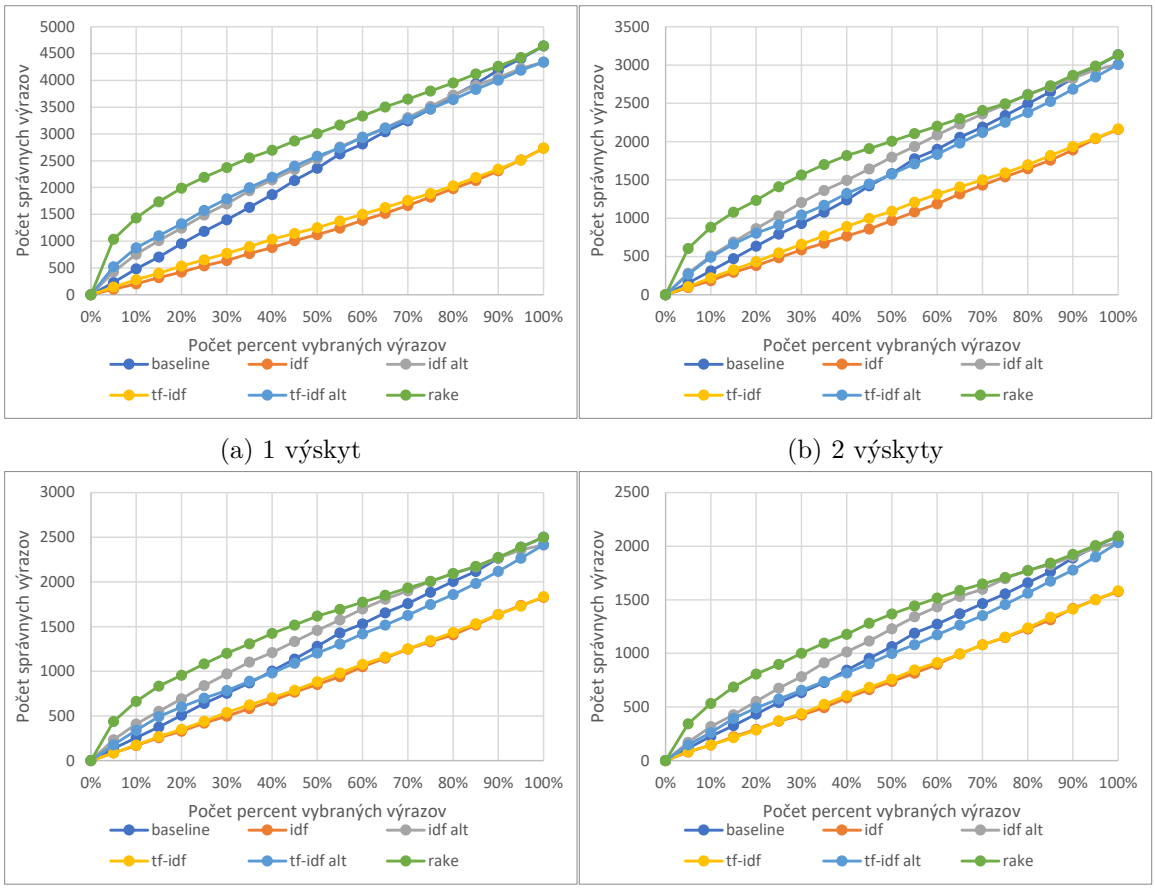
Z analýzy môžeme vyhodnotiť, že výrazne najlepšie výsledky vo väčšine prípadov dosahuje algoritmus TF-IDF. Jedinou výnimkou môže byť, ak užívateľ požaduje dosiahnuť čo najvyššiu hodnotu úplnosti, pretože tento algoritmus eliminuje niektoré správne výrazy. Keďže neexistujú štatistické metódy výpočtu skóre jednoslovných výrazov, bude tento algoritmus samostatne použitý pri zisťovaní vhodnosti jednoslovných kandidátov.

10.7 Lingvistické metódy skórovania dvojslovných výrazov

Pri dvoj a trojslovných výrazoch bude porovnávané väčšie množstvo metód, pretože budú pridané alternatívne spôsoby výpočtu hodnôt IDF na základe početností jednotlivých slov výrazu. V predošlej kapitole boli analyzované štatistické metódy výpočtu skóre, čo znamená, že v prípade dosiahnutia dobrých výsledkov lingvistických metód bude skúmaná kompatibilita sekvenčného použitia štatistických a lingvistických metód.

10.7.1 Lift grafy

Ako obvykle, metódy sú porovnávané v lift grafoch, ktoré sú znázornené nižšie:



(a) 1 výskyt

(b) 2 výskyty

(c) 3 výskyty

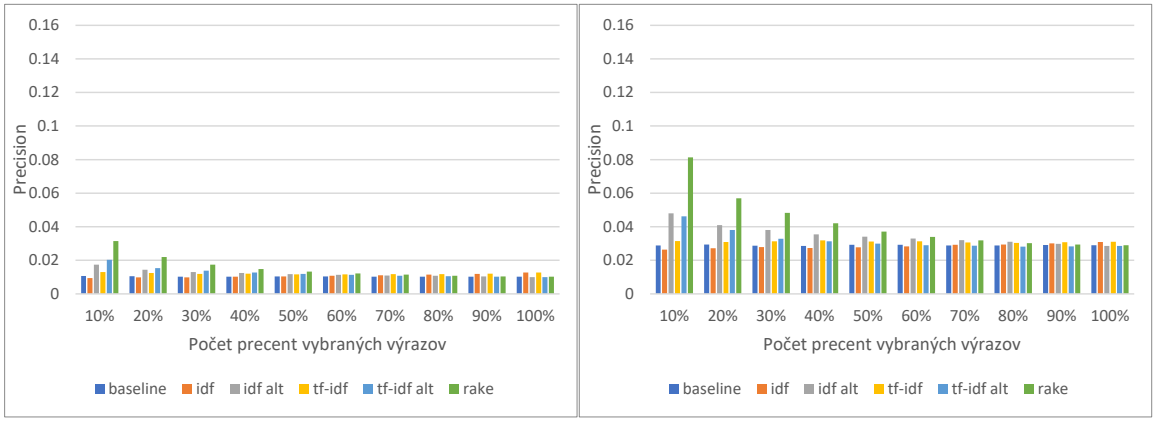
(d) 4 výskyty

Obr. 10.10: Lift grafy lingvistických metód dvojslovných výrazov

V grafe vidíme, že najlepšie si počína metóda RAKE, ktorá dosahuje najvyšší lift, a zároveň nestráca počet nájdených výrazov ako ostatné metódy. Lift oproti základnej krivke vidíme aj v metódach IDF a TF-IDF s alternatívnym spôsobom výpočtu IDF, pričom celkový rozdiel v nájdených výrazoch je podobný ako pri jednoslovných výrazoch. Pokles v nájdených výrazoch je však podstatne výraznejší v metódach IDF a TF-IDF, kde IDF bolo získané zo štatistík dvojslovných výrazov, z čoho vyplýva, že veľký počet výrazov nebol vôbec nájdený, čo má za následok podstatne horšie výsledky týchto metód. Grafy tiež ukazujú, že ako obvykle sa pri zvyšujúcej hodnote parametra znižuje efektívnosť jednotlivých algoritmov.

10.7.2 Grafy presnosti a úplnosti

Nasledujúce grafy ukazujú dosiahnuté hodnoty presnosti jednotlivých algoritmov:



(a) 1 výskyt

(b) 2 výskyty

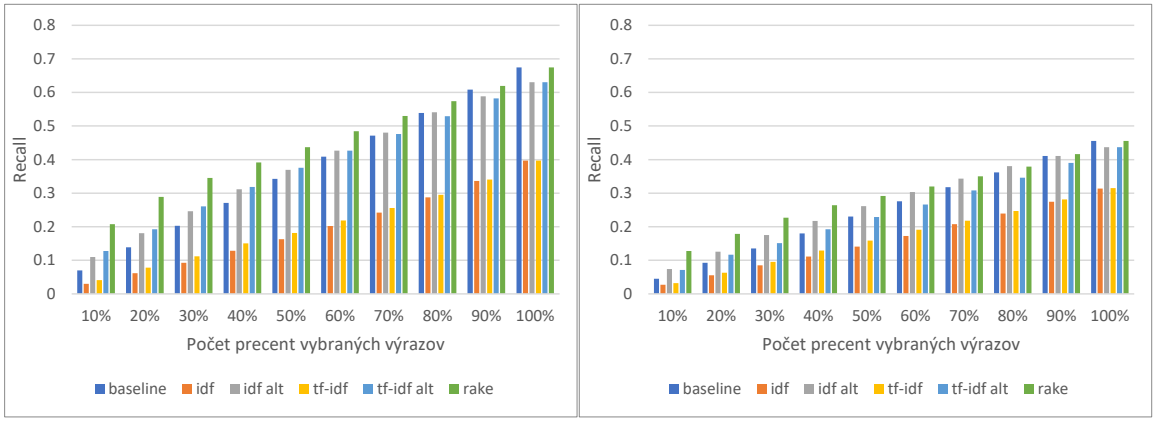
(c) 3 výskyty

(d) 4 výskyty

Obr. 10.11: Hodnoty presnosti lingvistických metód dvojslovných výrazov

Z grafoch vidíme, že výrazne najlepšie výsledky dosahuje algoritmus RAKE, pričom v najlepšom prípade ide o takmer 20 percent, čo sú výsledky porovnateľné so štatistickými metódami. Hodnoty IDF a TF-IDF sú podobné ako baseline, z čoho vyplýva, že bolo odstránené približne rovnako správnych a nesprávnych výrazov. Pri alternatívnom spôsobe výpočtu IDF dosahuje spočiatku lepšie výsledky TF-IDF, ale pri vyššom počte minimálnych výskytov výrazov má lepšie výsledky IDF, čo naznačuje, že počet výskytov stráca pri vyšších hodnotách relevanciu ako prediktor kľúčových výrazov. Napriek dobrým výsledkom RAKE je však prepad presnosti pri vyšších vybraných percentách výrazov podstatne rýchlejší ako u štatistických metódach.

Hodnoty úplnosti sú znázornené v nasledujúcich grafoch:



(a) 1 výskyt

(b) 2 výskyty

(c) 3 výskyty

(d) 4 výskyty

Obr. 10.12: Hodnoty úplnosti lingvistických metód dvojslovných výrazov

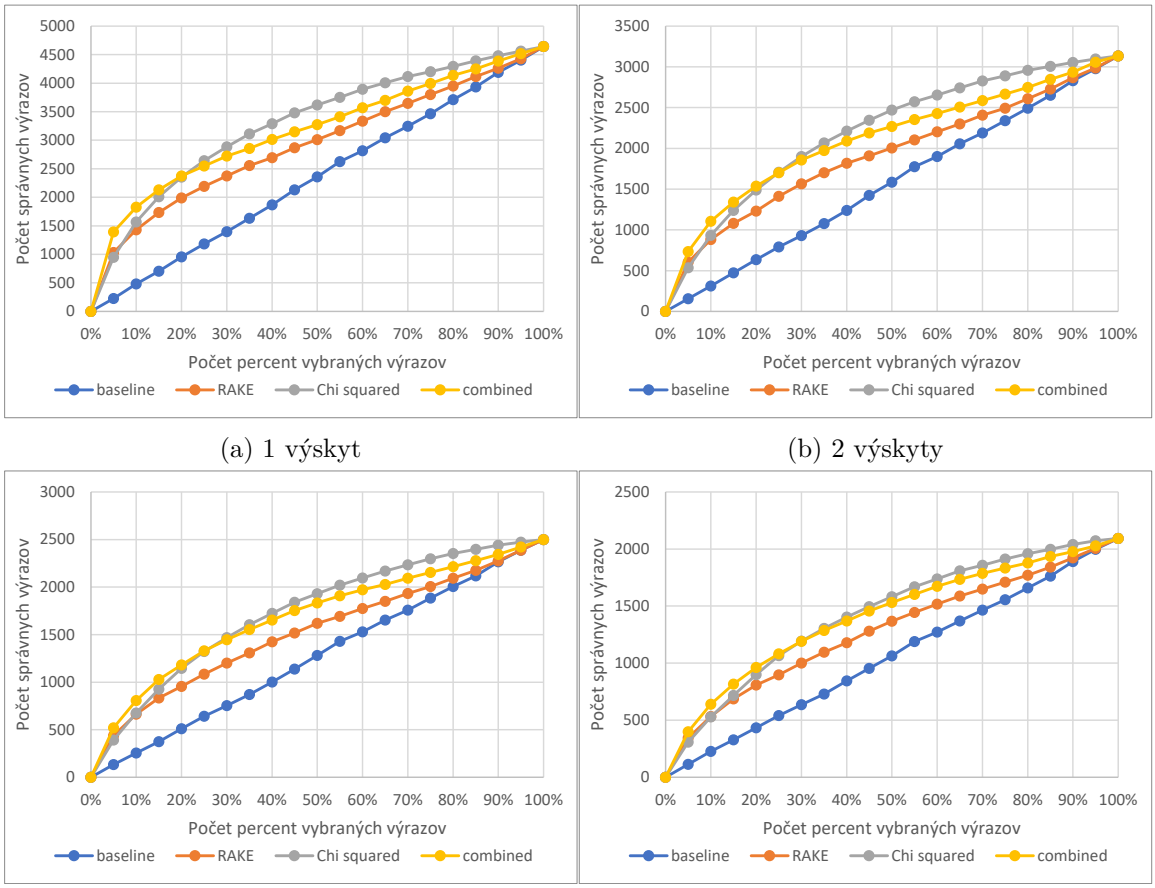
Vidíme, že rozdiel v hodnotách presnosti medzi algoritmom RAKE a základnou hodnotou je podstatne menší ako v štatistických metódach, čo znova ukazuje, že algoritmus RAKE dosahuje dobré výsledky len pri výbere malého percenta výrazov, a teda v kombinácii so štatistickou metódou bude mať potenciál zlepšiť hlavne hodnoty presnosti. Algoritmy IDF a TF-IDF s alternatívnym spôsobom výpočtu IDF dosahujú vo väčšine prípadov hodnoty podobné základu, a najhoršie výsledky sú dosiahnuté variáciami týchto algoritmov kde IDF je vypočítané ako početnosť dvojíc z dôvodov popísaných vyššie.

Najlepšie výsledky dosiahla vo všetkých kategóriách metóda RAKE, pričom najvýraznejšie rozdiely sú viditeľné pri nízkych percentách vybraných výrazov. Z toho dôvodu bude skúmaná kompatibilita kombinácie RAKE a štatistickej metódy Chi-Squared.

10.7.3 Analýza kombinácie algoritmov Ch-Squared a RAKE

Cieľom tejto časti je zistiť, či kombinácia najlepšie skórujúcich algoritmov štatistických a lingvistických prístupov bude znamenať celkové zlepšenie výsledkov, alebo bude musieť byť použitý len jeden z týchto algoritmov samostatne. Zo štatistických metód sa jedná o algoritmus Chi-Squred a z lingvistických metód ide o algoritmus RAKE. Keďže ide o fundamentálne rozdielne algoritmy, priemerná hodnota ich skóre nemusí byť rovnaká, čo môže mať za následok uprednostňovanie algoritmu s vyššou priemernou hodnotou skóre. Z toho dôvodu budú jednotlivé hodnoty skóre normalizované a nové skóre bude tvorené súčtom

týchto hodnôt. Výsledky budú zobrazené na lift grafoch, pričom budú porovnávané s výsledkami dosiahnutými jednotlivými algoritmami. Dosiahnuté výsledky algoritmov pri rôznych hodnotách parametra:



(a) 1 výskyt

(b) 2 výskyty

(c) 3 výskyty

(d) 4 výskyty

Obr. 10.13: Lift grafy pri kombinácii metód Chi Squared a RAKE

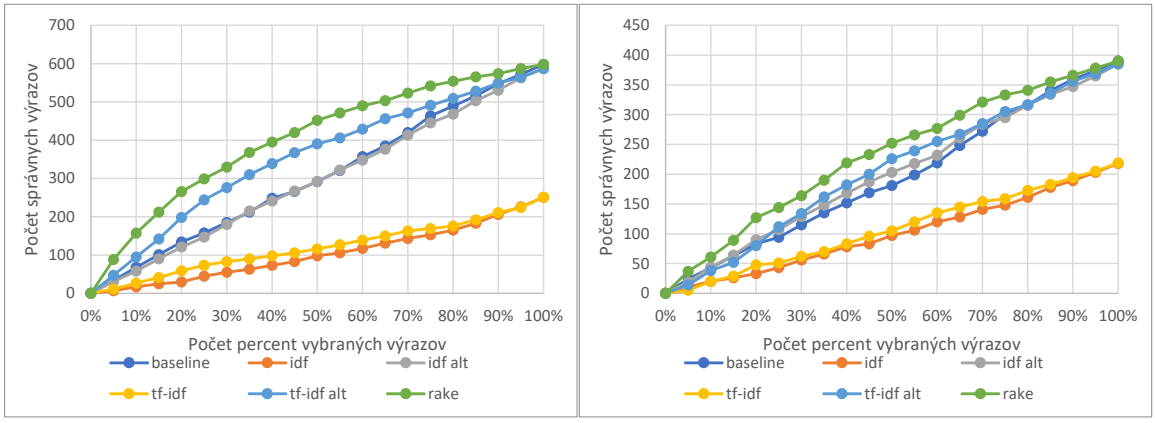
Z grafov vidíme, že kombinácia metód dosahuje lepšie výsledky len pri nižších percentách vybraných výrazov, pričom pri ostatných dosahuje lepšie výsledky Chi-Squared. Konkrétne sa hranica pri všetkých hodnotách parametru pohybuje okolo 20 percent, čo znamená, že využitie kombinácie jednotlivých algoritmov je efektívne pri snahe užívateľa dosiahnuť vysokú hodnotu presnosti. Použitý skórovací algoritmus teda bude rozdielny v závislosti na žiadanom počte percent vybraných výrazov.

10.8 Lingvistické metódy skórovania trojslovných výrazov

Podobne ako pri dvojslovných výrazoch boli tieto výrazy skórované štatistickými metódami, čo znamená, že môže pri dobrých výsledkoch dôjsť ku kombinácii týchto metód.

10.8.1 Lift grafy

Úspešnosti jednotlivých algoritmov sú znázornené v nasledujúcich lift grafoch:



(a) 1 výskyt

(b) 2 výskyty

(c) 3 výskyty

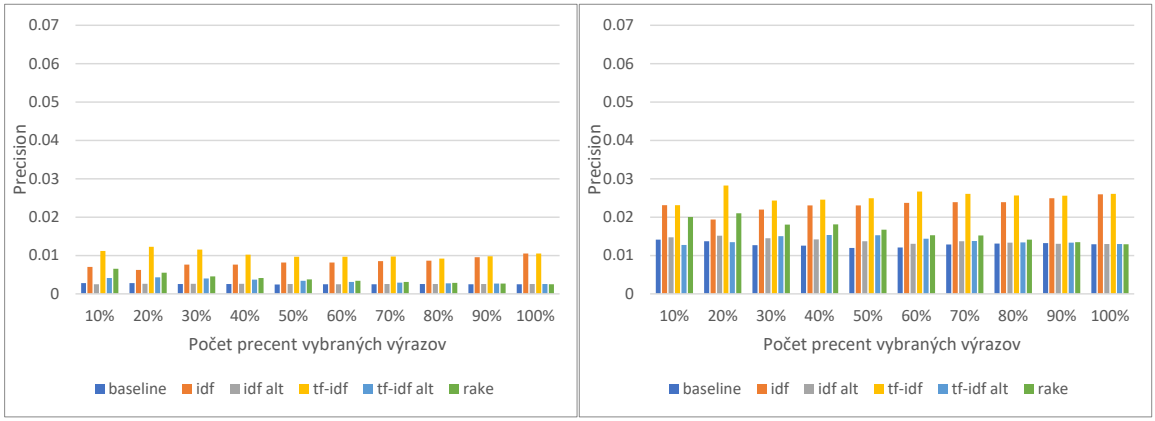
(d) 4 výskyty

Obr. 10.14: Lift grafy lingvistických metód trojslovných výrazov

Najlepšie výsledky dosiahla podobne ako v predošlom prípade metóda RAKE, avšak tieto výsledky sa výrazne zhoršujú pri vyšších hodnotách parametru. Takisto vidíme, že parameter IDF získaný alternatívnym spôsobom má pri trojslovných výrazoch takmer nulovú koreláciu s kľúčovosťou výrazu z dôvodu podobnosti priebehu so základnou krivkou, a teda vyšší lift metódy TF-IDF je možné pripísať len hodnote TF. Získavanie hodnoty IDF z početností trojslovných výrazov má za následok ešte vyšší počet nenájdených výrazov ako dvojslovné výrazy, kvôli čomu dosiahol algoritmus zlé výsledky.

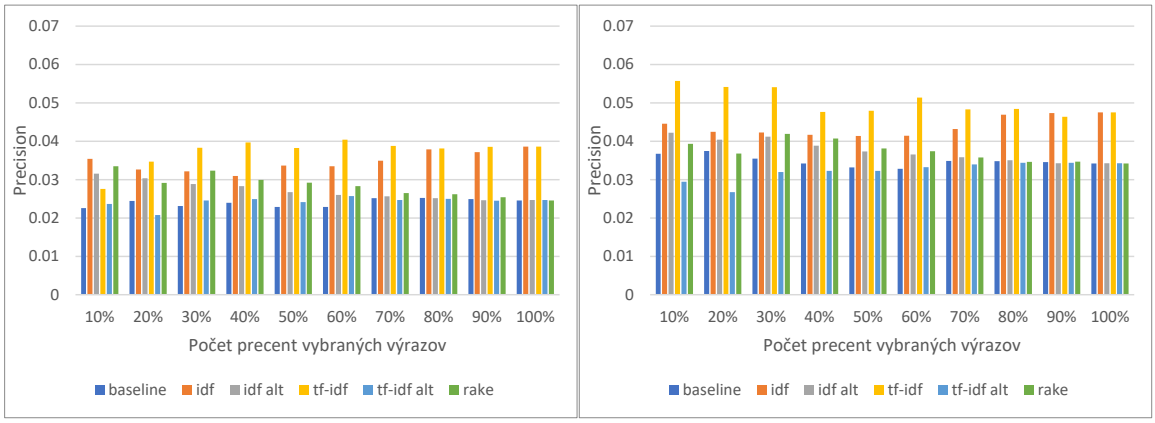
10.8.2 Grafy presnosti a úplnosti

Dosiahnuté hodnoty presnosti sú znázornené v nasledujúcich grafoch:



(a) 1 výskyt

(b) 2 výskyty



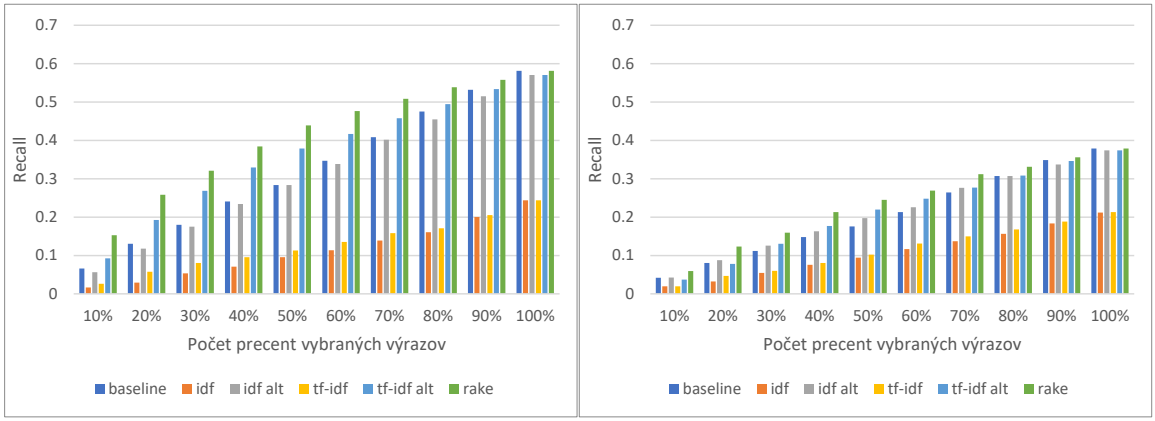
(c) 3 výskyty

(d) 4 výskyty

Obr. 10.15: Hodnoty presnosti lingvistických metód trojslovných výrazov

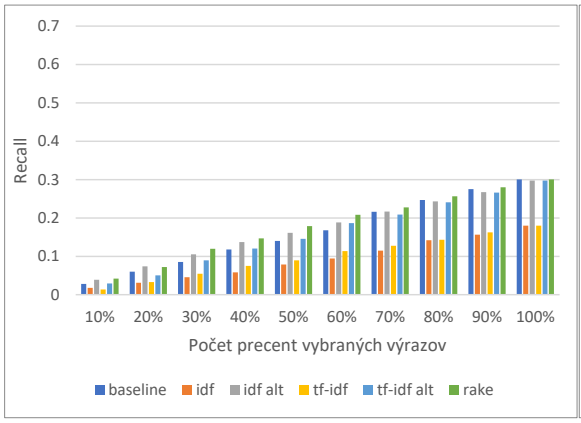
Prekvapivo najvyššie hodnoty sú dosiahnuté v algoritmoch IDF a TF-IDF, čo znamená, že z kandidátov bol odstránený väčší počet nesprávnych ako správnych kľúčových výrazov. Zlepšenie presnosti sa však pohybuje okolo hodnôt 1-2 percent, čo je príliš málo, hlavne v porovnaní so štatistickými metódami. Navyše sa dá z lift grafu predpokladať vysoký prepad hodnôt úplnosti v týchto metódach. Ostatné algoritmy majú z dôvodu malej miery zvýšenia lift len mierne lepšie hodnoty presnosti oproti základnej hodnote.

Nasledujúce grafy obsahujú hodnoty úplnosti jedotlivých algoritmov:

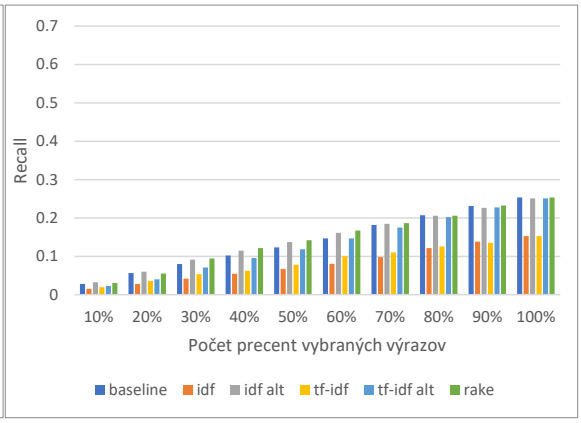


(a) 1 výskyt

(b) 2 výskyty



(c) 3 výskyty



(d) 4 výskyty

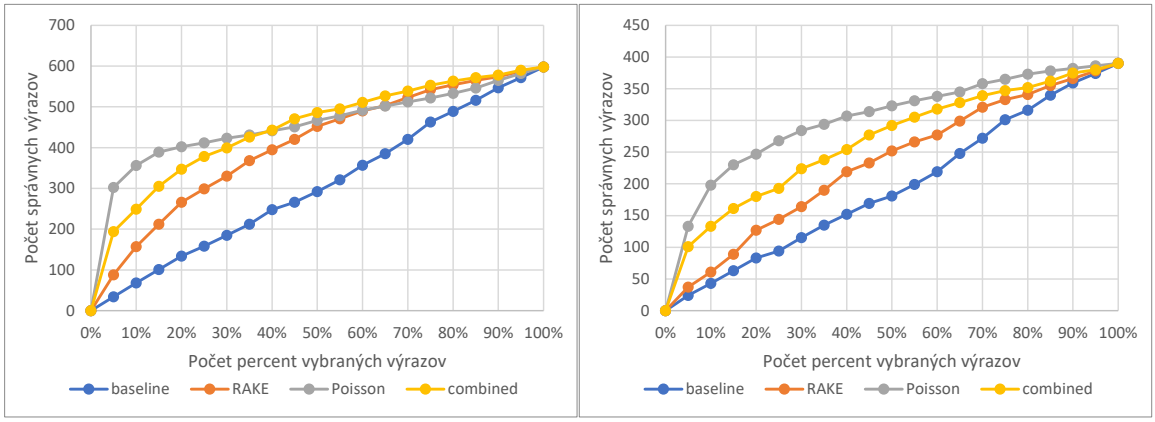
Obr. 10.16: Hodnoty úplnosti lingvistických metód trojslovných výrazov

Podľa očakávaní vidíme výrazný prepad hodnôt úplnosti u metód IDF a TF-IDF z dôvodu vysokého percenta eliminácie správnych výrazov. Algoritmus RAKE ukazuje určité zlepšenie hodnôt oproti základu, ide však o relatívne malé percento, hlavne v porovnaní so štatistickými metódami. Algoritmy IDF a TF-IDF s alternatívnym spôsobom výpočtu IDF ukazujú pri nižších percentách mierne zlepšenie a naopak pri vyšších percentách mierne zhoršenie výsledkov, pričom ich celková miera predikcie kľúčových výrazov je relatívne nízka.

V nameraných výsledkoch vidíme, že metóda ktorá dosiahla najlepšie výsledky je RAKE, teda bude testovaná jej kombinácia so štatistickou metódou Poisson-Stirling.

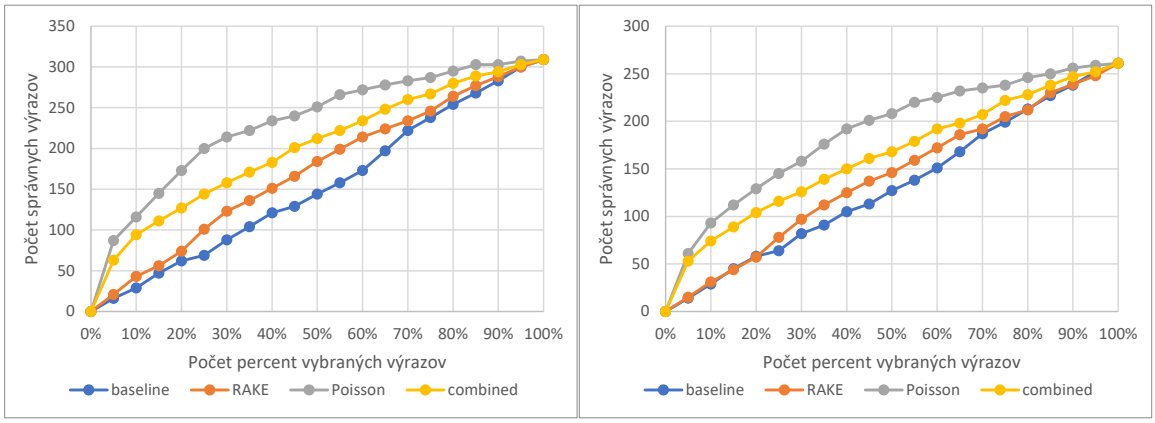
10.8.3 Analýza kombinácie algoritmov Poisson-Stirling a RAKE

Podobne ako u dvojslovných výrazoch budú testované a porovnávané výsledky kombinácie najlepšie fungujúcej štatistickej a lingvistickej metódy, konkrétne Poisson-Stirling a RAKE. Skóre bude opäť vypočítané súčtom normalizovaných hodnôt čiastkových skóre jednotlivých metód. Dosiahnuté výsledky sú ukázané na nasledujúcich grafoch.



(a) 1 výskyt

(b) 2 výskyty



(c) 3 výskyty

(d) 4 výskyty

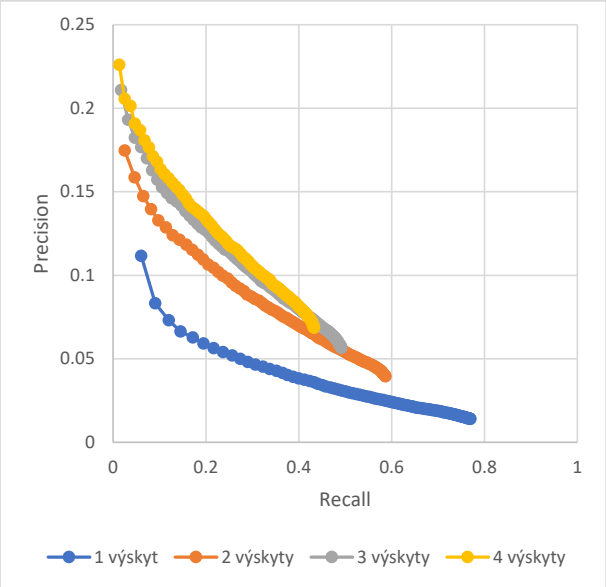
Obr. 10.17: Lift grafy pri kombinácii metód Poisson-Stirling a RAKE

Grafy ukazujú, že najvyšší lift dosiahol vo všetkých prípadoch algoritmus Poisson-Stirling, z čoho vyplýva, že jeho kombinácia s algoritmom RAKE nie je vhodná a teda na výpočet skóre trojslovných výrazov bude využívaný len štatistická metóda Poisson-Stirling.

Kapitola 11

Dosiahnuté výsledky

Výsledkom práce je program napísaný v jazyku Python, ktorý vykonáva automatickú extrakciu kľúčových výrazov z českých textov. Na obrázku 6.1 sú zobrazené dosiahnuté výsledky presnosti a úplnosti rôznych algoritmov na rozdielnych korpusoch v anglickom jazyku. Výsledky sa vzhľadom na testovaný korpus vyznačujú veľkou variabilitou v dosiahnutých hodnotách, čo indikuje vysokú mieru ovplyvniteľnosti dosiahnutých výsledkov výberom korpusu. Vytvorený algoritmus nie je možné testovať na týchto korpusoch, a teda porovnanie výsledkov bude mať len orientačný charakter. Dosiahnuté výsledky presnosti a úplnosti pri rastúcich hodnotách parametra sú zobrazené na nasledujúcom grafe:



Obr. 11.1: Závislosť dĺžky textu od počtu kľúčových výrazov

Na grafe vidíme, že medzi hodnotami úplnosti a presnosti existuje nepriama úmernosť, pričom presnosť v najlepšom prípade dosahuje 23 percent a úplnosť 78 percent. Dosiahnuté výsledky sú porovnateľné s výsledkami dosiahnutými v anglických textoch, sú však testované len na jednom korpuse, preto je ťažké posúdiť variabilitu výsledkov na iných rozsahoch textov a iných spôsoboch manuálneho výberu kľúčových výrazov. Z výsledkov je možné konštatovať, že implementované algoritmy sú efektívne pri identifikácii kľúčových výrazov a takisto aj pri predikcii ich kľúčovosti výpočtom skóre.

Kapitola 12

Záver

Cieľom práce bolo zoznámiť sa s možnosťami automatického výberu kľúčových výrazov, zhromaždiť literatúru na ich testovanie, analyzovať a porovnať jednotlivé metódy a vytvoriť nástroj pre automatický výber kľúčových výrazov. Boli skúmané rozdielne spôsoby výberu kandidátov a takisto aj rôzne štatistické a lingvistické metódy skórovania týchto kandidátov. Jednotlivé metódy boli analyované rôznymi kritériami, konkrétne porovnávaním hodnôt úplnosti a presnosti a analýzou kriviek v lift grafoch.

Na základe analýzy boli vybrané najvhodnejšie metódy pre jednotlivé fázy programu, konkrétne výber kandidátov je realizovaný na základe POS tagov, u jednoslovných kandidátov bolo skóre počítane pomocou lingvistickej metódy TF-IDF, dvojslovní kandidáti využívajú štatistickú metódu Chi-Squared, ktorú podľa požadovaných podmienok výstupu môžu kombinovať s lingvistickou metódou RAKE, a skóre trojslovných výrazov je počítané pomocou metódy Poisson-Stirling.

12.1 Možné vylepšenia

Ako vo všetkých projektoch, aj v tomto existujú možné vylepšenia funkcionality programu. Prvou možnosťou je implementácia a testovanie väčšieho množstva POS tagov, čo by mohlo mať za následok väčší počet nájdených výrazov. Program bol implementovaný spôsobom, ktorý podporuje jednoduché pridávanie a odoberanie POS tagov. Vytvorenie lepších IDF štatistík jednotlivých slov a výrazov by mohlo prispieť k zlepšeniu výsledkov algoritmov IDF a TF-IDF. Pre vierohodnejšie výsledky by bolo potrebné zhotoviť robustnejší korpus českých publikácií pre testovanie, a takisto objektívnejšie manuálne vytvorené zoznamy kľúčových výrazov. V neposlednom rade je potrebné testovať a porovnávať nové metódy extrakcie kľúčových výrazov.

Literatúra

[1] *A python implementation of the Rapid Automatic Keyword Extraction*. [Online; navštíveno 1.10.2016].
URL https://github.com/aneesha/RAKE

[2] Bougouin, A.; Boudin, F.; Daille, B.: *Topicrank: Graph-based topic ranking for keyphrase extraction*. 2013, [Online; navštíveno 14.4.2017].
URL http://www.aclweb.org/anthology/I13-1062

[3] Danesh, S.; Sumner, T.; Martin, J.: *SGRank: Combining Statistical and Graphical Methods to Improve the State of the Art in Unsupervised Keyphrase Extraction*. 2015, [Online; navštíveno 15.4.2017].
URL http://www.aclweb.org/anthology/S15-1013

[4] El-Beltagy, S.; Rafea, A.: *KP-Miner: A keyphrase extraction system for English and Arabic documents*. 2009, [Online; navštíveno 9.4.2017].
URL https://s3.amazonaws.com/academia.edu.documents/43133395/kpminer.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1493890281&Signature=YhznOcJcBAiBL9GbQtA5emD7DTE%3D&response-content-disposition=inline%3B%20filename%3DKP-Miner_A_keyphrase_extraction_system_f.pdf

[5] Evert, S.; Krenn, B.: *Computational approaches to collocations*. 2004, [Online; navštíveno 7.1.2017].
URL http://www.collocations.de/EK/Articles/MathAM.4up.pdf

[6] Hasan, K.; Ng, V.: *Conundrums in unsupervised keyphrase extraction: making sense of the state-of-the-art*. 2010, [Online; navštíveno 2.4.2017].
URL http://www.hlt.utdallas.edu/~vince/papers/coling10-keyphrase.pdf

[7] Jacquemin, C.; Bourigault, D.: *Term extraction and automatic indexing*. 2003, [Online; navštíveno 7.10.2016].
URL https://perso.limsi.fr/jacquemi/FTP/JacBourHandbookCL.pdf

[8] Justeson, J.; Katz, S.: *Technical terminology: some linguistic properties and an algorithm for identification in text*. 1995, [Online; navštíveno 2.12.2016].
URL https://anyall.org/JustesonKatz1995.pdf

[9] Lyse, G.; Andersen, G.: *Collocations and statistical analysis of n-grams*. 2012, [Online; navštíveno 10.1.2017].
URL http://dspace.uib.no/bitstream/handle/1956/11033/lyse-andersen-mwe-final.pdf?sequence=1&isAllowed=y

[10] Makhoul, J.; Kubala, F.; Schwartz, R.: *Performance measures for information extraction*. 1999, [Online; navštíveno 8.2.2017].
URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.28.276&rep=rep1&type=pdf

[11] Mihalcea, R.; Tarau, P.: *TextRank: Bringing order into texts*. 2004, [Online; navštíveno 13.4.2017].
URL https://digital.library.unt.edu/ark:/67531/metadc30962/m2/1/high_res_d/Mihalcea-2004-TextRank-Bringing_Order_into_Texts.pdf

[12] Pazienza, M.; Pennacchiotti, M.; Zanzotto, F.: *Terminology extraction: an analysis of linguistic and statistical approaches*. 2005, [Online; navštíveno 29.9.2016].
URL
https://art.torvergata.it/retrieve/handle/2108/46728/57697/ChurchH89

[13] Pianta, E.; Tonelli, S.: *KX: A flexible system for keyphrase extraction*. 2010, [Online; navštíveno 9.4.2017].
URL http://www.aclweb.org/anthology/S/S10/S10-1.pdf#page=192

[14] Ramos, J.: *Using tf-idf to determine word relevance in document queries*. 2003, [Online; navštíveno 19.11.2016].
URL
https://www.cs.rutgers.edu/~mlittman/courses/ml03/iCML03/papers/ramos.pdf

[15] Rose, S.; Engel, D.; Cramer, N.; aj.: *Automatic keyword extraction from individual documents*. 2010, [Online; navštíveno 17.10.2016].
URL
http://media.wiley.com/product_data/excerpt/22/04707498/0470749822.pdf

[16] Straková, J.; Straka, M.; Hajič, J.: Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore, Maryland: Association for Computational Linguistics, June 2014, s. 13–18.
URL http://www.aclweb.org/anthology/P/P14/P14-5003.pdf

[17] Vivaldi, J.; Rodríguez, H.: *Evaluation of terms and term extraction systems: A practical approach*. 2007, [Online; navštíveno 8.2.2017].
URL https://www.researchgate.net/profile/Jorge_Vivaldi/publication/233693935_Evaluation_of_terms_and_term_extraction_systems_A_practical_approach/links/00b495271268b9ec8d000000.pdf

[18] Vuk, M.; Curk, T.: *ROC curve, lift chart and calibration plot*. 2006, [Online; navštíveno 11.3.2017].
URL http://mrvar.fdv.uni-lj.si/pub/mz/mz3.1/vuk.pdf

[19] Xiaojun, W.; Xiao, J.: *Single document keyphrase extraction using neighborhood knowledge*. 2008, [Online; navštíveno 14.4.2017].
URL http://www.aaai.org/Papers/AAAI/2008/AAAI08-136.pdf

Prílohy

Príloha A

Manuál

V tejto prílohe je popísaná inštalácia potrebných balíkov a spustenie aplikácie. Predpokladá sa, že je nainštalovaný Python verzia 2.7. V nasledujúcej časti je popísaná inštalácia balíkov potrebných pre spustenie.

Inštalácia potrebných balíkov

- Morphodita: `sudo pip install ufal.morphodita`
- Six: `sudo pip install six`

Formát spustenia

```
python extract.py <filename> [-c <int>] [-p <int>]
```

Vysvetlenie parametrov

- `<filename>`: Meno súboru, v ktorom bude prebiehať extrakcia kľúčových výrazov. Súbor je v plaintext formáte.
- `-c <int>`: Nepovinný parameter. Určuje minimálny počet výskytov výrazov, aby mohli byť zaradené medzi kandidátov. Minimálna hodnota parametru je jeden. V prípade vynechania parametru pracuje program s hodnotou dva.
- `-p <int>`: Nepovinný parameter. Označuje, aký počet percent výrazov s najvyšším skóre bude selektovaný. Minimálna hodnota parametru je 1 a maximálna je 100. V prípade vynechania parametru pracuje program s hodnotou 100.

Príklad spustenia

```
python extract.py text.txt -c 2 -p 10
```

Výstupom programu je súbor `<filename>.keys` obsahujúci zoznam extrahovaných kľúčových výrazov podľa vybraných parametrov. V prípade, že má parameter `-p` hodnotu medzi 1 a 20, je pre výpočet skóre dvojslovných výrazov použitá kombinácia algoritmov Chi-Squared a RAKE, v ostatných prípadoch je použitý len algoritmus Chi Squared.

Príloha B

Obsah CD

- zdrojové súbory implementovaného nástroja v adresári `\src\`
- súbor dokumentácie `BC.pdf`
- zdrojové súbory dokumentácie v adresári `\doc\`
- súbor plagátu `poster.pdf`