VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

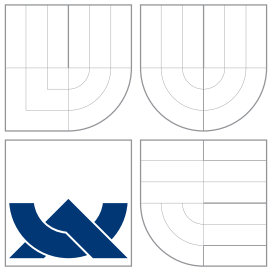
AUTOMATICKÁ IDENTIFIKACE KLÍČOVÝCH SLOV

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

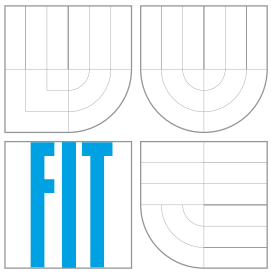
AUTOR PRÁCE
AUTHOR

MARCELA MAŠLÁŇOVÁ

BRNO 2007



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

AUTOMATICKÁ IDENTIFIKACE KLÍČOVÝCH SLOV

THE AUTOMATIC IDENTIFICATION OF KEYWORDS

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR

MARCELA MAŠLÁŇOVÁ

VEDOUCÍ PRÁCE
SUPERVISOR

doc. RNDr. PAVEL SMRŽ, Ph.D.

BRNO 2007

Automatická identifikace klíčových slov

Automatic Keyword Detection

Vedoucí:

Smrž Pavel, doc. RNDr., Ph.D., UPGM FIT VUT

Zadání:

1. Seznamte se s metodami vyhledávání klíčových slov.
2. Navrhněte a implementujte systém pro automatickou vyhledávání se zaměření na víceslovné výrazy.
3. Vyhodnoťte vytvořený systém pomocí standardních metrik.

Část požadovaná pro obhajobu SP:

1. prototyp systému

Kategorie:

Umělá inteligence

Literatura:

- podle dohody

Licenční smlouva

Licenční smlouva je uložena v archívu Fakulty informačních technologií Vysokého učení technického v Brně.

Abstrakt

Tato práce si klade za cíl zpracovat poznatky o značkování klíčových slov v textu a využít je v praxi pro automatické generování rejstříků. Důvodem pro automatizaci tvorby rejstříků je jejich vysoká náročnost a cena. Teoretická část práce se zabývá především metodami hledání vícenásobných výrazů, které jsou významné pro zpracovávaný text. Praktická část aplikuje vybrané metody na testovací data a shrnuje výsledky experimentů.

Klíčová slova

rejstřík, klíčová slova, vícenásobné výrazy, morfologická analýza, značkování

Abstract

The main goal of this work is to survey the field of the automatic keywords tagging in a text and apply this background for automatically generating back-of-the-book indexes. Human made indexes are expensive and that's why we are looking for (semi)-automatic methods indexes. The theoretical part of this thesis deals with collocations, which are an important part of generated indexes. The practical part of the work applies selected methods to testing data and summarize results of experiments.

Keywords

index, keywords, multi-words expresion, morphological analyse, tagging

Citace

Marcela Mašláňová: Automatická identifikace klíčových slov, diplomová práce, Brno, FIT VUT v Brně, 2007

Automatická identifikace klíčových slov

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracovala samostatně pod vedením pana doc. Pavla Smrže. Uvedla jsem všechny literární prameny a publikace, ze kterých jsem čerpala.

.....
Marcela Mašláňová
22. května 2007

Poděkování

Ráda bych poděkovala panu doc. Smržovi za vedení diplomové práce. Dále bych ráda poděkovala za rady Tomáši Janouškovi.

Obsah

Kapitola 1

Úvod

Zpracování textu se začalo ve větší míře zkoumat v 60. letech minulého století, kdy se výzkum v IT zaměřil na umělou inteligenci, zpracování řeči i textu. Od původní snahy využít nějaký racionální základ v pravidlech jazyků se upustilo a začalo se více využívat mechanického zpracování a statistických výsledků, které dosahují mnohem lepších výsledků. V současné době se automatické zpracování textu využívá v mnoha aplikacích. Většina textových editorů obsahuje alespoň nějakou jazykovou podporu, kontrolu pravopisu, dělení slov na koncích řádků, nabídku synonym atd. Některé programy poskytují i podporu tvorby rejstříků, která ovšem není příliš dokonalá. Obvykle dokáží pouze označená slova vypsat na konci textu s číslem příslušné strany, kde se výrazy vyskytují. I s programovou podporou pro vkládání rejstříku musí člověk obvykle sám vybrat slova a slovní spojení, která mají být do rejstříku zařazena. Pro rozsáhlé texty to může být časově náročné, zvlášť pokud rejstřík tvoří někdo jiný než autor.

Na profesionální úrovni se tvorbou rejstříků zabývají obvykle v nakladatelstvích specializovaných na literaturu populárně-naučnou, literaturu faktu nebo na učebnice. Rejstříky napomáhají orientaci v textu a urychlují vyhledávání zvoleného tématu.

Časová náročnost tvorby rejstříků vede ke snaze vytvořit nástroj, který by dokázal rejstříky tvořit automaticky anebo alespoň tvorbu rejstříku urychlil. S rychle narůstajícím množstvím článků není možné je všechny ručně procházet a vytvářet pro každý rejstřík. V textech se objevují stále nová jména, společnosti či výrobky, což dělá problém programům pro automatickou tvorbu rejstříků, která je založena na slovnících. Může se stát, že text je indexován pokaždé jinou společností pro jiné účely a pojmy v rejstříku se pro jeden text mohou diametrálně lišit. Pak je lepší i nedokonalý automatický rejstřík než několik různých. Tyto problémy vedly ke vzniku LinkIT [11], který měl za úkol vyřešit indexování univerzitních článků.

Firmy zabývající se tvorbou rejstříků používají profesionální programy, které nabízejí podporu tvorby rejstříků. Podpora tvorby znamená pouze nabídku možných klíčových slov, které by se v rejstříku mohly vyskytnout, programy nedokáží vytvořit celý rejstřík automaticky. Nabídku dokáží vytvářet obvykle podle některého z těchto kriterií:

- četnost slov — uživatel nastaví maximální počet výskytů slov. Např. nastaví výběr slov, která se vyskytují 50 a méně krát. Tím omezí množství výrazů, ze kterých se budou vybírat výrazy do rejstříku.
- seznam všech vlastních jmen vyskytujících se v textu.
- seznam označených slov na základě ručního výběru.

- seznam na základě frází — např. frázová slovesa, některé programy mají slovník víceslovných klíčových výrazů.

Všechny metody kromě ručního označení klíčových slov jsou nespolehlivé, proto je zde možnost výběru klíčových slov ponechána na uživateli.

Pokud uživatel sám označuje slova v anglickém textu, pak není takový problém s vyhledáním slov v různých tvarech, protože obvykle postačí odtrhnout koncovky jako *s* a varianty s apostrofy. Pro češtinu je nutné se zabývat koncovkami slov více, protože jedno slovo se může v textu vyskytovat v mnoha tvarech např. vystupňovaná přídavná jména, skloňovaná podstatná jména nebo časovaná slovesa. Problém v českém textu nastává také tehdy pokud jsou v textu použity pojmy z cizích jazyků nebo jsou používány poměrně nové, slovníku neznámé, termíny.

Problematika tvorby rejstříků bude podrobněji popsána v podkapitole 2.1. Automatická tvorba rejstříků je založeno na vyhledávání významných vícenásobných výrazů a jednoslovných klíčových slov, jejíž základy jsou popsány v podkapitole 2.2. Vyhledávání výrazů je založeno především na četnosti výskytu slov v textu. Metody, kterými je možné vícenásobné výrazy získat, jsou detailně popsány podkapitolou 4, hledáním jednoslovných výrazů se zabývá podkapitola 3.1. Tato kapitola se také zabývá problematikou relevance nalezených vícenásobných výrazů. V kapitole 3.2 jsou rozebrány metody, kterými se tyto výrazy dají vybrat. Testovací data jsou uvedena v kapitole 5.1, kde je také popsán systém, kterým jsou detekovány klíčová slova a generovány rejstříky. Předposlední kapitola 6 se zabývá metodologií hodnocení a jednotlivými experimenty prováděnými nad daty.

Tato práce si klade za cíl zjistit nakolik lze pomocí metod zpracování přirozeného jazyka zautomatizovat tvorbu rejstříků. Velký důraz je zde kladen na praktickou použitelnost výstupu a časovou náročnost. Cílem práce je zjistit, jestli použité metody dokáží urychlit tvorbu a zpracování rejstříku uživateli. Experimenty se zaměřily na to, které techniky jsou nezbytně nutné pro tvorbu rejstříků např. lemmatizace, určení slovních druhů aj.

Zkoumání, zda nabízený rejstřík obsahuje relevantní data, se bude ověřovat na literatuře s profesionálně vytvořenými rejstříky. Vzhledem k tomu, že literatury s kvalitně vytvořenými rejstříky je málo, musí se počítat s tím, že každý systém nabídne širší množství dat, než které je uváděno v knihách. Protože je každý rejstřík vytvořen poněkud jinak a s normou [1] se zachází spíše jako s doporučením, je třeba počítat s problémy *automatického hodnocení*. Dalším problémem je, zda kvalita automaticky generovaných rejstříků závisí na délce textu.

Text studijní opory k předmětu *Zpracování řečových signálů* by měl pomoci, odpovědět na některé z těchto otázek. Je na něm provedena případová studie, která je popsána v kapitole 6.5 sloužící jako závěrečný test pro vytvořený systém.

Kapitola 2

Vymezení pojmů

2.1 Rejstříky

Pro vyhledávání položek v rejstříku je nutné si uvědomit, jaká slova nebo víceslovné výrazy se v něm vyskytují. Rejstřík je podrobný, obvykle abecedně seřazený seznam pojmů publikace, např. knihy. Je vytvářen proto, aby pomohl čtenáři najít informace snadno a rychle. V ideálním případě není rejstřík pouhým seznamem základních pojmů publikace, ale uspořádaným seznamem položek zahrnujících křížové odkazy[1]. Rejstříky mohou obsahovat zkratky případně i jména citovaných autorů. Pokud je citovaných autorů nebo jmen v rejstříku více, používá se pro ně jmenný rejstřík, a ostatní klíčové výrazy jsou uvedeny v obyčejném rejstříku. Podle normy [1] by se měla dodržovat tato doporučení:

> Kvalitně vytvořený rejstřík vychází z toho, co budou potencionální čtenáři hledat. Délka rejstříku má dosahovat 10–15% stran pro vědecké publikace, 5% pro ostatní žánry. V rejstříku se doporučuje používat uspořádání slov — podstatné jméno, přídavné jméno následované číslem strany. Doporučuje se nepoužívat nebo ignorovat vlastní zájmena. Pokud existuje více položek druhého řádu, je přehlednější vytvořit jemnější strukturu s dalšími odrážkami jako v tabulce 2.1. Položky rejstříku by neměly začínat velkým písmenem pokud to nejsou vlastní jména. Pokud je uváděn i seznam ilustrací, musí mít vlastní rejstřík ilustrací.

Doporučované seřazení pojmů v rejstříku je uvedeno na příkladech tabulek 2.1 a 2.2, tabulkou 2.1 je ilustrován případ víceúrovňového rejstříku s jemnou strukturou.

programování	- genetické, 98, 112, 114
	- systémů, 56, 57
	- embedded, 56, 110, 111
	- operačních, 56, 59, 112
	- profesionální, 3

Tabulka 2.1: Základní tvar pojmů v rejstříku — typ I.

V rejstříku se obvykle slova uvádějí v základním tvaru. Základní tvar bývá v prvním pádu pro řídící podstatné jméno a infinitiv pro slovesa (pokud se sloveso vyskytne). V reálných rejstřících jako tabulka 2.2 se ovšem nejčastěji vyskytují jednoslovné nebo

[1] Křížové odkazy — obvykle se používají pro odkazování se na předchozí nebo následující části textu.

genetické programování, 98, 112, 114
grafická karta, 114, 115
⋮
programování operačních systémů, 3

Tabulka 2.2: Základní tvar pojmů v rejstříku — typ II.

dvouslovné výrazy. Pro dvojice jsou pak nejcharakterističtější výskyty z tabulky 2.3. Složitější a lépe vytvořené rejstříky mohou být víceúrovňové jak je uvedeno na příkladě 2.1. Tvorba víceúrovňových rejstříků je nejnáročnější, ale pro čtenáře nejpřehlednější. Taková struktura půjde v českém textu automaticky vytvořit velmi těžko kvůli poskládání slov z výrazu ve správným pořadí a tvaru. Bližší informace o tvorbě rejstříků lze najít pod normou ČSN ISO 999 - 1998 [1].

Tvary slov	$\sim$ %
přídavné jméno + podstatné jméno v prvním pádě	80
podstatné jméno + podstatné jméno v druhém pádě	18
ostatní tvary	2

Tabulka 2.3: Charakteristické tvary slov v rejstříku

2.2 Typy víceslovných výrazů

V literatuře se vyskytuje mnoho pojmů souvisejících s kolokacemi, jejichž definice se nemusí shodovat. Tato práce čerpala definice a rozdělení kolokací především z [8], podle které označujeme jako *kolokaci* několik slov, která na sebe mají syntaktickou a sémantickou vazbu. Tj. skupina slov, která popisuje nějakou skutečnost a přitom z jednotlivých slov nemusí být zřejmý jejich význam. Kolokace mají tyto charakteristické vlastnosti:

- omezenou kompozicionalitu — tzn. z jednotlivých slov není možné určit význam kolokace. Často citovaný příklad je *silný čaj*, kde silný vypovídá o kvalitě čaje a ne o fyzické síle. V některých případech se může význam jednotlivých slov naprosto lišit od významu kolokace. Takovým extrémním případem jsou idiomy, např. *dát si do těla.*

- omezenou substituovatelnost — nelze nahradit slovo jiným slovem, i když popisuje stejnou vlastnost např. *bílé víno* je spíše žluté barvy, ale nikdy o něm takto nemluvíme.

- omezenou modifikovatelnost (přizpůsobitelnost) — některé kolokace, především idiomy, nemohou být rozšířeny o další lexikální[2] jednotky, ani nemohou být ohýbány, protože by se změnil jejich význam. Např. přijít na buben — přijít na zelený buben.

Kolokace zahrnují pojmy jako: *klíčová slova, idiomy, termíny, typická spojení* a další.

Klíčová slova mohou být jednoslovná nebo víceslovná. Jsou to podstatné pojmy v textu, pojmenovávají problematiku, kterou se text zaobírá. Příkladem jednoslovného klíčového

[2]Tj. slova i slovní spojení

slova může být třeba *impresionismus* v Dějinách umění a pro víceslovná např. *renesanční malířství*. Pokud se klíčová slova týkají určité problematiky, mluvíme o termínech používaných v tomto oboru, např. *základní deska* v architektuře počítačů. Hlavní rozdělení kolokací založené na sémantice a syntaxi:

- typická spojení
- termíny
- idiomy

Co je to termín bylo uvedeno výše, zbylé dva pojmy se často překrývají.

Obvykle je idiomům věnována jen malá pozornost, protože se mohou překrývat s víceslovnými výrazy, a jen těžko lze určit přesnou hranici mezi nimi. Člověk určí význam idiomu podle sémantického významu, např. typický český idiom je *natáhnout bačkory, dát si do nosu*. Pokud člověk nezná význam idiomu, nepozná z jednotlivých slov, o co se jedná. Obvykle bývá idiom v každém jazyce odvozen od jiných slov (*natáhnout bačkory — kick the bucket*), proto je předmětem mnoha výzkumů vyhledávání idiomů a víceslovných výrazů. Ve slovnících jsou potřeba nejen idiomy a víceslovné výrazy, ale i volnější víceslovné výrazy jako případy, která předložka se pojí s kterým slovesem. Automaticky se dají získávat i takovéto výrazy a jejich hledáním se mimo jiné zabývá [5]. Mezi idiomy nepatří už takové víceslovné výrazy, které mají příliš volný význam. Na základě jednotlivých slov se nedá pochopit jejich smysl, ale vazba není tak pevná jako u idiomů uvedených výše např. frázové sloveso [13] a podmět *zavolat lékaře — call for - doctor*. Z jednotlivých slov tady lze určit smysl, přesto je frázové sloveso ustálená vazba. Existují i slovní spojení (především v angličtině) jako *fire away*[3], která mají naprosto odlišný význam od významu jednotlivých slov.

Typické výrazy — víceslovné výrazy lze ilustrovat na výrazu *dobře (s někým) vycházet* (*get away with*), což jsou víceslovné slovesné výrazy (multi-word verbs). Takové výrazy lze rozdělit do podtříd na frázová slovesa (typická především pro angličtinu) a slovesa pojící se s předložkami. Všechny tyto výrazy se chovají jako jedno slovo. Termín „slovo“ je často používané nejen v morfologickém smyslu, ale také pro položky, které se chovají jako jedna entita lexikálně a syntakticky. Pojem víceslovné výrazy a další podtřídy byly zavedeny, pro výstižnější pojmenování skupin chovajících se podle určitých pravidel nebo charakterizovaných podobnými vlastnostmi.

Víceslovné výrazy založené na předložkách nepokládáme za idiomy, přestože jsou na sebe pevně vázány. Je totiž těžké odlišit, co je pouze předložka a co je víceslovný výraz např. *narozdíl od, co se týče* (*apart from, as for*). K podobné situaci dochází u podstatných jmen jako *asistent ředitele — assistant director*. Podstatná jména mají ustálený výraz, ale není to idiom narozdíl od spojení typu *ruku v ruce - arm in arm*. Zvláštní postavení v kolokacích zaujímají vlastní jména. Patří mezi kolokace, ale často jsou nežádoucí a je třeba je odfiltrovat.

Kolokace se také dají rozdělit podle toho, jak se vyskytují ve větách. Obvykle se jednotlivá slova kolokace vyskytují vedle sebe, některé kolokace mohou být oddělené dalšími slovy tzv. *kolokace s dírami*. Takový příklad je uveden níže ve třetí větě.

Příklady vět s různě spořádanými klíčovými slovy:

- Objekty lze najít v budově Akademie věd na Národní třídě, *vstup* je *volný*.

[3] fire away — spustit palbu, doslovně by to bylo pálit do dálky nebo daleko

- *Volný vstup* vám nemůžeme zaručit.
- *Vstup* na výstavu je *volný* pouze v pondělí.

Klíčovými slovy (tedy kolokací) je v tomto případě *volný vstup*.

S automatickým zpracováním jazyka jsou spojeny další pojmy jako *lemmatizace* — převod (nejen) kolokace na základní tvar, např. *operačními systémy* na *operační systém*. Zpracování textu se provádí na rozsáhlých textech — *korpusech*. Korpusy jsou texty upravené s ohledem na automatické zpracování textu. Obvykle jsou uloženy ve formátu slovo na řádek. Pro zachycení vnitřní struktury textu jako jsou třeba nadpisy se používají značkovací jazyky např. SGML[4], takové korpusy jsou (zatím) spíše vyjímkou. Korpusy mohou být vytvořeny z knih, novinových článků, z internetových článků aj. Podle toho, k čemu je korpus potřeba, se vybírají data, případně jazyk nebo jazyky.

2.3 Předzpracování textu

Jednoslovné výrazy mohou tvořit značnou část rejstříku. Jejich vyhledávání je založené na četnosti výskytu slov v textu. Takto se na výstup dostane značné množství pomocných slov, která se dají odfiltrovat *stop-listem*[5], který obsahuje nejčastěji používaná slova, především *funkční slova*[6]. Odfiltrování slov můžeme založit na označení slov slovními druhy, a pak vynechat pomocná slova. Pro další zúžení výběru slov lze použít některé z dále popsaných metod z kapitoly 4.

O něco složitěji jsou zpracovávány **Víceslovné kolokace**, ale také se zde uplatňuje filtrace stop-listem. Obecné rozdělení metod vyhledávání víceslovných kolokací:

- frekvenční vyhledávání (n-gramy mohou být vytvářeny podle různých pravidel)
- hledání na základě lingvistických poznatků — gramatické vzorce
- kombinace metod

Vyhledávání bigramů[7] je také založeno na frekvenci výskytu v textu. Pouhým frekvenčním hledáním lze získat kolokace jako v tabulce 2.4. Jak je vidět mezi prvních pár nejčastěji se vyskytujících bigramů, se nedostala žádná slova, která bychom mohli považovat za typickou kolokaci. Tento problém se běžně řeší tak, že se vytvoří stop-list všech často se opakujících spojení, které nám nevyhovují. Je to nejrychlejší řešení, ale je třeba dobrý slovník se všemi pomocnými slovy, ve všech tvarech, který se na odfiltrování použije. Další možností je použít značkovač, který určí slovní druhy bigramů. Pak stačí označit slovní druhy, které budou z výstupu odfiltrovány.

Stop-list založený pouze na odstranění pomocných slov by odstranil např. *se* nebo v, ale bigramy jako *v roce* by zůstaly, což můžeme za klíčové považovat jen těžko. V angličtině je to celkem triviální problém, protože pro určení slovních druhů existuje jen pár pravidel. V češtině je to mnohem komplikovanější kvůli ohýbání slov [8].

[4] Standard Generalized Markup Language
[5] Stop-list — seznam slov, která jsou vyřazena z dalšího zpracování.
[6] Předložky, spojky, některá zájmena, pomocná slovesa, aj.
[7] Kolokace, která je dvouslovným výrazem
[8] Skloňování, časování

Počet výskytů	Bigram
4403	v roce
3286	a to
3101	je to
2880	a v
2743	více než
2628	se v
2624	ale i
210	microsoft windows
210	musí mít
210	mají být
209	systémů a
209	Hradec Králové
209	o výkonu
209	určené pro

Tabulka 2.4: Bigramy — frekvenční vyhledávání

Frekvenční vyhledávání je možné založit na prostém procházení textu nebo použít „okénko“ — vytvářející bigramy ze slov před (za) aktuálním slovem. Metod získávání bigramů je více. Po odfiltrování pomocných slov by mělo být jasnější, zda některá metoda dává výrazně lepší výsledky nebo jestli vycházejí zhruba stejně.

V tabulkách 2.4 a 2.5 byla nalezena vlastní jména. V některých textech nemusí být žádaná, a v takovém případě je třeba mít rozsáhlý slovník, který je dokáže rozlišit.

Většina citované literatury se zabývá především hledáním bigram. Kolokace jsou často složeny z více než dvou slov, a proto jsou v této práci zmiňovány trigramy a delší kolokace. V [8] je zmiňováno hledání bigramů a trigramů. Pokud je z jednoho textu vygenerován seznam bigramů i trigramů, pak mohou být vygenerovány bigramy, které jsou částmi trigramů. Např. v případě *hrubého domácího produktu* z tabulky 2.6 je plnohodnotnou kolokací trigram, protože z bigramu *hrubý domácí* nelze určit smysl. U ostatních dvou mohou být správně bigramy i trigramy. Automatické zjištění, jestli je správný delší nebo kratší termín, lze založit na četnosti výrazů.

Po vygenerování bigramů (n-gramů) je na výstupu spousta zbytečných slov, která se dají odfiltrovat na základě:

- slovníku
- seznamu výrazů
- podle slovních druhů

Jednou možností je vytvořit stop-list, který je třeba naplnit všemi kombinacemi předložek a spojek, zájmeny a dalšími. Pro vytvoření je možné použít již hotový stop-list pomocných slov a doplnit jej nejčastějšími slovy. Po odstranění slov ze stop-listu by zbyly pouze výrazy vyhodnocené jako kolokace, které se převedou do základního tvaru. Daly by se převést pouze ty nejčastější, ale některá klíčová slova se v textu mohou vyskytovat pouze zřídka a nemusely by se mezi nejčastějšími objevit.

Další možný přístup je určit slovní druhy a určit takové kombinace slovních druhů, které se nebudou v bigramech vyskytovat. Typickým příkladem bude bigram obsahující

Počet výskytů	Bigram
6006	že a
5933	že v
5658	že se
4860	že na
4314	v a
4180	že je
3720	v se
79	microsoft windows
79	Čsn en
79	pro windows
78	v zájmu
78	kontakt s
78	Ctibor Čejpa
78	pro grafiky

Tabulka 2.5: Bigramy hledané „okénkem“ — příklad z testovacích dat

Trigramy	Bigramy A	Bigramy B
microsoft windows nt	microsoft windows	windows nt
ochrana životního prostředí	životní prostředí	ochrana životní
hrubý domácí produkt	domácí produkt	hrubý domácí

Tabulka 2.6: Trigramy vs. bigramy

slova — předložka následovaná spojkou, libovolný slovní druh následovaný předložkou atd. Pro přesně definovaný problém je možné použít opačný přístup a vytvořit vzorce slovních druhů. Pokud by bylo cílem, třeba v anglickém textu, získat frázová slovesa [13], pak by se dal vytvořit vzorec: *Sloveso Předložka Libovolný slovní druh*, kde po předložce následuje slovo libovolného slovního druhu.

Značkování slovních druhů lze provádět různě složitými metodami, záleží na požadované úspěšnosti. S dostatečně velkým označkovaným korpusem dat je možné vytvořit si *trénovací a testovací množinu*. K datům z testovací množiny pak stačí vyhledat stejná slova v trénovací množině a přiřadit slovní kategorie. Problém nastane, pokud se jedná o trénovací množině neznámé slovo.

Mezi další jednodušší postupy patří již dříve zmiňovaný morfologický analyzátor. **Morfologie** (tvarosloví) je věda zabývající se ohýbáním a odvozováním slov pomocí předpon a přípon. Slova každého jazyka jsou sestavená z jednoho či více *morfémů* — nejmenší jazykové jednotky s identifikovatelným významem. Touto cestou je možné podle koncovek jednotlivých slov zhruba určit slovní druhy, např. *barevný* podle koncovky *-ný* bude pravděpodobně přídavné jméno.

Určováním slovních druhů se zabývalo již mnoho výzkumů. Mezi české morfologické analyzátory patří projekt Masarykovy univerzity — [10], který je založen na morfologii českého jazyka. Na Univerzitě Karlově vznikl projekt — PDT [7].

PDT pracuje se slovníkem příslušného jazyka. Zvláštní pozornost tvorbě *morfologického analyzátoru* byla věnována předponám, které se ve slovanských jazycích hodně vyskytují

Počet výskytů	Trigram
1890	v současné době
1103	v České republice
709	v roce 1994
694	v roce 1995
556	jedná se o
550	na rozdíl od
525	ve srovnání s
177	ochrany životního prostředí
176	a to v
174	a východní Evropy
173	Čr a Sr
172	v některých případech
172	ale i v
172	o více než

Tabulka 2.7: Trojice po sobě jdoucích (sousedících) slov

(nej-, ne-), a proto je použitelný obecněji pro slovanské jazyky.

Značkovačem se myslí program, který dokáže slovům přiřadit mluvnické kategorie. Morfologický analyzátor dokáže pouze navrhnout kategorie a lemmu. Při zpracování značkovačem za použití slovníků může opět nastat problém s neznámými slovy, která se dají řešit těmito způsoby:

- přiřadit nejčastější slovní druh z trénovací množiny
- vytvořit bigramy/trigramy, které rozhodnou na základě pravděpodobností po sobě jdoucích slov, o jaký slovní druh se bude jednat
- „uhodnout“ slovní druh z koncovky na základě morfologické analýzy

Neznámá slova podle statistiky bývají především:

- vlastní jména
- přídavná jména
- ostatní — zkratky, značky aj.

Největší část neznámých slov tvoří vlastní jména. Částečně se problém jejich určení dá vyřešit použitím slovníku jmen, ale stále zůstává velká část vlastních jmen jako jsou názvy, např. společností nebo výrobků. Druhou kategorii neznámých slov tvoří především cizí slova ve formě přídavných jmen, např. *aglutinační*. Taková slova se dají aspoň částečně určit podle koncovky, čímž se kategorie neznámých slov opět zužuje. Zbývající část je zlomkem oproti dvěma prvním kategoriím. Do této kategorie spadají zkratky, které mohou být často používané v daném oboru, z kterého text čerpá, ale v běžném slovníku se vůbec nemusí vyskytovat. Totéž platí i o podstatných jménech, která budou běžná pro určitý obor, ale slovníku naprosto neznámá.

Kapitola 3

Základní techniky vyhledávání kandidátů na rejstříková hesla

Rejstříky obsahují především jednoslovné a dvouslovné výrazy. Zjednodušeně lze říct, že problematika vyhledávání výrazů v rejstříku je problematikou hledání klíčových slov — unigramů a kolokací (bigramů, trigramů).

3.1 Jednoslovné výrazy

Jednoslovné výrazy, jinak nazývané unigramy podle množství slov ve výrazu, jsou v textu hledány na základě frekvence výskytu. Problém nastane u příliš krátkých textů nebo třeba manuálů[1], kde je zmíněno klíčové slovo pouze v nadpisu.

Pro takové případy je možné hledané unigramy více ohodnotit, to platí např. pro nadpisy nebo slova vysázené tučně (pokud bude zpracovávaný text takto označen). Hodnota skóre se zvýší poměrně k obodování ostatních vyhledaných výrazů. Dá se očekávávat, že nadpisy budou stejně jako text obsahovat spoustu pomocných slov a nadpisy jako *úvod* nebudou vhodnými kandidáty rejstříku. Proto se seznam unigramů vyfiltruje stop-listem často užívaných slov nebo se použijí modely popsané v podkapitole 3.2

3.2 Víceslovné výrazy

Pro zúžení množství vyhledaných n-gramů z jednoho korpusu lze použít filtraci **backgroundovým modelem**, který může být: *korpusový* — obecný anebo *doménovy* — specializovaný. Korpusový background model — se vytváří z dostatečně velkého obecného korpusu (milióny slov). Pokud se v obecném korpusu některé slovo vyskytuje jen výjimečně a ve zkoumaném textu mnohokrát, pak zřejmě půjde o text zabývající se problematikou vztahující se k tomuto slovu a takové slovo je vhodným kandidátem na rejstříkové heslo.

Backgroundový model vychází ze znalosti Zipfových zákonů [14]. Platí rovnice $f.r = k$, která se dá vyložit na seznamu nejčastěji se vyskytujících slov v textu takto: f — 50-té nejčastější slovo vyskytující se v textu bude zastoupeno r — $\sim$ 3-krát častěji než k — 150-té nejčastější slovo. Takové rozložení neodpovídá normálnímu Gaussovu rozložení, ale hyperbolickému Paretovu rozdělení. Podle nich lze určit:

[1] Manuál — návod k použití kde se mohou v textu vyskytovat klíčová slova (vhodní kandidáti do rejstříku) pouze jako nadpisy následované vysvětlením. V takovém případě se klíčová slova mohou v textu objevit jen jednou.

- četnost termu
- dokumentová četnost

Četnost termu závisí na počtu výskytu slova v dokumentu d_j, kde platí vztah

$$f(k, s, N) = \frac{\frac{1}{k^s}}{\sum_{n=1}^{N} \frac{1}{n^s}}$$

kde N je počet elementů, k je jejich hodnost a s je exponent, který popisuje rozložení. Vzorec lze chápat jako zlomek, který popisuje četnost výskytu k-tého nejčastějšího slova. Normalizované rozložení lze zapsat jako:

$$\sum_{n=1}^{N} f(k, s, N) = 1$$

Korpusový model je tedy vytvářen ze seznamu všech slov nezávislého korpusu zbaveného pomocných slov, který se podle četnosti (pořadí) porovná se slovy zpracovávaného textu jako je tomu v tabulce 3.1. V prvním sloupci jsou uvedené dvojice obecného korpusu seřazené podle četnosti, v druhém sloupci jsou dvojice zpracovávaného textu také řazené podle četnosti. Slova, která se nalézají na konci seznamu obecného korpusu, jsou mnohem častější ve zpracovávaném textu. Výjimku tvoří slova jako *tučnice obecná*, která je v obecném korpusu i ve zpracovávaném textu nalezena zřídka, a proto bude vhodné posunout tohoto kandidáta na rejstříkový výraz v seznamu více dopředu. V případě popsaném

obecný korpus	zpracovávaný korpus
déšť měkká voda subtropy	masožravé rostliny subtropy pěstování rostlin
⋮	⋮
plastové nádoby masožravé rostliny tučnice obecná ⋮	tučnice obecná plastové nádoby měkká voda ⋮

Tabulka 3.1: Korpusový (backgroundový) model

tabulkou by slova jako *masožravé rostliny* nebo *tučnice obecná* byly korpusovým modelem vyhodnoceny jako klíčová slova vhodná ke vložení do rejstříku.

Doménový model je založen na oblasti, kterou se dokument (korpus) zaobírá. Např. pokud půjde o knihu biologie — *masožravé rostliny*, pak tento bigram není vhodným kandidátem rejstříkového výrazu, zato kolokacemi budou jednotlivé druhy např. *tučnice obecná* nebo *subtropické rosnatky*. Podle domény se pak vyberou slova, která budou při vybírání kolokací ignorována.

Pro doménový model je třeba získat seznam slov a slovních spojení seřazených podle četností platných pro danou doménu. Takový seznam se dá vytvořit ručně při malém objemu dat. Při větším množství slov by to bylo nepohodlné, např. vše co se týká práva Evropské unie — práva, paragrafy, nařízení č. 1680, evropské fondy, evropská ústava. Tento model byl

použit systémem *Computer-aided Document Indexing System* [6], kde použili pro rozdělení slov do domén tezaurus[2]. Po aplikaci modelu tato slova vypadnou a zůstanou pouze pojmy, kterými se tyto zákony zabývají.

3.3 Existující systémy pro generování rejstříků

Komerční software
Pravděpodobně nejznámější textový editor **MS Word** řeší tvorbu rejstříků tak, že si uživatel musí označit klíčová slova v textu, což je pro češtinu značný problém. Jak bylo řečeno v úvodu, kvůli koncovkám je tvorba rejstříků pro český jazyk tímto způsobem dost náročná. Každé slovo by muselo být označeno ve všech tvarech a v rejstříku by se všechny tyto tvary objevily. Podobný postup využívá i kancelářský balík **OpenOffice**. Tyto programy nenabízejí možnost jednotlivé tvary „sjednotit“ podle lemmy.

Pro firmy, které se zabývají tvorbou rejstříků na profesionální úrovni, jako jsou nakladatelství specializovaná na literaturu faktu, vznikly programy nabízející asistovanou tvorbu rejstříků. Např. software **Sonar Bookends InDex Pro** nabízí výrazy vhodné pro rejstříky na základě výběru vlastních jmen z textu, nejčastějších výrazů, frází apod. Problém asistované tvorby rejstříků spočívá v příliš velkém množství výrazů, které program nabízí. Uživatel si musí vymezit hranice pro optimální množství výrazů, které chce použít pouze na základě zkušeností s tvorbou indexů.
Univerzitní studie
Na Kolumbijské univerzitě v USA se zaměřili na generování rejstříků k internetovým článkům. Potřebovali vyřešit problémy se vznikem stále nových článků, které neměli čas „ručně“ indexovat, a bez rejstříků bylo vyhledávání v takovém množství textů velmi obtížné. Své vyhledávání postavili na dokumentové četnosti víceslovných výrazů tj. nepoužívali korpusový model.

Automatické hledání klíčových slov generovalo velké množství nevyhovujících slovních spojení, které se rozhodli eliminovat projektem **LinkIT** [11]. Postupovali tak, že náhodně vybrali 0,025% termů z korpusu o velikosti 250MB a vyhodnotili tyto termy na základě soudržnosti[3]. Tato přípravná studie ukázala, že 90% náhodně vybraných termů je dobrým kandidátem na klíčové slovo a tedy úspěšný systém potřebuje vyhodnocování s maximálně 10% nepoužitelného výstupu. Tato míra byla vyhodnocena jako postačující, protože se předpokládá, že uživatelé, kteří s texty budou pracovat, dokáží poznat a ignorovat výrazy, které nejsou klíčové a v rejstříku jsou jen proto, že mají vysokou hodnotu dokumentové četnosti. Strojové zpracování nachází více termů než člověk, protože je méně „vybíravé“. Samotný projekt funguje tak, že se nejdřív označkují slova slovními druhy. Vytvoří se termy[4] typu **NP** *noun phrase* — jmenná fráze založené na řídícím podstatném jméně, protože většina klíčových slov ho obsahuje. V takovém termu má jedno slovo řídící postavení např. *kávový filtr*, *olejový filtr*, *uhlíkový filtr* – vedoucí slovo je zde *filtr*. Složitější je to s termy, které mají více vedoucích slov jako *druh rakoviny - způsobený azbestem* — tam jsou vedoucí slova *rakovina* a *azbest*. Po vyhledání termů se vytvoří databáze pojmů, kde se k řídícím slovům ukládají zbytky termů. Takto uložené termy se vyhodnotí a označkují na základě frekvence. Základní pojmy si nadefinovali takto:

[2]Tezaurus — nabízí synonyma, tezaurus použitý zmiňovaným projektem měl slova rozdělené do kategorií — domén, podle toho do jaké oblasti slovo spadá, např. mléko — potravina.

[3]Soudržnost termů — společný výskyt slov těsně vedle sebe, jak často se vyskytovaly v tzv. kolokacích s dírami atd.

[4]Term je slovo nebo skupina slov, která se používá v specifickém kontextu

- klíčová slova jsou identifikována četností slov v dokumentu.
- technické termíny jsou NP nebo části NP opakované víc než dvakrát v dokumentu.
- řídící NP jsou identifikovány metodou, ve které jsou termy seřazeny podle vedoucího slova. Termy jsou oznámkovány a dál seřazeny podle vzestupně podle četnosti.

Touto metodou získali pouze 6,5% nepoužitelných výrazů, proto je tento způsob zpracování prakticky využitelný pro jejich typ hledání klíčových výrazů v textu.

Většina systémů generujících rejstříky je pouze *počítačem podporovaná* jako projekt **Computer-aided Document Indexing System** [6] univerzity v Záhřebu. Stejně jako čeština má i chorvatština různé morfologické tvary především u podstatných a přídavných jmen. Problematika různých tvarů je vyřešena generováním dvou nabídek. Jedna pro všechny tvary slova s uvedeným počtem jejich výskytů a druhá zobrazuje pouze lemmata a jejich množství. Program vyhledává i n-gramy a to o dvou až čtyřech slovech. Jako vstupní formát je použit jazyk XML, na kterém zakládají vyhledávání slov a jejich ukládání do seznamu. Hledání klíčových slov je založeno na tezauru podle kterého se určují klíčová slova. Další řazení se řídí frekvencí výrazů v dokumentu. Pro projekt byl použit tezaurus EUROVOC, který je vícejazyčný a zahrnuje 6000 tříd rozdělených do 21 oblastí — politika, věda, finance aj. seřazených hierarchicky do osmi tříd. Byl navržen pro Evropské společenství, takže program dokáže indexovat pouze témata blízká problematice tezauru. Kromě závislosti indexování tématu může být problémem pomalé vyhledávání (časová složitost může být až exponenciální).

Univerzita v Severním Texasu v USA se rozhodla vytvořit vlastní testovací sadu automatického generování rejstříků a to na základě **zlatého standardu** [3], který závisí na parametrech jako jsou délka rejstříku, délka vstupních položek, rozsah pokrytých témat. Víceslovné výrazy si rozdělili na:

- n-gramy — všechny generované n-gramy z dokumentu obvykle pokryjí n-gramy v rejstříku.
- NP — výrazy ukládané podle řídících slov
- syntaktické výrazy — fráze založené na znalostech lingvistiky.

Hledání klíčových výrazů založili na několika metodách a zkoušeli, jak a zda-li vůbec se vyhledávání výrazů zlepší. Testovali četnost, délku výrazů v rejstříku a gramatické vzorce. Materiály nasbírali na stránce projektu Gutenberg[5]. Hlavním problémem, ostatně jako vždy při zpracování rejstříků, jsou nedostatečná vstupní data. Jen malá část knih byla vložena s rejstříkem anebo rejstřík vůbec neobsahovaly.

Rejstřík vytvářeli v různě jemných strukturách, protože testovali i to, jak jemná struktura půjde vytvořit. Vkládali nejdřív řídící slovo výrazu a zbytek výrazu vložili podle gramatických pravidel. Utvořili si skupiny slov, na která aplikovaly pravidla např. výrazy s předložkou, předložka následuje po slovese např. takto: *Acetate, of Ammonium Solution.*

Zlatý standard vytvářeli pro různě podrobné rejstříky. Pro každý text vytvořili dva rejstříky a to: jednoduchý index založený na řídícím slově a dlouhý rejstřík založený na plně rekonstruovaných položkách rejstříku s různou úrovní granularity[6]. Vytvoření některých termínů rejstříků může být dost složité, a proto pro nejjemněji strukturovaný rejstřík bylo

[5] http://www.gutenberg.com — je zde možné zdarma stáhnout knihy v různých jazycích.
[6] Granularita — víceúrovňová nebo-li jemná struktura rejstříku.

použito vyhledávání gramatických výrazů přes web — AltaVista [12], na kterém ověřili, zda výraz existuje. Webem ověřené výrazy ponechali v seznamu, zbytek zahodily a úspěšnost hledání vzrostla z 30,34% na 54,78%. Důležitým parametrem vyhodnocení je délka textu vzhledem k délce rejstříku. Pro vyhodnocení vzali poměr počtu slovních jednotek v dokumentu, vzhledem k položkám v rejstříku. Jednoduchý rejstřík obsahuje asi 0,44% slovních jednotek, které odpovídají zhruba jedné frází v rejstříku pro každých 227 slov v textu. Jemné rejstříky mají poměr 0,7%, což odpovídá frázi na každých 140 slov.

Další práci podobnou automatickému generování rejstříků je organizování a linkování spřízněných webových stránek [9]. Nástroj je založen na novém typu hypertextu:
HC — hypertextová konkordance je hypertextový rejstřík, který řadí pojmy podle kontextu stejně jako konkordance. Konkordance sloužily k rychlému vyhledávání podobných a souvisejících pasáží v rozsáhlém textu. HC je charakterizována těmito vlastnostmi:

- termíny k indexování jsou vybrány terminologickým extrakčním algoritmem.
- výskyty indexovaných termínů v dokumentu jsou provázány odkazem do rejstříku
- termíny jsou uváděny ve stylu konkordance
- každý termín v indexu je provázán se svým dokumentem

Program by měl být schopen indexovat i dokumenty, které nebyly napsány ve značkovacím jazyce jako jsou HTML, XML a jiné SGML jazyky. Termíny jsou automaticky extrahovány Damerauovou metodou [2].

$$score(word) = \frac{\frac{f(word,coll1)}{f(coll1)}}{\frac{f(word,coll2)}{f(coll2)}}$$

Metoda porovnává relativní frekvence termínu v dokumentu *coll1* s relativní frekvencí termínu v referenční kolekci *coll2*[7].

Výhodou této metody jsou:

- jednoduchost algoritmu, který provádí srovnání s obecným korpusem a tím vyhodnocuje, které termíny jsou zajímavé
- algoritmus se dá aplikovat i na krátké texty

Systém funguje tak, že uživatel zadá webové stránky, které chce indexovat. Pro stránky se vyhledají termíny vzhledem k obecnému korpusu a vyberou se všechny unigramy a bigramy s vyjímkou těch, které jsou ve stop-listu. Nejlépe ohodnocené termíny jsou indexovány. Množství termínů odpovídá délce dokumentu a pravděpodobnostnímu ohodnocení. Konkrétní úspěšnost není zmíněna, pouze uvádějí, že systém může dosahovat horších výsledků než běžné zpracování, ale urychluje to práci, takže se nepřesnosti vyplatí.

[7] Ve zlomcích výrazu jsou udány maximální pravděpodobnosti výskytu slova v textu a kolekci.

Kapitola 4

Určování relevantních slovních spojení

Nejjednodušší metoda jak získat slovní spojení z textu je vytvořit všechny možné dvojice sousedících slov. Tato kapitola uvádí další možnosti, jakými lze získat bigramy, i když spolu jednotlivá slova bigramu přímo nesousedí, také se zabývá často používanými statistickými metodami pro získání relevantních slovních spojení.

4.1 Směrodatná odchylka

Kolokace se dobře hledají na základě četnosti v textu. V případě kolokací s dírami to nemusí stačit. Pro takové případy lze zjistit zda se jedná o kolokaci ze směrodatné odchylky. Pro „volný vstup“ podkapitoly 2.2 je směrodatná odchylka spočtena takto:
Nejdřív se určí průměrná vzdálenost slov ve větách:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{3}(2+1+2) = \frac{5}{3}$$

Směrodatná odchylka se určí ze vztahu:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(d_i - \mu)^2}{n-1}$$

kde n je kolikrát se slova vyskytla společně, d_i je vzdálenost vzájemného výskytu vzhledem k pozici i, μ je průměrná vzdálenost výskytu. Vyčíslení pro vzorové věty pak vypadá takto:

$$\sigma = \sqrt[2]{\frac{1}{2}((2-\frac{5}{3})^2 + (1-\frac{5}{3})^2 + (2-\frac{5}{3})^2)} \approx 0.57$$

Pokud je hodnota směrodatné odchylky blízká nule jako v tomto případě, tak se pravděpodobně jedná o kolokaci. V případě, že by směrodatná odchylka byla rovna nule, pak by se slova kolokace vyskytovala pouze spolu (vedle sebe).

I přes vysokou četnost a nízkou standardní odchylku se slova spolu mohou vyskytovat pouze náhodou. Aby se možnost náhody zcela vyloučila, provádí se testování nulovou hypotézou. Nejprve se nadefinuje problém jako nulová hypotéza H_0 a k ní se ustanoví inverzní hypotéza H_1.

- H_0 — slova v bigramu se vyskytují společně pouhou náhodou
- H_1 — slova v bigramu jsou kolokací

Pro slova vyskytující se spolu náhodně platí:

$$P(s^1 s^2) = P(s^1)P(s^2)$$

kde s je slovo a p celková pravděpodobnost, která je dána pravděpodobnostmi výskytu jednotlivých slov v textu. Pokud je pravděpodobnost jevu H_0 velmi malá, tj. když se pravděpodobnost p pohybuje v intervalu 0 – 0,0005, je možné hypotézu zavrhnout a tím potvrdit H_1.

4.2 T - test

Hodnoty jednotlivých pravděpodobností se určí t-testem.

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

kde $\bar{x}$ je střední hodnota vzorku, s^2 je odchylka vzorku, N je velikost vzorku a μ je střední hodnota celé množiny, ze které se vzorek vybírá. Jestliže je t dost velké, pak můžeme nulovou hypotézu zavrhnout. K hodnotám t se vyhledají stupně významnosti v statistických tabulkách. Jestliže t je větší než vyhledaný stupeň, pak můžeme nulovou hypotézu zavrhnout s pravděpodobností závislou na rozdílu t a stupně významnosti. Pro výpočet t - testu kolokace je třeba nejdřív určit průměr a odchylku ze vzorku. Jako vzorek se vezme sekvence N bigramů a jednotlivé části se označkují 1 nebo 0 podle toho zda se jedná o část testovanou jako kolokace nebo nikoli.

$$P(\textit{tučnice}) = \frac{4675}{14307668}$$

$$P(\textit{obecná}) = \frac{15828}{14307668}$$

Jak už víme nulová hypotéza pro tento případ je, že slova jsou nezávislá.

$$H_0 : P(\textit{tučnice obecná}) = P(\textit{tučnice})P(\textit{obecná})$$

$$P(\textit{tučnice obecná}) = \frac{4675}{14307668}\frac{15828}{14307668} \approx 3.615.10^{-7}$$

Pokud je nulová hypotéza pravdivá, pak bigram *nová rosnatka* v náhodně generovaných bigramech dostane přiřazenou 1 a ostatní bigramy 0 s pravděpodobností $p = 3,615.10^{-7}$. Průměr pro toto rozložení je μ a odchylka je $\sigma^2 = p(1-p) \approx p$ což je zhruba p. Vyčíslení σ^2 vrací pro většinu bigramů malé hodnoty p. Vypadá to, že bigram *tučnice obecná* se ve vzorku vyskytl 8 - krát vzhledem k celkovému množství bigramů. Pak tedy průměr:

$$\bar{x} = 8\textit{*ostatní bigramy} \approx 5.591.10^{-7}$$

a celkové vyčíslení t-testu:

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} \approx \frac{5.5910^{-7} - 3.610^{-7}}{\sqrt{\frac{5.5910^{-7}}{1430000}}} \approx 0.10065$$

Hodnota t je menší než udávaný kritický stupeň pro $\alpha = 0.005$, takže nulovou hypotézu nelze zavrhnout a bigram je hledanou kolokací, což je správně. T-test ovšem neřeší problém bigramů jako jsou *v roce* apod. Vyhodnocuje je jako kolokace, a proto je stále zapotřebí stop-list. Tato statistika se v přirozeném zpracování řeči dá použít i pro jiné problémy jako např. jakým způsobem určit, zda je bigram relevantním slovním spojením.

4.3 Míra vzájemné informace

Velmi často se k vyjádření vztahu mezi dvěma proměnnými x' a y' používá míra informace. Pro zpracování výskytu slov v kolokaci lze použít vyjádření:

$$I(x', y') = log_2 \frac{P(x', y')}{P(x')P(y')} = log_2 \frac{P(x', y')}{P(x')} = log_2 \frac{P(y', x')}{P(y')}$$

Obvykle bývá míra vzájemné informace vyjádřením vztahu mezi náhodnými proměnnými a ne vazbou mezi jejich hodnotami. Pro případy kolokace míra mezi hodnotami udává, o kolik se zvýší hodnota pravděpodobnosti, že se na pozici $i \pm 1$ bude vyskytovat proměnná y', když pozice proměnné x' je i. Pokud je bigram kolokací dochází k extrémnímu případu, platí:

$$I(x', y') = log_2 \frac{P(x', y')}{P(x')P(y')} = log_2 \frac{P(x')}{P(x')P(y')} = log_2 \frac{1}{P(y')}$$

Takže čím je četnost kolokace nižší, tím vyšší hodnotu bude mít míra vzájemné informace, tj. výrazy s malou četností jsou preferovány před četnými. Z toho plyne, že to není příliš dobrý popis jevu, protože četná kolokace bude méně ohodnocena než neobvyklá kolokace. Je to přesný opak toho, jak je třeba mít jev ohodnocen. Pro opačnou mezní situaci, kdy jsou slova naprosto nezávislá platí:

$$I(x', y') = log_2 \frac{P(x', y')}{P(x')P(y')} = log_2 \frac{P(x')P(y')}{P(x')P(y')} = log_2 1 = 0$$

Míra vzájemné informace je vhodná spíše pro vyloučení kolokací. Pro míru vzájemné informace lze použít úpravy, které vedou k charakterističtějším popisu jevu. Pro vyšší váhy četnějších kolokací stačí zavést mocniny četnosti n až do stupně deset. Na korpusech se experimentálně ověří, která mocnina je pro korpus nejlépe použitelná.

$$I(x', y') = log_2 \frac{P(x', y')}{P(x')P(y')} = log_2 \frac{C(x'y')^n N}{C(x')C(y')}$$

kde C je kubická míra a n řád mocniny. Po takové úpravě lze mluvit o míře vzájemné informace vyšších řádů.

4.4 Pearsonův X^2 test

Další možnou metodou hledání relevantních slovních spojení je X^2 test. Je vhodnější než t-test, protože ten se zakládá na normálním pravděpodobnostním rozložení, což neodpovídá tak zcela povaze textových korpusů. Tento test závislostí slov v bigramech se nezakládá na normálním rozložení. Zjednodušeně řečeno je pro X^2 test vytvořena tabulka 4.1. Základ testu spočívá v porovnání frekvencí bigramů v tabulkách s frekvencí očekávanou pro nezávislá slova bigramu. Pokud je rozdíl mezi těmito frekvencemi příliš velký, pak je možné

zavrhnout nulovou hypotézu. X^2 je asymptoticky rozložená χ^2, takže čím větší čísla, tím větší šance, že X^2 má rozložení χ^2.

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

kde i udává počet řádků tabulky, j je počet sloupců, O_{ij} je právě zkoumaná buňka tabulky a E_{ij} je očekávaná hodnota. Proč nám toto rozložení vyhovuje? Očekávaná hodnota E_{ij} je

	$s_1 = nový$	$s_1 \neq nový$
$s_2 = typ$	114	101
	nový typ	*tento typ*
$s_2 \neq typ$	1011	249
	nový systém	*tento systém*

Tabulka 4.1: Výskyty kolokací pro X^2 test

určena z okrajových pravděpodobností, tedy z proporciolizovaných součtů řádků a sloupců.

$$\frac{114 + 101}{N} \frac{114 + 1011}{N} N \approx 971.325$$

kde N je počtem všech bigramů a ostatní hodnoty jsou doplněny z tabulky 4.1. Očekávaná hodnota první buňky je určena okrajovou pravděpodobností *nový* a druhou částí bigramu *typ*. Tato hodnota bude platit pro čistě náhodný výskyt těchto dvou slov v bigramu společně. Hodnota χ^2 se získá takto:

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})} \approx 12.3$$

kde N je množství všech bigramů, O_{ij} jsou počty výrazů pro kombinaci slov z tabulky 4.1. Hodnota potvrzuje, že se slova v bigramu nacházejí spolu čistě náhodně, neboť výsledek je výrazně nad stupněm významnosti pro $\alpha = 0.005$. Výsledek je stejný jako u t - testu, jejich výsledky se o mnoho neliší. Důvodem proč se více používá X^2 test je to, že X^2 lze uplatnit i tam, kde je operováno s vysokými mírami pravděpodobnosti, kde by běžný t-test selhal[1].

4.5 Pravděpodobnostní podíly

Pravděpodobnostní podíl umožňuje zjistit snáz než X^2 zda se jedná o kolokaci, protože její výstup je v uspořádaném pořadí a není nutné hledat v statistických tabulkách mezní hodnoty, o které budou data oříznuta. Ze dvou navržených hypotéz je hned zjevné, která je pravděpodobnější. Hypotézy pro rozptýlená data bigramu $s_1 s_2$:

- $H_1 : P(s^2|s^1) = p = P(s^2|\neg s^1)$
- $H_2 : P(s^2|s^1) = p_1 \neq p_2 = P(s^2|\neg s^1)$

[1] X^2 test se používá např. pro překlady z cizích jazyků na základě stejné frekvence výskytu slov v korpusech.

První hypotéza popisuje nezávislá, druhá závislá slova v bigramu.

$$p = \frac{c_2}{N} = \frac{22}{2450}$$

$$p_1 = \frac{c_{12}}{c_1} = \frac{10}{20}$$

$$p_2 = \frac{c_2 - c_{12}}{N - c_1} = \frac{22 - 10}{2450 - 22}$$

kde se používají maximální hodnoty pravděpodobností p, p_1, p_2 a pro slova v korpusu s_1 a s_2 a jejich bigram $s_1 s_2$ se zapisují hodnoty c_1 má 20 výskytů, c_2 jich má 22, c_{12} jich má 10. Vyhodnocují se binomickým rozložením:

$$b(k, n, x) = \binom{n}{k} x^k (1 - x)^{(n-k)}$$

Nyní se před připravené hypotézy konečně podělí:

$$log\lambda = log\frac{L(H_1)}{L(H_2)}$$

$$log\lambda = log\frac{b(c_{12}, c_1, p)b(c_2 - c_{12}, N - c_1, p)}{b(c_{12}, c_1, p_1)b(c_2 - c_{12}, N - c1, p_2)}$$

$log\lambda = logL(c_{12}, c_1, p) + logL(c_2 - c_{12}, N - c_1, p) - logL(c_{12}, c_1, p_1) - logL(c_2 - c_{12}, N - c_1, p_2)$

kde $L(k, n, x) = x^k(1-x)^{n-k}$. Síla podílových pravděpodobností spočívá v snadno čitelném výsledku a přesnějším zpracování řídkých (rozptýlených) dat.

4.6 Metody založené na gramatice

Kromě výše uvedených statistických metod lze použít zpracování založené na znalosti lingvistiky. Pro hledání kolokací je možné použít gramatické vzorce, jak to bylo provedeno v [5]. Vyhledávání je založené na korpusu, z kterého se automaticky vypíší všechny víceslovné pojmy odpovídající gramaticky a kolokačně. Kostry jsou hledány na základě gramatických a kolokačních vlastností. Původně existovaly kostry pouze pro angličtinu, práce [5] je univerzální pro všechny jazyky. K zadanému korpusu se přidají gramatické vzorce platné pro jazyk, v jakém byl vytvořen korpus, a systém dokáže vytvořit seznam víceslovných výrazů a synonym. Pro víceslovné pojmy dokáže vygenerovat seznam rozdílů v použití blízkých pojmů.

Kromě frekvenčního vyhledávání jsou v kostrách použity gramatické vzorce. Spíš než frekvenční hledání klíčového slova jsou vyhledávány všechny relace, ve kterých se slovo vyskytuje. Slova jsou předzpracována značkovačem, který jim určí slovní druh, a na výstupu lemmatizována nebo převedena do správného slovního tvaru pokud nejdou lemmatizovat. Víceslovné pojmy jsou pak vygenerovány do jednotlivých seznamů podle příbuznosti použitých slov. Gramatické vzorce mohou vypadat třeba takto:

$$1 : V(DET|NUM|ADJ|ADV|N)^*2 : N$$

kde 1 a 2 jsou slova určená slovními druhy, ke kterým se hledají všechny možné relace a obsah závorek jsou všechny možné slovní druhy, které se mezi slovy 1 a 2 mohou vyskytovat[2].

[2]Závorky ohraničují výraz, který může být 0 – n - krát zopakován, opakování značí *, — je značka pro nebo.

Kapitola 5

Testovací data a vytvořený systém

5.1 Testovací data

Pro testování systému bylo použito volně dostupných knih v elektronické podobě. Získání vstupních dat bylo značným problémem, přestože je knih v elektronické podobě dost, existuje jich jen málo s kvalitním rejstříkem. Běžná literatura rejstříky neobsahuje vůbec, proto bylo hledáno mezi populárně - naučnou literaturou a učebnicemi, a nakonec bylo použito většinou učebnic, které mají obsáhlejší rejstříky. Některé byly ve formátu e-book, takže se velmi obtížně převáděly do textové podoby, u dalších byl problém s českými znaky nebo s dost často se vyskytujícími latinskými symboly jako např. ϱ. Prvním problémem tedy bylo překonvertovat data do formátu, který bude dále snadno zpracovatelný. Většina materiálů byla k dispozici ve formátu PDF[1], pro který existuje sice mnoho programů pro převod na obyčejný text, ale ne všechny se hodí pro češtinu[2].

Jako nejlepší řešení se nakonec ukázalo použití *OCR*[3] programů, které ovšem převádí text s chybami jako, např. špatné načtení některých speciálních symbolů např. $\bar{a}$, které je zapsáno jako dvě písmena, která jsou později chybně rozpoznána jako zkratka. Takto zpracovaný text byl uložen ve formátu *txt*. Z knih byly odděleny rejstříky do samostatných souborů pro další zpracování, ve kterém byly rejstříky rozděleny podle délek n-gramů. Téměř každá kniha měla jiný tvar rejstříku a doporučení uvedená v normě nebyla příliš dodržována. Pro přehlednost je doporučováno psát rejstřík s odrážkami viz. tabulka 2.1, ale pro automatické vyhodnocování výsledků to přínosem nebylo. Takto vytvořené rejstříky se „rozpadly“ na unigramy.

Občas se v rejstřících vyskytla synonyma nebo anglické ekvivalenty za klíčovým slovem rejstříku. Takové položky byly rozděleny na dva samostatné výrazy a hodnoceny samostatně, to se opět projevilo negativně na hodnocení, protože synonyma se obvykle v knize vůbec nevyskytla a nemohla být nalezena. Např. pro *jmenný server* se často používá anglický název *name server*, který je v této knize uveden jen v rejstříku. Na každé straně e-book se opakovaly výrazy, které sloužily pro orientaci knihou např. *rozcestník* nebo *obsah*. V knihách jsou často hlavičky stránek se jménem kapitol např. Lieova grupa, což

[1] PDF — Portable Document Format

[2] Firma Adobe, která přišla se standartem PDF, na svých stránkách poskytuje službu vygenerování prostého textu z PDF formátu, ale přestože jim bylo zasláno několik různých souborů, žádné nebyly zpracovány.

[3] Optical Character Recognition — optické rozpoznávání znaků je metoda, která umožňuje digitalizaci tištěných textů nebo textů ve formátech určených k tisku jako je PDF. Převedený text je závislý na kvalitě předlohy, protože OCR program nerozeznává všechna písmena správně.

nebývá problém ve vyhodnocení, protože to je klíčové slovo. Horší je pokud je v hlavičce např. jméno autora. Odstranit takové výrazy z celého dokumentu bylo zavrhnuto, protože takový zásah by mohl ovlivnit i text knihy, která by pak mohla být ochuzena o některá klíčová slova. Např. automatickým odstraněním slov z e-book jako *rozcestník* by byly tyto slovní spojení odstraněny i z textu, což by mohlo vést k odstranění klíčového slova. Při zkoumání výsledků byly tyto n-gramy ponechány v textu a ignorovány s tím, že uživatel je dokáže poznat a odstranit.

Kromě textů knih byl použit obecný korpus z podkapitoly 2.2, který využívá text obsahující deset miliónů slov. Obecné korpusy obsahují různé texty, které by měly vytvořit dostatečně velkou množinu slov pro statistiku četnosti slov v textu.

5.2 Vytvořený systém

Systém je založen na použití frekvenčních metod a určování slovních druhů souborem programů PDT popsaným v kapitole 2.3. Z programů PDT je pro tento systém používán parser, jenž dokáže text převádět na CSTS[4] validní formát. CSTS je formát založený na SGML[5] a byl hlavním formátem dat ve verzi PDT 1.0. Ačkoliv byl v PDT 2.0 nahrazen PML[6], některé nástroje jej stále výhradně používají. CSTS může reprezentovat jen morfologickou a analytickou anotaci, kdežto PML je formát dat založený na XML, navržený pro reprezentaci lingvistické anotace textů jako jsou morfologické značkování, závislostní stromy apod. Pro tento systém úplně stačí CSTS formát.

PDT očekává vstupní textové soubory ve formátu *iso-8859-2*, které převádí do jazyka CSTS pomocí parseru. Parser neznačkuje pouze slova a hranice vět, ale dokáže označit také odstavce a nadpisy. Takto upravený text projde nejprve morfologický analyzator, určí všechny možné značky, o které by se mohlo jednat, a z nich značkovač vybere tu nejpravděpodobnější. Značkovač lze spustit se dvěma různými parametry:

- T — Tagger, jenž značkuje neznámá slova pomocí X
- TG — Tagger - Guesser, který neznámá slova *uhodne*, ale i přesto mohou být některá slova označena X jako neznámá.

V systému byla použita data generovaná s parametrem *TG*, pokusy s oběma parametry neprokázaly velké rozdíly mezi testovacími daty. Původní předpoklad, že slova označená jako neznámá budou téměř všechna klíčová, se nepotvrdil, jak lze vidět v tabulce 5.1.

Většinu neznámých slov tvoří zkratky jako jsou značky chemických prvků a označení množin ($\sim 65\%$)[7]. Další slova bez rozpoznaných slovních druhů jsou použita ve zvláštním tvaru např. *elektro chemický*, kde není snadné „uhodnout“ slovní druh, anebo se může jednat o vlastní jména, která nejsou ve slovníku ($\sim 35\%$). Značnou skupinu tvoří cizí (hlavně anglická) slova např. *machine learning*, u kterých je velká pravděpodobnost, že se jedná o klíčová slova, a měla by se v rejstříku vyskytnout. Proto jsou slova s neznámým slovním druhem v dalším zpracování zvýhodňována.

Značkovač přiřazuje slovům 16-ti znakovou značku, která je definována v [4]. V systému jsou využívány především značky na první, druhé a třetí pozici. První pozice uvádí slovní

[4] CSTS — Czech sentence tree structure

[5] Standard Generalized Markup Language — rodina jazyků používaná pro značkování textů.

[6] PML — Prague Markup Language

[7] Výskyt tak velkého množství chemických prvků a označení množin je způsoben použitými testovacími daty.

tvar v textu	základní tvar - lemma	značka	příklad použití
Tm	thulium	Xx----------	vzácný chemický prvek Tm
Cf	californium	Xx---------	
Xe	Xe	Xx---------	
Ck	Ck	Xx---------	množina Ck
NP	Np	Xx---------	NP úplný problém
Anthony	Anthony	XX---------	jménem Anthony
learning	learning	XX---------	používá se machine learning
tělese	těleso	XX---------	
pH	Ph	Xx---------	neutrální Ph
elektro	elektro	XX---------	elektro chemická

Tabulka 5.1: Příklady slov s neurčeným slovním druhem

druh viz. A.1, druhá udává podrobnější informace o slovním druhu a třetí jmenný rod např. *Laplaceův rozvoj* je označen *AUIS1M--------- NNIS1-----A----*, kde *AU* znamená přídavné jméno přivlastňovací, *I* mužský rod, *S* číslo jednotné, *1* – pád první, *M* přivlastňovací rod mužský a zbylé pozice nejsou určené, u druhého slova značka *NNIS1* znamená podstatné jméno obyčejné, mužský rod v jednotném čísle, prvním pádě, *A* udává afirmativ[8].

Systému je dán vstupní text, který je programy PDT 2.3 parsován a označkován. Dále je zpracováván skripty, které pracují se vstupním textem i korpusy a vytváří z nich seznam klíčových slov, který se v jednotlivých experimentech liší podle nastavených parametrů. O podrobnostech jednotlivých experimentů bude pojednávat kapitola 6.

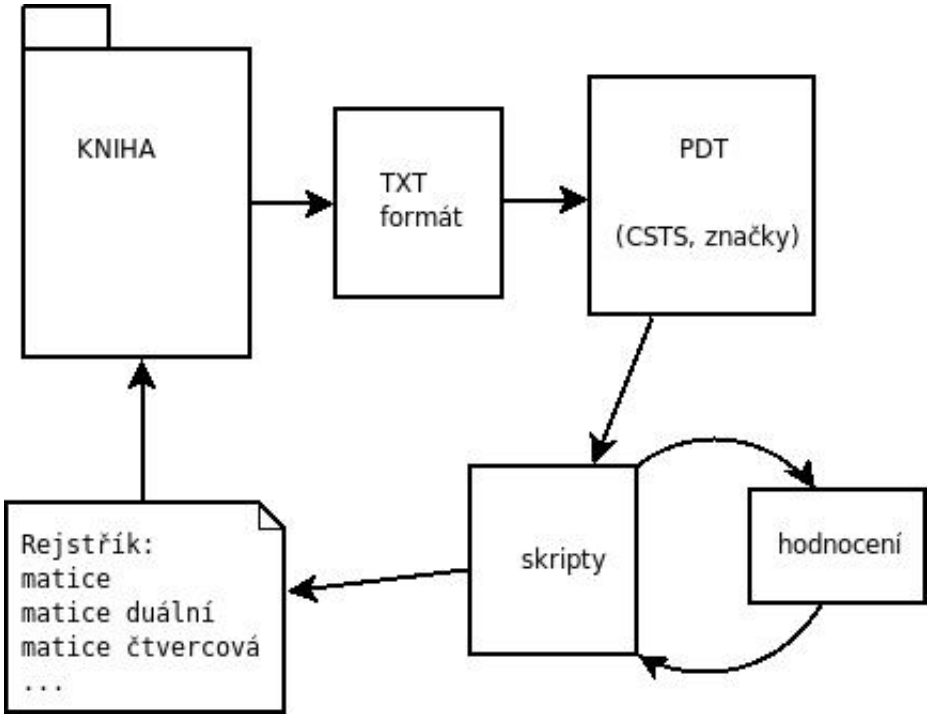
počet	lemma	dvojice v textu	slovní druh 1	slovní druh 2
104	matice a	matice a	NNFS5-----A----	J,----------
104	a e	a e	J,----------	NNNXX-----A----
86	e v	e v	NNNXX-----A----	RR--6----------
73	o o	o o	RR--6----------	RR--6----------
65	v a	v a	RR--6----------	J,----------
63	v e	v e	RR--6----------	NNNXX-----A----
58	pro všechny	pro všechen	RR--4----------	PLYP4----------
58	e k	e k	NNNXX-----A----	RR--6----------
57	a je	a být	J,----------	VB-S---3P-AA---
56	u e	u e	RR--2----------	NNNXX-----A----

Tabulka 5.2: Označkované bigramy vytvořené pouze na základě četnosti

V první fázi pokusů s texty byly utvořeny bigramy a trigramy pouze na základě frekvence (četnosti). Jak lze vidět v tabulce 5.2 bylo zde příliš mnoho pomocných slov. Ty se dají odstranit pomocí stop-listů[9], ale protože bylo třeba ponechat na výstupu především podstatná a přídavná jména, bylo v dalších krocích použito rovnou filtrace slov podle slovních druhů. Tabulka 5.2 obsahuje označkovaná slova, která potvrzují předpoklad, že nejčastějšími výrazy jsou podstatná jména a pomocná slova.

[8]Afirmativ — slovo bez negativní předpony „ne-“

[9]Stop-list — seznam výrazů, které budou vynechány v dalším zpracování.



Obrázek 5.1: Diagram systému

5.3 Zpracování n-gramů

Označkované texty jsou nyní připraveny pro vytváření testovacích n-gramů. Většinou se v rejstřících vyskytují především pojmy o délce jednoho až tří slov, proto se experimenty zabývají těmito n-gramy. U delších n-gramů je už obtížné nastavit pravidla, která určí zda se jedná o klíčové slovo nebo náhodně se spolu vyskytující slova.

Během jednotlivých experimentů byly nalezeny „chyby“ značkovače, které byly řešeny dalšími skripty používanými pro generování n-gramů. Značkovač vracel některá častá slova „špatně“ označená – *a, ale, aby, či* byla označena jako podstatná jména. Uvedená slova mohou být podstatnými jmény, ale přece jen nejčastěji se tato slova vyskytují v textech jako spojky, a proto byly pro další zpracování přeznačeny na *J,-----------*. Podobná situace nastala pro *s, k, v, o*, které byly rozpoznány jako zkratky podstatných jmen, a proto byly přeznačeny na mnohem obvyklejší tvar — předložky *RR--6----------*.

Problémy způsobovala také lemmatizace 2 slov, která negované výrazy převáděla na výrazy bez předpony „ne-“, např. *nerovnost* byla převedena na *rovnost*. Podobně tomu bylo u stupňovaných přídavných jmen, např. *nejvyšší doména* má lemmu *vysoká doména*, ale pro potřeby rejstříku nebylo žádoucí používat lemmatizované tvary v těchto případech. Pro tato slova byl použit místo lemmy tvar, který se vyskytoval v textu, např. *nejvyšší doména*. Další úpravy se přímo dotýkaly provedených experimentů nebo délky n-gramů, a proto budou podrobněji popsány v kapitole 6.

Lemmatizace n-gramů byla nutná ze dvou důvodů:

- přehlednější nabídka n-gramů v základním tvaru
- sjednocení n-gramů v různých tvarech pod jedno klíčové slovo

N-gramy se v textu vyskytují v různých tvarech, jak je uvedeno v tabulce 5.3, a tak by se v rejstříku vyskytly několikrát. Takto se všechny tvary „uložily“ pod jednu lemmu, která byla ohodnocená součtem výskytů jednotlivých tvarů. Některé n-gramy mohou být převedeny na lemmu chybně, proto se původně zamýšlelo použít jako základní tvar v rejstříku nejčastější tvar n-gramu v textu nebo ten, co bude v prvním pádě. V textu se mnoho n-gramů v prvním pádě nevyskytovalo takže nejčastější tvar n-gramu také nebyl vhodný. Podle experimentů bylo nejjednodušší provést úpravy na již lemmatizovaných n-gramech.

jednotlivé tvary	značka 1	značka 2	počet výskytů
zvolené bázi	AAFS6----1A----	NNFS6-----A----	8
zvolená báze	AAFS1----1A----	NNFS1-----A----	1
zvolené báze	AAFP4----1A----	NNFP4-----A----	1
zvolené bázi	AAFS2----1A----	NNFS4-----A----	1
zvolených bázích	AAFP6----1A----	NNFP6-----A----	1
zvolené bázi	AAFS3----1A----	NNFS3-----A----	1
zvolenou bází	AAFS7----1A----	NNFS7-----A----	1
báze zvolená		se celkem vyskytla	14

Tabulka 5.3: Uložení tvarů podle lemmy

Jednotlivá slova jsou systémem vyhledány pouze na základě četnosti. V rejstříku by měly být jako **unigramy** uvedena pouze podstatná jména v prvním pádě. Pro další omezení velkého množství výrazů bylo experimentováno s použitím obecných korpusů. Unigramy nejsou tak častá klíčová slova jako bigramy, ale používají se v jemněji strukturovaných rejstřících pro větší přehlednost např.

- děj
- děj — adiabatický
- děj — izotermický a izobarický
- děj — termodynamický

Automatickým generováním není lehké dosáhnout tak jemné struktury, a proto jsou v rejstříku ve stejném množství jako bigramy, aby rejstřík poněkud rozčlenily.

Bigramy vytvořené systémem byly do rejstříku uloženy ve tvaru podstatné jméno následované přídavným tak, aby bylo možné co nejlépe dodržet normu. Problémy nastaly u dvojice podstatných jmen, kde je těžké rozhodnout, které ze jmen je řídící, pokud ani jedno není v prvním pádě. Proto zůstaly takové dvojice v pořadí v jakém byly načteny ze souboru, pouze se kontrolovalo, zda se už někde nevyskytly v opačném pořadí, v takovém případě se v seznamu výrazů uvádí častější varianta.

Z bigramů byly odstraněny veškeré číslovky, částice, spojky, zájmena, předložky, citoslovce, příslovce, slovesa a interpunkce. Samozřejmě jednodušší by bylo pouze povolit výskyt podstatných a přídavných jmen, ovšem pro testy bylo užitečné si prohlédnout jak se výstup změní s odstraněním každého dalšího druhu, a zda-li by se to nějak nedalo využít pro delší n-gramy. Odstranění příslovcí je sporné, protože některé rejstříky je používají, přestože se v literatuře [1] doporučuje převádět takové výrazy na přídavná jména. Např. *lineárně nezávislá* by měla správně být v rejstříku uváděna jako *matice - lineárně nezávislá*. Zrovna

v tomto případě je těžké vyhnout se používání příslovcí, proto také norma jejich používání pouze *nedoporučuje*. V rejstřících se občas mohou vyskytnout i číslovky, např. *Karel IV.*, jejichž problematika je zkoumána experimenty v podkapitole 6.6. Z dalšího zpracování byly vynechány, protože jejich používání přineslo více bigramů, které se v rejstříku nedaly použít.

Na výstupu zůstaly pouze dvojice složené z neurčených slov a podstatných a přídavných jmen, ze kterých byly vyloučeny dvojice složené pouze z přídavných jmen. Tyto dvojice se mohly hodit jako klíčová slova, ale těžko by se hledalo řídící podstatné jméno, pokud s ním již neutvořily trigram. Ne všechny takové dvojice byly odstraněny, a to kvůli špatnému určení některých slovních druhů např. *nezáporná samoadjungovaný*
AAFP1----1N---- NNFS2-----A----. Takový případ byl pozorován pouze jednou, takže se tento problém dá při hodnocení zanedbat. Další filtr byl zaveden pro jednopísmenné

slovní druh 1	slovní druh 2	původní koncovka	nová koncovka
-	AAF	(ý|é)	á
NNF	A	(ý|é)	á
NNF	A	(ův|ových)	ova
NNN	AUF	ův	ovo
NNN	A	ý	é
-	AAI	á	ý
-	AAI	(á|é)	ý

Tabulka 5.4: Úpravy koncovek bigramů

dvojice tj. *X Y NNFP2-----A---- NNFP2-----A----*, které byly také zcela odstraněny, protože se převážně jedná o části vzorců nebo příkladů. Vzorce v textech způsobily mnoho problémů, např. rovnice

$$A.(B.C) = \sum(a_{i,j})$$

se zdeformuje na *(A ˇ (B ˇ C) = (Jednotkový prvek (O3) a11.* Největší problémy způsobují vzorce obsahující znaky jako suma $\sum$ nebo matice, které se rozloží na více řádků a jsou narušeny okolním textem.

Přestože byly bigramy uvedeny v základním tvaru, tak potřebovaly ještě upravit koncovky podle pravidel uvedených v tabulce 5.4. Typickým příkladem může být *zobrazení tečný*, které se převede na *zobrazení tečné* podle pravidla pro *NNN A* v tabulce 5.4. V některých případech nestačilo změnit koncovku jen podle kategorií přídavného jména, ale měnila se podle rodu řídícího podstatného jména, především to platí pro dvojice podstatné jméno a přídavné jméno přivlastňovací. Tyto změny se vždy nesetkaly s úspěchy, protože přivlastňovací přídavná jména jsou často chybně vyhodnocena jako podstatná jména např. *Cauchyova věta* má správný tvar lemmy *Cauchyho věta*. U většiny takových dvojic byla lemma zapsána ve správném tvaru, hlavní problémy u podobných případů způsobila neznalost správného základního tvaru vlastního jména např. *Jordanova tvaru* je lemmatizováno na *Jordanov tvar* namísto *Jordanův tvar*.

Pro **trigramy** byla experimentálně zjištěna poněkud jiná pravidla než pro bigramy. V rejstřících se občas vyskytují trigramy, které obsahují zájmeno, předložku či spojku, ale norma [1] jejich používání nedoporučuje. Původně bylo zamýšleno vybudovat filtry, které by povolily určité slovní druhy jen na některých pozicích, ale protože filtry nepřinesly očekávané zlepšení, byly takové trojice úplně odstraněny. Po provedení několika experimentů s použitím filtrů vzniklo velké množství kombinací využívajících ukazovací zájmena nebo

předložky ve tvarech jako, např. *toto lineární zobrazení* nebo *informace o základech*. Mezi použitelné trigramy se nakonec zařadily pouze podstatná a přídavná jména, která se nesmí vyskytovat v žádném z uvedených pořadí:

- přídavné jméno — podstatné jméno — přídavné jméno
- jakákoli dvojice — přídavné jméno
- trojice, kde se na dvou libovolných pozicích nachází jednopísmenné slovo např. *A matice B*

Trigramy byly pro tisk uspořádány ve tvaru podstatné jméno z konce trigramu jako první, následované zbylými slovy. Toto pravidlo bylo změněno pouze v případě, že se jednalo o zkratku, např. *lineární zobrazení f* se zapíše jako *zobrazení lineární f*. Toto uspořádání nemusí být ideální pro všechna data, ale pro testovaná data se osvědčilo při tvorbě strukturovaného rejstříku.

Typy rejstříků Pro hodnocení úspěšnosti je důležité vědět, jaké typy rejstříků měla

kniha	obyčejný	strukturovaný	strukturovaný jemně	seznam	komentář
A		√		√	seznam nadpisů
B	√				chybně vytvořený
C	√				podrobný, téměř bez chyb
D			√		rozpad struktury na unigramy
E	√				podrobný, téměř bez chyb
F	√				malé množství výrazů, mnoho chyb

Tabulka 5.5: Typy rejstříků

zpracovaná data. V příloze jsou uvedeny konkrétní knihy, které byly použity jako testovací data a v tabulce 5.5 jsou uvedeny typy rejstříků, které jednotlivé knihy obsahují. Obyčejný rejstřík není strukturovaný, výrazy jsou zapsaný tak, jak se vyskytují v textu. Strukturovaný rejstřík má jednu úroveň, jemně strukturovaný je členěn do dvou a více úrovní. Seznam nadpisů nebo kapitol se používá hlavně u manuálů a může být doplněn o některá častěji používaná slova. Příliš krátké nebo špatně zpracované rejstříky měla většina testovaných knih, a proto se experimentování řídilo především podle knih *C* a *E*.

Kapitola 6

Výsledky experimentů

6.1 Metodologie vyhodnocování

Experimenty byly prováděny na knihách s různě kvalitními rejstříky, což se negativně projevilo na vyhodnocení výsledků. Ideální vyhodnocení by bylo provedeno několika uživateli, kteří by podle svého subjektivního dojmu určili, jak moc jim nabídka klíčových slov ulehčila práci s tvorbou rejstříku, a jak ji zrychlila.

Rejstříky byly pro vyhodnocování rozděleny na rejstřík unigramů, bigramů a trigramů. Každá část rejstříku se vyhodnocovala zvlášť, protože hledání n-gramů nemá stejnou úspěšnost např. pro bigramy a trigramy. V mnoha případech se vyskytl některý unigram jako součást bigramu nebo trigramu. Obvykle se uvádí unigram a za ním delší výrazy se stejným řídícím podstatným jménem. U bigramů a trigramů je to sporné, protože např. *teorie neuronových sítí* a *neuronové sítě* jsou obě položkami rejstříku, ale např. *horní trojúhelníková matice* položkou je a *trojúhelníková matice* není. V takovém případě je třeba, buď uvést oba výrazy jako v experimentu 6.2, anebo se řídit množstvím výrazů v textu jako v experimentu 6.2. Uváděný příklad lze snadno rozhodnout ve prospěch trigramu, který se opakuje 11krát a bigram 12krát. Z tohoto opakování je zřejmé, že trigram se v textu nevyskytl tolikrát pouze náhodně a bigram je jeho součástí. Jiné případy nemusí být tak jednoznačné.

Pro jednotlivé experimenty byly upravovány rozsahy výstupních rejstříků s cílem dosáhnout co nejmenšího počtu výrazů a nejvyšší úspěšnosti. Podrobné změny jsou uváděny u jednotlivých experimentů. Data byla zpočátku vyhodnocována pouze na rovnost dvou n-gramů, ale protože se mnoho výrazů lišilo pouze v koncovce, bylo jako úspěch hodnoceno i nalezení „podobného“ n-gramu s odlišnou koncovkou, např. *adjungovaná zobrazení* = *adjungované zobrazení*. Bigramům byla odebírána koncovka o délce jednoho znaku, trigramům koncovka až o délce tři znaky pro případy jako, např. *model neuronových sítí* = *model neuronové sítě*.

K hodnocení se tedy využívá procentuální shoda rejstříku knihy s vygenerovaným rejstříkem. Takové hodnocení je ovšem vhodné pouze jako známka toho, zda další experiment vedl k lepším výsledkům. Popsané vyhledávání není dokonalé, protože rejstříky dodané v knihách bývají neúplné, používají nedoporučovaná slova a mohou obsahovat i jiné např. gramatické chyby. Pro uživatele je spíš podstatné množství výrazů, které bude muset smazat jako nehodící se. Protože většina rejstříků obsahuje *mnohem* méně výrazů, než by měla, je možné, že uživatel by z nabídky zadal více klíčových slov než kdyby dělal rejstřík ručně. Pro další hodnocení jsou uváděny procentuální shody s originálním rejstříkem kromě části 6.5, která je zaměřena na vyzkoušení tvorby rejstříku tímto systémem.

6.2 Vyhodnocení provedené na základě četnosti výrazů

Experiment č. 1

Pro první experiment byla použita všechna data, vygenerovaly se všechny bigramy a trigramy, které se porovnaly oproti rejstříkům s úspěšností uvedenou v tabulce 6.1. Ze zpra-

název knihy	nalezeno bigramů z rejstříku	celkem nabídnuto dvojic	úspěšnost hledání v %	nalezeno trojic z rejstříku	celkem nabídnuto trigramů	úspěšnost hledání v %
A	6/156	2472	3.85	5/39	3219	12.82
B	137/170	17598	81.18	22/29	25108	75.86
C	45/48	3493	93.75	13/23	5252	56.52
D	88/110	12776	80.00	24/29	17728	82.76
E	125/151	5916	82.78	30/40	10452	75.00
F	28/33	4766	84.85	19/25	7033	76.00

Tabulka 6.1: Experiment 1 — četnost nad všemi výrazy

cování byly vynechány n-gramy obsahující interpunkci, pomocná slova, a bigramy tvořené jednopísmenými výrazy, např. *X Y*. Tímto experimentem byla zjištěna maximální možná úspěšnost hledání v rejstříku a základní druhy chyb, které znemožňují nalezení výrazů uvedených v rejstříku. Např. kniha *B* nemůže mít úspěšnost hledaných bigramů větší než 74%, důvody které zabránily nalezení všech výrazů uvedených v rejstříku jsou popsány v podkapitole 6.2 stejně jako data, která tvoří těch zbylých 26%.

V další části experimentu byly odstraněny výrazy, které se v textu vyskytly pouze jednou. Že bude třeba se zaměřit na detekci klíčových slov s nízkou četností plyne z tabulky 6.2 vzhledem k tomu, že bigramů bylo nalezeno zhruba o 30% méně a trigramů až o 50% méně než v předchozí tabulce 6.1. V rámci prvního experimentu bylo zkoumáno použití číslovek,

název knihy	nalezeno bigramů z rejstříku	celkem nabídnuto dvojic	úspěšnost hledání v %	nalezeno trojic z rejstříku	celkem nabídnuto trigramů	úspěšnost hledání v %
A	2/156	298	1.28	0/39	180	0.00
B	126/170	3211	74.12	15/29	2326	51.72
C	31/48	549	64.58	7/23	355	30.43
D	59/110	2456	53.64	14/29	1861	48.28
E	82/151	1048	54.30	8/41	1001	19.51
F	19/33	694	57.58	5/25	439	20.00

Tabulka 6.2: Experiment 1 — čestnost nad výrazy kromě těch, co se vyskytly jednou

protože v některých případech mohou být klíčovými slovy, např. *Karel IV.*, ale většinu číslovek v experimentálních datech tvořila čísla stran, obrázků anebo části výpočtů takže výstup byl zahlcen výrazy jako *168 gramů*. O odstranění číslovek ze seznamu rozhodlo množství výrazů obsahujících číslovky, např. kniha *D*, kde existuje celkem *42478* bigramů, z toho *5638* bigramů obsahuje číslovkou, ale jen *sedm* číslovek se vyskytuje v rejstříku. Tento nepříznivý poměr dvojic rozhodl pro odstranění veškerých číslovek.

Experiment č. 2

Značná část n-gramů se v základním tvaru lišila pouze koncovkou, a proto se další experiment zabýval především úpravou koncovek popsaných tabulkou 5.4 tak, aby byl

název knihy	nalezeno bigramů z rejstříku	celkem nabídnuto dvojic	úspěšnost hledání v %	nalezeno trojic z rejstříku	celkem nabídnuto trigramů	úspěšnost hledání v %
A	7/156	8518	4.49	5/39	4174	12.82
B	138/170	12949	81.18	20/29	6940	68.97
C	45/48	7835	93.75	13/23	4075	56.52
D	85/110	8716	77.27	8/29	4898	27.59
E	131/151	7704	86.75	28/41	4057	68.29
F	28/33	9692	84.85	18/25	5235	72.00

Tabulka 6.3: Experiment 2 — upravené výrazy

zredukován počet nabízených výrazů jinou metodou než „oříznutím“ výrazů podle četnosti výskytu. Do zpracování byly opět zařazeny n-gramy vyskytující se v textu pouze jednou. Porovnání tabulek 6.1 a 6.3 ukazuje, že došlo ke snížení počtu nabízených výrazů o čtvrtinu až polovinu. V některých případech došlo i k snížení úspěšnosti vyhledávání, ale to bylo způsobeno pouze těmi výrazy, které se v rejstříku nevyskytovaly v lemmatickém tvaru[1].

Experiment č. 3 — korpusový model

Vzhledem k množství výrazů, které jsou nabízeny, je stále potřeba snížit jejich počet, a proto byl pro další experiment použit *korpusový model* popsán kapitolou 3.2. Z korpusu byly vygenerovány n-gramy, které byly aplikovány na n-gramy z předchozího experimentu. Bigramům a trigramům byly experimentálně stanoveny meze, podle kterých jsou výrazy zvýhodňovány

kniha	nalezeno bigramů z rejstříku	celkem nabídnuto dvojic	úspěšnost hledání v %	nalezeno trigramů z rejstříku	celkem nabídnuto trigramů	úspěšnost hledání v %
A	7/156	8437	4.49	5/39	4174	12.82
B	135/170	12727	79.41	20/29	6939	68.97
C	45/48	7776	93.75	13/23	4075	56.52
D	84/110	8644	76.36	8/29	4898	27.59
E	131/151	7642	86.75	28/41	4057	68.29
F	28/33	9609	84.85	8/29	4898	27.59

Tabulka 6.4: Experiment 3 — backgroundový (korpusový) model

na základě experimentů s modelem. Nabízené n-gramy byly přesunuty na konec seznamu, pokud se v obecném korpusu vyskytovaly příliš často tzn. pokud se nalezly *10krát* a více v korpusu. Původně byly tyto výrazy zahozeny, ale ukázalo se, že to může vést k odstranění některých klíčových slov z nabídky jak uvádí tabulka 6.4. Pokud se výrazy v obecném korpusu vyskytly více než *dvakrát* a méně jak *devětkrát*, bylo jejich ohodnocení navýšeno tak,

[1] Jedná se o chybu způsobenou dodaným rejstříkem.

aby se posunuly na začátek seznamu výrazů. Mezi výrazy vyskytujícími se v textu pouze jednou byly časté slovní obraty jako např. *způsob definice*, ale také se mezi nimi vyskytují klíčová slova např. *určitý integrál*. Konkrétně tyto příklady nebylo možné korpusovým modelem rozlišit.

Experiment č. 4 — unigramy

Výběr unigramů je zcela postaven na korpusovém modelu, protože na první pozice nabízených unigramů se dostanou ta nejběžněji používaná slova, která určitě nejsou klíčová. Jako nejpravděpodobnější klíčové slovo se jeví unigramy, které se v korpusu vyskytly

kniha	nalezeno unigramů z rejstříku	počet výrazů	%
A	23/68	3405	33.82
B	77/88	4487	87.50
C	14/15	2648	93.33
D	267/461	2644	57.92
E	34/48	2380	70.83
F	2/3	3006	66.67

Tabulka 6.5: Experiment 4 — unigramy, korpusový model

1 - 10krát[2] nebo v korpusu nebyly nalezeny. Unigramy, které se vyskytly v korpusu *11 - 1350krát*, jsou vyhodnoceny jako méně pravděpodobná klíčová slova. Uvedené skupiny jsou upřednostněny v seznamu klíčových slov před ostatními. Problém se slovy které v korpusu vůbec nebyly nalezeny nebo se vyskytly jen jednou spočívá v tom, že mezi klíčová slova jsou zařazeny i slova s gramatickou chybou nebo nevhodní kandidáti do rejstříku, např. *další příklad*.

Rozbor chyb

Rozdělení chyb na druhy bylo prováděno na knize *E*, protože měla nejlépe zpracovaný rejstřík a nejvyšší úspěšnost. Při rozboru chyb nad jinými knihami by bylo potřeba projít mnohem více textu při hledání chybějících výrazů. Rozbor chyb na ostatních knihách, s vyjímkou *C* a *E*, by znamenalo procházet chyby rejstříků.

Unigramy

Unigramy z tabulky 6.6 nebyly v rejstříku knihy *E* nalezeny. Tato tabulka slouží jako ilustrace problémů spojených s hledáním unigramů. Všechny unigramy by měly být podstatná jména, a právě porušení tohoto pravidla způsobuje nejvíc chyb (chyba typu 2), protože v originálním rejstříku knihy je mnoho klíčových slov, které tuto podmínku nesplňují. Problémem zůstávají stále slova, která se v textu vůbec nevyskytla, a jsou zjevně synonymem nějakého pojmu vyskytujícího se v textu (chyba typu 3). Poslední druh chyby jsou slova, která obsahují gramatickou chybu nebo nějaký neobvyklý znak (chyba typu 1), zde je to ä, které se chybně přepsalo z původního $\bar{a}$ a ani takový tvar by nemusel být nalezen. Např. zápis takového znaku v LATEXu se dává do složených závorek, které by se odfiltrovaly jako interpunkce a výraz by se vůbec nezpracoval.

[2] Meze byly zvoleny experimentálně.

unigram	chyba
ä-matice	1
ekvivalentní	2
homogenní	2
indefinitní	2
kolmé	2
nilpotentní	2
různoběžné	2
regulární	2
rovnoběžné	2
rozložitelné	2
rozpadlé	2
sesquilineární	2
uúhel	3
uúsečka	3

Tabulka 6.6: Seznam nenalezených unigramů

Bigramy

V tabulce 6.7 jsou popsány druhy chyb, kvůli kterým nebyly nalezeny některé pojmy z rejstříku v knize *E*. Nejčastější chybou při hledání bigramů jsou výrazy, které se nevyskytují v textu (chyba typu 1). Tento problém lze rozdělit na dvě části:

- výrazy, které se v textu vůbec nenalézají, může jít o synonyma nebo názvy, které popisují probíraný jev.
- výrazy, které se vyskytly jako tzv. kolokace s dírami zmiňované podkapitolou 2.2. Např. z výrazu *permutace se nazývá sudá* vznikl výraz *sudá permutace*.

Převážná většina (9/11) výrazů se v textu vůbec nevyskytly, o ostatních výrazech se dá mluvit jako o *kolokaci s dírou*, ale pravděpodobně by nebyly nalezeny ani při použití metod detekujících takové kolokace, jak bylo popsáno v kapitole 2. Např. výraz *ortogonální doplněk podmnožiny* je klíčové slovo samo o sobě a těžko z toho určit, že klíčovým slovem má být pouze *ortogonální podmnožina*, se *svazkem přímek* je podobný problém, protože v textu se vyskytuje pouze jednou jako *přímkou p (tzv. svazek rovin)*, a z takových výrazů by měl i člověk problém nalézt uvedená klíčová slova, pokud by neznal problematiku. Posledním problematickým pojmem byla *uúplná kontrakce*, která se v textu nalézá pouze jako *úplná kontrakce*.

Ve třech případech (chyba typu 2) byly „chybně“ určeny slovní druhy, a tak byla některá slova vyloučena z nabídky. Např. *hermiteovskými* bylo určeno jako příslovce, které byly z nabídky vyřazeny. Problém se správným určením slovního druhu se vyskytuje především u přivlastňovacích přídavných jmen. Určení příslovce místo podstatného nebo přídavného jména není chybou, protože tato slova jako příslovce existují, problém je spíše v tom, že zrovna zde mají funkci podstatného nebo přídavného jména a zrovna v těchto případech vyvolá příslovce chybu.

Šestkrát nebyly výrazy nalezeny, protože rejstřík nebyl vytvořen podle normy (chyba typu 3) a používal příslovce.

výraz v rejstříku	druh chyby
antisymetrický tensor	1
charakteristická čísla	1
hermiteovské zobrazení	2
lineárně nezávislá	3
lineárně závislá	3
neřešitelná soustava	1
nedourčená soustava	1
negativně definitní	3
negativně semidefinitní	3
nehomogenní soustava	1
oddělující nadrovina	1
ortogonálně diagonalizovatelné	3
ortogonální podmnožiny	1
positivně definitní	3
pseudoinverzní matice	2
sudá permutace	1
svazek přímek	1
uúplnou kontrakci	1
určená soustava	1

Tabulka 6.7: Chybějící výrazy v rejstříku — bigramy

Trigramy

Najít trigramy je mnohem obtížnější, protože není jisté, v jakém tvaru budou slova uspořádána a značná část trigramů se v textu nevyskytuje přímo tak, jak jsou zapsány v rejstříku. Je tedy možné, že nalezených klíčových výrazů je ve skutečnosti víc, než udává tabulka 6.3. V další tabulce 6.8 jsou probrány chyby, ke kterým dochází u trigramů. I zde byly největším problémem výrazy, které se v textu vůbec nevyskytly (chyba typu 1) — 12 výrazů. Jednou se zde vyskytl špatně určený slovní druh (chyba typu 3), kde *adjungovaná* byla určena jako příslovce. Jako chybně vytvořený výraz se dá chápat i *inverze v permutaci*, protože používání předložek není doporučeno normou [1].

6.3 Vyhodnocení na základě X^2

V kapitole 4.4 byl zmíněn Pearsonův test X^2, kterým lze rozhodnout, jestli se slova v bigramu vyskytují pouze náhodně. Pro zpracování textů se používá míra rizika $\alpha = 0,05$ s prvním stupněm volnosti, který odpovídá hodnotě *3,48*. Všechny výrazy, které jsou ohodnoceny méně, jsou hledanými výrazy. Prvních patnáct nalezených bigramů je uvedeno v tabulce 6.9. Tato metoda není pro hledání kandidátů do rejstříku vhodná, protože mezi nalezenými výrazy nebyly ty nejčastější, které se podařilo nalézt seřazením výrazů podle četnosti. Pokud se jedno slovo z dvojice vyskytovalo často a druhé téměř vůbec je pravděpodobné, že bude mezi slovy vybranými touto metodou. Takže ve výsledku byly vynechány výrazy, které se vyskytují velmi často jako např. *lineární soustava* v knize *D — Lineární algebra* a naopak vybrány výrazy, které se vyskytly jen jednou jako např. *matice*

výraz v rejstříku	chyby typu
algebraicky adjungovaná matice	3
fundamentální soustava řešení	1
inverze v permutaci	1
kladné samoadjungované zobrazení	1
matice soustavy rovnic	1
metrické klasifikace forem	1
nezáporné samoadjungované zobrazení	1
obecná poloha bodů	1
singulární čísla matice	1
střed dvojice bodů	1
submaticí matice a	1
vzájemně kolmé projektory	1

Tabulka 6.8: Chybějící výrazy v rejstříku — trigramy

složená, kde se *matice* vyskytuje v textu velmi často, ale *složená* pouze třikrát. Takové dvojice by ze statistického hlediska měly být klíčové, ale přesto se téměř žádná z nalezených nevyskytla v rejstříku knihy. Většina rejstříkových výrazů byla vybrána z nejčastěji se vyskytujících výrazů nebo z těch, co se v textu vyskytly pouze jednou. Úspěšnost Pearsonova testu se pohybovala řádově v procentech, a protože další statistické metody jsou založeny na podobných principech bylo dále pracováno pouze s metodami založenými na četnosti, které přinášely lepší výsledky.

6.4 Odstranění redundantních výrazů

V podkapitole 2.3 byl zmiňován problém redundance kratších n-gramů v delších. Jejich odstraněním by mělo dojít k nabídce menšího počtu výrazů. Unigramy není třeba odstraňovat, protože většinou bývají v rejstřících ve větším množství kvůli přehlednosti a vytvoření strukturovaného rejstříku podle vzoru tabulky 5.3. Sledováním klíčových slov v trigramech a bigramech bylo rozhodnuto, že výrazy, které se vyskytovaly téměř stejně často jako dvojice i jako trojice budou pravděpodobně nadbytečné a budou ponechány pouze trigramy, ze kterých si uživatel případné přebývající slovo může odmazat, což je snazší než si domýšlet chybějící slovo bigramu.

Odstraněním redundantních výrazů zbude podle tabulky 6.10 stále ještě velké množství výrazů pro ruční výběr. Toto množství lze zredukovat tak, aby počet výrazů v nabídce odpovídal normě [1]. Norma uvádí počet stran rejstříku 5 – 15% z počtu stran knihy. Pro další experimenty byla navržena horní mez rejstříku 15% stran, protože se nepředpokládá, že by všechny nabídnuté výrazy byly vhodnými kandidáty. Pro výsledný výstup systému byly použity výrazy, které byly redukovány korpusovým backgroundem a redukcí bigramů na trigramy. Na výstupu je uvedeno množství výrazů, které byly nabídnuty a procentuální úspěnost vzhledem k rejstříku v testované knize. Opět je třeba mít na paměti, že ke zpracování nebyla použita žádná kniha, u které by rejstřík odpovídal normě. Procentuální úspěšnost je pouze orientačně ukazuje jestli nebylo ztraceno velké množství dříve úspěšně detekovaných výrazů. Procentuální úspěšnost experimentu č.3 byla zachována

hodnota	výraz
3.457	prvek vzhledem
3.431	generátor vektorový
3.428	souřadnice stejná
3.428	souřadnice reálná
3.421	rozklad následující
3.421	popis adsorpční
3.420	napětí vztah
3.412	matice složená
3.412	matice operace
3.371	mezi reagující
3.368	vodivost roztok
3.354	při řešení
3.346	obsah lineární
3.329	množina daná
3.323	báze výsledná

Tabulka 6.9: Prvních 15 bigramů nalezených pomocí X^2 v datech E

kniha	počet nabízených výrazů	počet výrazů po redukci
A	16100	15923
B	24379	23908
C	14561	14378
D	16100	15923
E	14143	13954
F	17936	17654

Tabulka 6.10: Korpus — odstranění bigramů, které jsou součástí trigramů

pouze s vyjímkou knihy B, kde se dva bigramy přesunuly mezi trigramy.

Pro snížení počtu výrazů podle normy [1] uvažujme rejstřík, kde na stránce jsou dva sloupce textu, tedy na jednu stranu vychází zhruba *110* výrazů. Poměr unigramů, bigramů a trigramů není normou definován, protože závisí na typu textu, a tak byla z každé skupiny n-gramů vybrána třetina z celkového množství výrazů. Pro automatický výběr lze navrhnout oddělení nejméně frekventovaných výrazů z nabídky a podle spokojenosti uživatele vybrat požadovaný rozsah. Předpokládá se, že v rejstříku se nevyskytnou výrazy, které se v textu vyskytly pouze jednou, protože pokud se takové výrazy pomocí doménového modelu nepřesunuly v seznamu na začátek, budou zahozeny. V tabulkách B.1, B.2 a 6.11 jsou zachyceny procentuální úspěšnosti jednotlivých rozsahů pro každý n-gram zvlášť. Protože nejúspěšnější byla nejvyšší mez, je použita pro vytvoření systému.

6.5 Případová studie

Případová studie byla provedena na textech k předmětu *Číslicové zpracování řeči*. Text obsahoval značné množství anglických slov, které způsobily problémy popisované v dalším odstavci. Kromě těchto slov jsou zde i LATEXové značky, které by měly být z textu také

odstraněny. Vzhledem k tomu, že jich není tolik, bylo ponecháno na uživateli jejich odstranění. Text má rozsah 215kB, (130 stran textu), kvůli problémům s velkými a malými písmeny, byl převeden na malé znaky. Vytvoření rejstříku trvalo půl hodiny na nabídnutém rozsahu rejstříku 5% stran knihy, tzn. nejmenší nabízený rozsah *1669* výrazů. Ručně vybraní kandidáti na rejstříkové výrazy mají *441* výrazů. Mnohem delší nabídka sice nabízí více výrazů, ale její zpracování trvá hodinu. Záleží tedy na uživateli, jak velkou nabídku chce projít. V příloze B.3 je uveden příklad rejstříku vytvořeného z nabídky.

Rozbor chyb

Největší skupinu chyb tvoří gramatické chyby (151). Protože zde bylo velké množství zkratek, které slovník nerozeznal nebylo možné se bez důkladné znalosti textu rozhodnout co je zkratka, a co je část vzorce. Jako gramatické chyby jsou hodnoceny výrazy typu: *zprac, zakladnich, vzokr*. Další velkou skupinu tvořily anglické výrazy (83), které sice mohou být klíčovým slovem v rejstříku, ale kvůli různým tvarům, v kterých se v textu vyskytly, jsou v nabídce víckrát, např. *filter, filters*. Dále se tam vyskytují i slova jako *we*, kterým s českým slovníkem nelze správně přiřadit slovní druhy a odfiltrovat je. Především vlastním jménům jsou přiřazovány špatné lemmy (39), případně jsou zde slova dvakrát, protože jim pokaždé byla vytvořena jiná koncovka v základním tvaru. Zlomek chyb tvoří výrazy jako jsou zpodstatnělá přídavná jména (31), např. *vysoká*, může jít o zvěř nebo o školu (podstatná jména), anebo o přídavné jméno a ta se do rejstříku jako unigramy nehodí.

Hlavní problém při tvorbě rejstříku bylo rozlišit zkratky od gramatických chyb. Pro správný výběr klíčových výrazů by bylo třeba znát podrobně text. Problémy také působí tvary trigramů, protože zatím nebyla nalezena pravidla, která by jim dokázala přiřadit správné koncovky. Např. *pulsní kódová modulace* je ve správném tvaru, ale pokud šlo o tvary, kde se vyskytují dvě podstatná jména, dochází k problémům, např. *perioda základní tón*. Výrazy byly nabízeny ve tvaru, který měl dát základ víceúrovňovému rejstříku.

6.6 Zhodnocení nalezených výsledků

Nejkvalitněji zpracovaný rejstřík měly knihy *C* a *E*, což vysvětluje překvapivě dobré výsledky ve srovnání s ostatními knihami. Typy rejstříků uvedené v tabulce 5.3 značně ovlivnily vyhodnocování. Kniha *A* má rejstřík vytvořený především z nadpisů, jedná se o manuál, pro kterou by bylo vhodnější vytvořit seznam nadpisů a doplnit je nejfrekventovanějšími slovy. Kniha *D* má rejstřík vytvořený formou jemné víceúrovňové struktury, který se ovšem vlivem automatického zpracování rozpadl na unigramy. Rejstřík knihy *F* obsahuje malé množství výrazů a nebyl vytvořen podle doporučení uvedených v normě [1]. Rejstříky jsou rozdílné, takže lze pozorovat chování systému a chyby na různých typech.

Nejlépe hodnocený experiment byl založené na četnosti bez vynechání výrazů, ten byl užit jako míra maximální úspěšnosti, kterou již nelze zvýšit. Nenalezené n-gramy zde byly většinou zastoupeny výrazy, které se v textu vůbec nevyskytly. Takovou chybu nelze odstranit automatickým zpracováním, takže nezbývá než, aby si uživatel tyto výrazy sám doplnil. Po vynechání výrazů vyskytujících se v textu pouze jednou, úspěšnost výrazně poklesla, takže byly provedeny další pokusy, které měly detekovat málo frekventované výrazy. Jedním z pokusů byla snaha využít statistických metod např. Pearsonův test. Pomocí tohoto testu byly nalezeny výrazy s úspěšností v rozmezí 3–9%. Tak nízkých hodnot bylo dosaženo proto, že většina statistický metod vybírá jako klíčová slova ta, která nejsou v textu frekventována příliš často a při hledání kandidátů vhodných do rejstříku jsou použity především nejvíce

frekventované výrazy, a pak výrazy vyskytující se velmi málo, třeba jen jednou v celém textu. Po neúspěchu tohoto experimentu byly zavrženy další metody jako je t-test, protože jsou založeny na stejném principu.

Dále bylo experimentováno s backgroundovým modely. Korpusovým modelem byly posunuty výrazy tak, aby se málo frekventované dostaly na začátek seznamu, pokud zde byla šance, že by to byl dobrý kandidát na rejstříkový výraz. V podstatě všechny unigramy, byly zpracovány korpusovým modelem, protože obsahovaly nejvíce výrazů, které by nebylo možné jinak vytřídit. Srovnání úspěšnosti korpusového modelu můžeme provést s experimentem 1 z tabulky 6.2, kde byly odstraněny výrazy vyskytující se pouze jednou. Úspěšnost systému nelze měřit procentuální úspěšností hodnocení, ale množstvím nabízených výrazů a tím kolik klíčových slov bylo ztraceno oříznutím seznamu.

kniha	unigram nabízené výrazy	unigram nalezené výrazy	unigram %	bigram nalezené výrazy	bigram %	trigram nabízené výrazy	trigram nalezené výrazy	trigram %
A	506	10/68	14.71	4/156	2.56	366	0/39	0.00
B	3080	75/88	85.23	123/170	72.35	2310	18/29	62.07
C	440	8/15	53.33	40/48	83.33	366	9/23	39.13
D	2768	267/461	57.92	64/110	58.18	2126	6/29	20.69
E	1027	34/48	70.83	103/151	68.21	770	17/41	41.46
F	660	0/3	0.00	21/33	63.64	476	8/25	32.00

Tabulka 6.11: Výsledný systém — nabízené výrazy a úspěšnost hledání

Experimentováním byla stanovena hranice, kdy dochází k minimálním ztrátám klíčových slov a rozsah jejich nabídky je překročen pouze o 5% nad maximální doporučovanou délku rejstříku. Unigramy nedosáhly vysoké úspěšnosti, ale aspoň došlo ke snížení počtu nabízených výrazů. V některých případech zůstalo stejné množství výrazů jako před oříznutím, protože množství výrazů odpovídalo pětině stran z počtu stran knihy. Ztráta výrazů se při pohledu na výsledky nezdála velká. Unigramy a bigramy potřebovaly nastavit 20% stran z počtu stran knihy, což přesahuje doporučení, pro trigramy stačila horní mez 15%. Bigramy dosáhly o 10% vyšší úspěšnosti, než tomu bylo v *experimentu 1B* při nižším nebo téměř stejném počtu výrazů. Oříznutím v *experimentu 1B* a v tomto posledním experimentu se množství výrazů rovnalo, ale použití korpusového modelu vedlo jednoznačně k lepším výsledkům. Trigramy dosáhly dokonce vyšší úspěšnosti s větším oříznutím, což bylo zapříčiněno tím, že jich bylo vytvořeno méně už na počátku zpracování[3]. Vzhledem k tomu, že množství výrazů bylo sníženo oproti *experimentu 1B* a bylo dosaženo lepších výsledků, je použití korpusového modelu užitečné.

Přestože byly rejstříkové výrazy nalezeny, zůstává zde ještě problém koncovek výrazů. Mezi unigramy se vyskytovaly zpodstatnělá přídavná jména, ty pak měly celý tvar slova chybně. Větší problémy se vyskytly u bigramů, kde byly sice koncovky přeznačkovány, ale pokud byl chybně určen rod, tak je koncovka chybná, např. *budova nový*. Koncovky trigramů nebyly nijak filtrovány, protože není lehké zjistit v jakém pořadí mají být slova trojice uspořádány a podle jakého slova se budou řídit jejich koncovky. Úprava pořadí slov a koncovek trigramů je ponechána na uživateli.

[3]Trigramy byly mnohem přísněji odfiltrovány od nepřípustných slovních kombinací.

Kapitola 7

Závěr

Tato práce měla prozkoumat možnosti hledání klíčových výrazů v textu pro využití při automatické tvorbě rejstříků knih. Hlavní otázkou bylo, s jakou úspěšností lze detekovat klíčová slova, a jaké metody jsou nejúspěšnější. Také bylo zkoumáno jaké množství výrazů bude nabídnuto, a jestli navržený systém usnadní práci uživatelům.

Před zpracováním vlastním systémem byly zkoumány metody zúžení nabídky výrazů a zavrhnutí výrazů, které se podle normy [1] v rejstříku nemají vyskytovat. Na základě výzkumu bylo rozhodnuto používat program, který dokáže určit slovní druhy a nabídnout výrazy v základním tvaru.

Nejúspěšnější metoda hledání klíčových slov je vytvoření seznamu výrazů podle četnosti, ale protože je takto nabídnuto velké množství výrazů, bylo experimentováno s metodami, které měly zredukovat jejich množství. Použití statistických metod nepřineslo očekávané zlepšení, a proto byl výsledný systém doplněn pouze backgroundovým korpusovým modelem. Ten měl zvýhodnit klíčové výrazy, které se v textu vyskytovaly pouze zřídka. Ale protože model měl za následek pouze přeskládání pořadí výrazů a jejich počet nezmenšil, bylo třeba množství výrazů snížit na počet doporučovaný normou. Některé klíčové výrazy tím byly ztraceny, ale zato zbylo menší množství výrazů, které je už přijatelné pro uživatelské zpracování. Cílem bylo získat takový poměr počtu výrazů a vhodných kandidátů pro rejstřík, aby výsledný systém uživatele nezahltil nadbytečným množstvím výrazů, které nebude schopen zpracovat.

Dále bylo vyvráceno, že by délka knihy ovlivňovala úspěšnost hledaných klíčových slov. Pro testování byly použité různě dlouhé knihy, a bylo zjištěno, že nezávisí ani tak na délce knihy, ale na tom, jakým stylem je napsaná. Pokud se jedná o návod nebo manuál není systém moc úspěšný a dokáže najít pouze pár nejfrekventovanějších slov, která můžeme označit jako klíčová.

Pro další zlepšení systému by bylo potřeba získat texty s kvalitně vyrobenými rejstříky, aby bylo možné lépe otestovat změny v úspěšnosti. Největší problém způsobily výrazy, které se v textu vůbec nevyskytly, a přesto byly uvedeny jako klíčová slova v rejstříku. Další výzkum v této oblasti by se tedy měl zabývat hlavně hledáním málo frekventovaných klíčových slov v textu. Velkým problémem byly také koncovky výrazů, především u trigramů, a právě uspořádáním slov a koncovkami trigramu nebo delších n-gramů by se mohli zabývat další práce.

Dodatek A

Seznam používaných zkratek z projektu PDT

značka	latinský název	český název
A	adjektivum	přídavné jméno
C	numerál	číslovka nebo výraz s číslicemi
D	adverbium	příslovce
I	interjekce	citoslovce
J	konjunkce	spojka
N	substantivum	podstatné jméno
P	pronomen	zájmeno
R	prepozice	předložka
T	partikule	částice
V	verbum	sloveso
X	neznámý, neurčený, neurčitelný slovní druh	
Z	interpunkce, hranice věty	

Tabulka A.1: Značky slovních druhů

kniha	označení v textu
Stručný úvod do LATEXu	A
Velký průvodce protokoly	B
Lieovy algebry	C
Fyzikální chemie	D
Lineární algebra	E
Hakl	F

Tabulka A.2: Data — použité knihy pro testování

Dodatek B

Výsledky experimentů a případová studie

Experimentování s různými rozsahy nabízených rejstříků.

kniha	unigram nabízených výrazů	unigram nalezených výrazů	%	bigram nalezených výrazů	%	trigram nalezených výrazů	%
A	109	1/68	1.47	0/156	0.00	0/39	0.00
B	769	38/88	43.18	103/170	60.59	14/29	48.28
C	109	2/15	13.33	13/48	27.08	6/23	26.09
D	695	114/461	24.73	43/110	39.09	5/29	17.24
E	255	8/48	16.67	53/151	35.10	13/41	31.71
F	145	0/3	0.00	6/33	18.18	5/25	20.00

Tabulka B.1: Rozsah — minimální množství nabízených výrazů

kniha	unigram nabízených výrazů	unigram nalezených výrazů	%	bigram nalezených výrazů	%	trigram nalezených výrazů	%
A	255	1/68	1.47	2/156	1.28	0/39	0.00
B	1539	58/88	65.91	116/170	68.24	17/29	58.62
C	219	2/15	13.33	15/48	31.25	8/23	34.78
D	1392	189/461	41.00	56/110	50.91	5/29	17.24
E	512	16/48	33.33	71/151	47.02	15/41	36.59
F	329	0/3	0.00	11/33	33.33	7/25	28.00

Tabulka B.2: Rozsah — dvojnásobné množství nabízených výrazů vzhledem k B.1

filtr
filtrace
filtrace temporální
filtr perceptuální
fir
foném
fonetika
fonologie
forma
forma fonetická
fourierka
frekvence
frekvence normovaná
frekvence vysoká
frekvence vzorkovací
frekvence základní tón
frikativa
frownie
funkce
funkce autokorelační
funkce hustoty
funkce přenosová
funkce spektrální
funkce typ
fysiologie
gain
gaussovka
gramatika
gsm
sada
samohláska
sc viterbi
searching
segmentace
sekvence
sekvence stavová
sekvence vektorů
sémantika
semiokluzíva
separace
setrvačnost
shluk
shlukování
signál
signál celý
signál chybový
signál náhodný
signál periodický
signál řečový
signal-to-nois
signál vzorkovaný
signum
simulace
sinusovka
skrytý markovův model
slovník velký
slovo
slovo izolované

Tabulka B.3: Ukázka rejstříku studijní opory

Literatura

[1] Čsn iso 999-1998 zásady zpracování, uspořádání a grafické úpravy rejstříků, 1998.

[2] Apte, C. a Damerau, F. a Weiss, S. M. Text mining with decision trees and decision rules. (květen 2007).

[3] Csomai, A. a Mihalcea, R. Creating a testbed for the evaluation of automatically generated back-of-the-book indexes. http://www.cs.unt.edu/∼rada/papers/csomai.cicling06.ps. (prosinec 2006).

[4] Hajič, J. Popis morfologických značek. (květen 2007).

[5] Kilgarriff, A. a Rychly, P. a Smrz, P. a Tugwell, D. The sketch engine. In *Proceedings of the Eleventh EURALEX International Congress*, pages 105–116, 2004.

[6] Kolar, M. a Vukmirovic, I. a Bašić, B. D. a Šnajder, J. Computer-aided document indexing system. http://cit.zesoi.fer.hr/downloadPaper.php?paper=773.pdf/, 2005. (květen 2007).

[7] Kolektiv autorů ÚFAL MFF, Karlova Univerzita. Prague dependency treebank. (květen 2007).

[8] Manning, C. D. a Schütze, H. *Foundations of Statistical Natural Language Processing*. The MIT Press, June 1999.

[9] Schütze, H. The hypertext concordance: A better back-of-the-book index. In Didier Bourigault, Christian Jacquemin, and Marie-Claude L'Homme, editors, *Proc. of Computerm '98*, pages 101–104, Montreal, Canada, 1998.

[10] Sedláček, R. Morfologický analyzátor češtiny ajka, 1999. (květen 2007).

[11] Wacholder, N. a Evans, D. K. a Klavans, J. L. Automatic identification and organization of index terms for interactive browsing. In *JCDL '01: Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*, pages 126–134, New York, NY, USA, 2001. ACM Press.

[12] WWW stránky. Alta-vista. http://www.altavista.com/. (květen 2007).

[13] WWW stránky. Wikipedia, 2007. (květen 2007).

[14] Zipf, G. K. *Human Behavior and the Principle of Least Effort*. Addison-Wesley (Reading MA), 1949.

Index