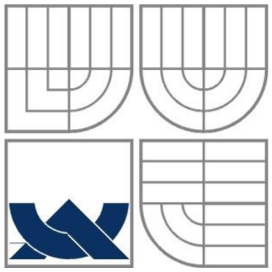
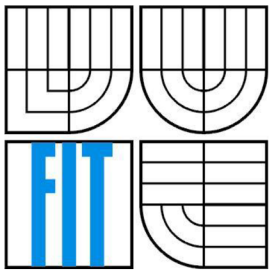


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

SLOVENSKÁ LEMMATIZACE

SLOVAK LEMMATIZATION

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

Šimon Lipták

VEDOUCÍ PRÁCE
SUPERVISOR

Doc. RNDr. PAVEL SMRŽ, Ph.D.

BRNO 2016

Zadání bakalářské práce/18337/2015/xlipta02

Vysoké učení technické v Brně - Fakulta informačních technologií

Ústav počítačové grafiky a multimédií Akademický rok 2015/2016

Zadání bakalářské práce

Řešitel: **Lipták Šimon**
Obor: Informační technologie
Téma: **Slovenská lemmatizace**
Slovak Lemmatization
Kategorie: Umělá inteligence

Pokyny:

1. Seznamte se s metodami používanými pro morfologickou analýzu a lemmatizaci (případně jen určování kmene slov) v jazycích s bohatou morfologií, pro efektivní ukládání slovníků a výkonné zpracování textů, např. pro indexování.
2. Navrhněte a implementujte systém, který s využitím existujících datových zdrojů dokáže lemmatizovat slovenská slova, která nejsou ve slovníku, a generovat správné tvary slov pro daný větný kontext.
3. Zpracujte slovenská data pro vybranou implementaci "stemmingu" (např. Snowball Stemmer - http://snowball.tartarus.org/).
4. Vyhodnoťte úspěšnost vytvořeného systému a porovnejte s dostupnými alternativami.
5. Diskutujte výhody a nevýhody zvoleného přístupu a možnosti automatického získávání morfologické informace z korpusových textů.

Literatura:

- Manning, C. D., Schütze, H., Foundations of Statistical Natural Language Processing, MIT Press, 1999, ISBN 0-262-13360-1.

Pro udělení zápočtu za první semestr je požadováno:

1. Funkční prototyp

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

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

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

Vedoucí: **Smrž Pavel, doc. RNDr., Ph.D.**, UPGM FIT VUT
Datum zadání: 1. listopadu 2015
Datum odevzdání: 18. května 2016

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

L.S.

doc. Dr. Ing. Jan Černocký
vedoucí ústavu

Abstrakt

Cieľom tejto bakalárskej práce bolo zoznámiť sa s nástrojmi a metódami pre morfologickú analýzu a lematizáciu slov, navrhnúť a implementovať systém, ktorý dokáže lematizovať slovenské slová, ktoré sa nenachádzajú v slovníku a následne vypísať vyskloňované tvary, spracovať slovenské dáta pre implementáciu stemmingu. Na záver vyhodnotiť úspešnosť na základe testovania a porovnať s dostupnými alternatívami.

Abstract

Aim of this bachelor thesis was to become familiar with the tools and methods for morphological analysis and lemmatization of words, to design and to implement a system for lemmatization of slovak words, which are not in dictionary and then to write their forms, to process slovak data for implementation of stemming. At the end to score prediction based on testing and to compare with available alternatives.

Kľúčové slová

lematizácia, lema, systém pre lamatizáciu slov, morfologická analýza, automatická indexácia, stemming

Keywords

lemmatization, lemma, system for lemmatization of words, morphological analysis, automatic indexation, stemming

Citácia

LIPTÁK, Šimon. *Slovenská lemmatizace*. Brno, 2016. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Pavel Smrž.

Slovenská lemmatizace

Prehlásenie

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

.....
Šimon Lipták
17. mája 2016

Poďakovanie

Touto cestou by som sa chcel poďakovať pánovi docentovi Smržovi za odbornú pomoc pri písaní tejto bakalárskej práce a za poskytnutie slovníka, s ktorým systém lematizácie slov pracuje.

Obsah

1 Úvod

Jazyk je jeden z najdôležitejších systémov, ktorý nás sprevádza počas celého života, je našou súčasťou. Sprevádza nás od narodenia až po koniec nášho života na Zemi. Pomocou jazyka udržiavame vzťahy, vyjadrujeme svoje pocity, či sa vzdelávame. Je veľmi dôležitým nástrojom predávania a posúvania informácií.

Ak sa pozrieme na jazyk ako na dorozumievací systém, dá sa povedať, že je špecifickým nástrojom komunikácie, ktorý sa využíva len medzi ľuďmi. Aj ostatní príslušníci živočíšnej ríše majú schopnosť komunikovať. No iba ľudia majú schopnosť kreatívne produkovať a prijímať jazyk. Jazyk je najväčšia vymoženosť človeka, je to vec, ktorá ho odlišuje od ostatných živočíchov na Zemi.

V období elektronických médií, modernej výpočtovej techniky a internetovej komunikácie si len veľmi ťažko vieme predstaviť imaginárny priestor bez znakov, textov, jazykov, priestor, v ktorom by existovali bytosti bez možnosti vzájomnej komunikácie.

Z historického hľadiska sa slovenský jazyk vyvíjal nepretržite od 5. storočia ako osobitný slovanský jazyk, ale dlhé storočia zostal jazykom ľudu a ľudovej slovesnosti. Ako administratívny, liturgický jazyk, či jazyk vedy sa na území dnešného Slovenska používali iné jazyky (staroslovienčina, latinčina, čeština/slovakizovaná čeština, neskôr i nemčina a maďarčina). Slovenský jazyk patrí spolu s češtinou, maďarčinou, poľštinou a srbčinou medzi západoslovanské jazyky, ktoré majú bohatú morfológiu. Konkrétne v slovenčine, ale nepochybne aj v iných jazykoch, sa jednotlivé slová môžu vyskytovať vo viacerých tvaroch, čo výrazne komplikuje prácu so slovenským textom. Okrem rôznych tvarov má slovenčina veľké množstvo výnimiek a musia sa dodržiavať konkrétne pravidlá pri odvodzovaní jednotlivých slov.

Preto pri práci s textom je nevyhnutným krokom slová lematizovať, čiže upravovať na základný tvar slova, tzv. lemu. Ale ani táto metóda nie je niekedy postačujúca, pretože ďalším problémom je polysémia (mnohoznačnosť), čiže jedno slovo nadobúda viacero významov. Pre takéto slová potrebujeme poznať celý kontext.

Je podstatné uvedomiť si, prečo je lematizácia dôležitou časťou práce s textom. Odpoveďou bude nasledujúci príklad. Predstavme si, že potrebujeme zistiť nejaké informácie na tému „zberný dvor“. Slovné spojenie zadáme do vyhľadávača a ak budeme mať šťastie, zobrazia sa nám výsledky vyhľadávania, avšak môže nastať situácia, kedy žiadne stránky nenájdeme a dôvodom bude, že žiadne takéto stránky neexistujú. Ale je možné, že existujú stránky, kde sa slovné spojenie vyskytuje v inom tvare ako je základný tvar, čiže nominatív jednotného čísla. Práve v tejto chvíli je dôležité, aby sme text lematizovali a tým urýchlili vyhľadávanie konkrétnej informácie.

Táto bakalárska práca je zameraná na vytvorenie kompletného postupu lematizácie. Skladá sa z teoretickej a empirickej časti.

V teoretickej časti sa nachádza postup morfologickej analýzy, čo sú to morfologické značky a z čoho sa skladajú, postup automatickej indexácie textov a na čo slúži, vysvetlená morfológia slovenčiny a spôsob tvorenia slov v rôznych jazykoch.

V empirickej časti je vytvorený systém pre lematizáciu slov, spracované slovenské dáta pre implementáciu stemmingu a v závere sú vypracované štatistiky lematizácie, ktoré sú porovnané s existujúcimi systémami.

2 Analýza témy

V tejto kapitula sa vysvetľuje termín morfologická analýza, rozdiel medzi lematizáciou a stemmingom, čo je to strojový slovník, k čomu slúži a čo je jeho obsahom, čo je to automatická indexácia, ako sa delí a popisuje jej postup, rozoberá morfológiu slovenčiny a iných jazykov.

2.1 Morfologická analýza

Pretým ako sa začneme venovať morfologickej analýze musíme si priblížiť, čo je to vlastne morfológia jazyka. Morfológia vznikla z dvoch gréckych slov „morphé“ (tvar) a „logos“ (slovo, náuka). Morfológia, čiže tvaroslovie je jazykovedná náuka o gramatických tvaroch slov, ako aj o slovách, ktoré majú funkciu tvarov. Morfológia je teda náuka o tvarovej rovine v systéme jazyka. Jazykový systém sa skladá z niekoľkých rovín. Sú to roviny, ktoré predstavujú osobitné čiastkové systémy, ktoré sa navzájom dopĺňajú. Rozlišuje sa zvuková (fonetická), slovníková (lexikálna), tvarová (morfologická), skladobná (syntaktická) rovina a popri nich aj slohová (štýlová) rovina.

Každá morfologická jednotka má dve stránky: formu a obsah. Forma a obsah sú paralelné, nie sú však symetrické. Forma aj obsah tvaru gramatického slova i morfémy majú svoju vlastnú vnútornú zákonitosť, ale súčasne rešpektujú zákonitosť svojho náprotivku. Štruktúra morfologickej roviny je komplexná, je vybudovaná na slovných druhoch, na morfologických kategóriách a na rozličných tvarotvorných postupoch. [3].

Morfologická analýza je dôležitou etapou pri strojovom spracovaní textu z hľadiska informatických potrieb. Poznáme dva typy morfologickej analýzy:

1. Úplná analýza
2. Lematizácia

Pri automatickej morfologickej analýze sa vychádza z dvoch základných súborov dát. Prvým súborom je slovník kmeňov slov. Druhým súborom je zoznam koncoviek slov, zoznam prípon pre stupňovanie.

Koncovky sú uložené vo forme tabuliek a každej koncovke je priradený údaj o tom pre aký vzor, druh, osobu a podobne je koncovka prípustná. Postup pri analýze riadi algoritmus morfologickej analýzy, ktorý pozostáva z čítania textu, segmentácii vety, hľadania správneho tvaru v slovníku kmeňov a v zozname koncoviek. Vstupný tvar sa hľadá v slovníku kmeňov a postupne sa porovnáva s konkrétnymi kmeňmi slov až pokiaľ sa nenájde najdlhší tvar , do ktorého sa analyzovaný slovný tvar zmestí [14].

2.1.1 Úplná analýza

Úplná analýza je typ morfologickej analýzy, ktorej úlohou je získať pre zadané slovo úplné morfologické informácie, čiže informácie o základných kategóriách ako sú: rod, číslo, pád, osoba, čas,

stupeň atď. Výsledky úplnej analýzy slúžia ako vstupné informácie pre ďalší stupeň analýzy a tou je syntaktická analýza [14].

2.1.2 Lematizácia

Lematizácia je súčasť automatickej morfologickej analýzy. Je to proces vytvárania základného slovníkového tvaru slov, tzv. lemy, za pomoci morfologickej analýzy a slovníka. V rámci vytvárania lemy môžeme využiť len tzv. paradigmatickú morfológiu, čo znamená, že môžeme použiť skupinu slov s rovnakým ohýbaním, alebo môžeme zahrnúť aj lexikálnu deriváciu , napríklad tvaru „písaný“ priradíme infinitív „písať“. Lematizáciu je možné vykonávať tromi spôsobmi:

1. **pomocou slovníka kmeňov alebo koreňov** – výhodou tejto metódy je minimálna chybovosť, nevýhodou rozsiahlosť slovníka a jeho prípadné obmedzenia na špecifický odbor.
2. **odstránením afixov** – tzn. sufixov (prípon) a prefixov (predpôn). Ide o najčastejšie používanú metódu s tým, že príslušný algoritmus je zvyčajne schopný zohľadňovať aj nepravidelnú flexiu (napríklad hláskové zmeny - soli x soľ). Afixy môžu byť odstraňované na základe zoznamov sufixov a prefixov alebo na základe pravidiel, podľa ktorých sú konkrétne afixy generované.
3. **štatisticky na základe variety** (verzia jazyka so špeciálnou komunikačnou funkciou) **po sebe nasledujúcich písmen v slove** (letter successor variety stemmers) – kedy sa pomocou frekvencie jednotlivých zhlukov písmen stanovuje, či sa jedná o prefix, koreň alebo sufix. Táto metóda je nezávislá na jazyku a dokáže pružne zohľadňovať nové dokumenty v databáze, nedokáže však rozlíšiť inflexné a derivačné (slovotvorné) afixy [12].

Ľudia si lematizáciu veľmi často zamieňajú s termínom stemming. Stemming je veľmi podobný lematizácii. Líši sa od lematizácie hlavne tým, že nevytvára základný slovníkový tvar slova, ale vytvára koreň slova. Napríklad z anglického slova „cats“ (po slovensky „mačky“) odoberieme príponu „-s“ a dostaneme koreň slova „cat“, čo v tomto prípade je koreňom, ale aj základným tvarom. V slovenčine keď od slova mačky odstránime koncovku „-y“ dostaneme kmeň slova „mačk“ - toto by nám spravil nástroj stemmer. Pri lematizácií dochádza k výmene koncovky „-y“ za „-a“ a výsledkom je základný slovníkový tvar „mačka“.

2.2 Pravidlový a slovníkový systém

Lematizácia môže byť vykonávaná dvoma spôsobmi: na základe pravidiel alebo vyhľadávaním v slovníku. Nutnou súčasťou slovníkového systému je zoznam slov, ktorým sú priradené určité lemy. Tento zoznam slov sa nazýva slovníkom, ktorý daný program používa. Program zadané slovo vyhľadáva v slovníku a pri úspešnom nájdení slova okamžite priradí lemu. Naopak pri pravidlovom systéme je nutné vytvoriť pravidlá, na základe ktorých sa zadané slovo spracuje a vytvorí lemu. Tieto pravidlá pozostávajú z afixov jednotlivých slovných druhov, prípadne na akú príponu sa majú zmeniť.

Morfologické značkovače (morphological taggers) využívajú slovníkový systém (morfologický analyzátor Majka) alebo pravidlový systém alebo oba systémy (morfologický značkovač MorphoDiTa). MorphoDiTa využíva voľne dostupný slovník *MorfFlexCZ* a pre slová nenachádzajúce sa v slovníku využíva pravidlový systém, na základe ktorého odhaduje správnu lemu [23].

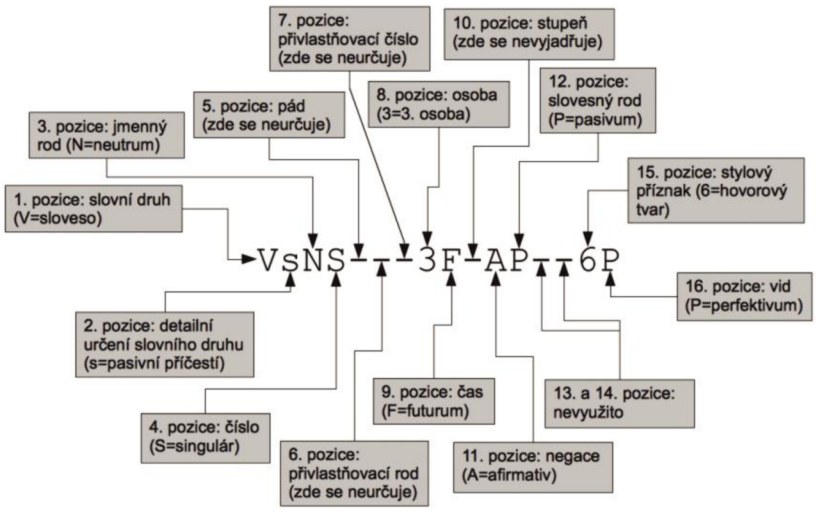
2.3 Strojový slovník

Pri práci s morfologickou analýzou algoritmus pracuje s rôznymi kmeňmi a koncovkami, ktoré musia byť uložené v súbore. Týmto súborom je strojový slovník. Slovník je lingvistickou bázou automatického morfologického analyzátora. Hlavnou stavebnou jednotkou slovníka je heslo. Obsahuje podoby kmeňov, z ktorých má každý kmeň priradený morfologický vzor (pravidlo). Podľa týchto pravidiel sa generujú trojice: základný tvar (lema), generovaný tvar (slovný tvar) a morfologická značka, tzv. tag [17].

Strojový slovník by mal byť navrhnutý tak, aby bol čo najjednoduchší a pritom aby bolo možné jeho dáta ďalej používať pre iné lingvistické bádania a experimenty. Teda strojový slovník je textový súbor, je ľahko editovateľný bežnými textovými súbormi. Koncovka tohto textového súboru je `.dic` [16].

2.4 Morfologické značky

Výstupom morfologickej analýzy, ktorá pracuje s izolovanými slovnými tvarmi (bez ohľadu na ich kontext), sú morfologické značky, tzv. tagy. Druhou časťou výsledku je základný tvar (lema). Morfologické značky slúžia k ľahšiemu hľadaniu v korpusoch a nesú informácie o slovných druhoch a gramatických kategóriách, ako sú napríklad rod, číslo, pád, osoba a podobne.



Obrázok 2.3: Morfologický reťazec

Každá značka je reťazcom 16. znakov. Každej pozícii odpovedá jedna morfologická kategória. Každej hodnote v danej kategórii odpovedá jeden znak, prevažne veľké písmeno abecedy, niekedy aj iný znak. Hodnota, ktorá nedáva zmysel je reprezentovaná znakom pomlčka „-“ [27].

Pozície v reťazci:

1. Slovný druh – podstatné meno (N), prídavné meno (A), ...
2. Detailné určenie slovného druhu – slúži k zachyteniu ďalších relevantných morfologických kategórií, ktoré sú uvedené na ďalších pozíciách
3. Rod – mužský (M), ženský (F), stredný (N), ...
4. Číslo – singulár (S), plurál (P), ...
5. Pád – nominatív (1), genitív (2), datív (3), ...
6. Privlastňovací rod – vyjadruje rod subjektu alebo objektu, ktorému sa zámeno privlastňuje, napr.: feminum – ženský rod (F), maskulinum animatum – mužský rod (M)
7. Privlastňovacie číslo – kategória, ktorá sa vyjadruje pri zámenách, napr.: singulár (S), plurál (P)
8. Osoba – 1. osoba (1), 2. osoba (3), 3. osoba (3), ...
9. Čas – budúci (F), minulý (R), prítomný (P), ...
10. Stupeň – určuje sa pri stupňovaní, napr.: 1. stupeň (1), 2. stupeň (2), 3. stupeň (3)
11. Negácia – bez negácie – afirmatív (A), s negáciou – negatív (N)
12. Aktívum / pasívum – aktívum (A), pasívum (P)
13. Táto pozícia nie je zatiaľ použitá.
14. Táto pozícia nie je zatiaľ použitá.
15. Štýlový príznak – hodnoty priradené jednotlivým javom sú založené na hodnotení, ktoré sa objavuje v jazykovedných príručkách, napr.: veľmi archaický tvar (3), skratky (8), ...
16. Vid – určuje sa pri slovesách, napr.: dokonavý (P), nedokonavý (I), ...

2.5 Korpusové morfologické analyzátory

Najznámejšie morfologické analyzátory (značkovacie programy) spracovávajú dáta v korpuse tak, že každému slovnému tvaru priradia jeho morfologický reťazec (tag). Zvyčajne sa značkujú vybrané časti korpusu v rozsahu do 10 mil. slovných tvarov. Vzniknuté súbory sú zhruba trikrát až štyrikrát väčšie než pôvodný, čo znamená, že pri ich ďalšom spracovaní vznikajú časové problémy.

1. **Probabilistický analyzátor Claws:** autor tohto analyzátora je R. Garside z Lancasteru. Tento analyzátor má vysokú úspešnosť, dosahuje len 1,7% chýb. Celkovo je Claws hybridný a pracuje s anotovaným lexikónom, ktorého súčasťou je aj zoznam základných anglických idiómov. Značkovanie sa vykonáva v niekoľkých fázach, používa sa tiež Viterbiho algoritmus, ktorý spracováva pravdepodobnosti prechodu medzi vetnými zložkami. Probabilistický prístup je motivovaný tým, že je blízky psychológii človeka.

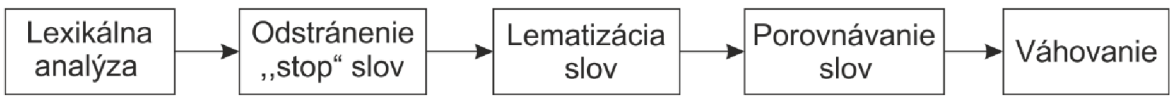
2. **Analyzátor vytvorený J.Clearem:** tento analyzátor bol vytvorený v birminghamskom COBUILDe (Collins Birmingham University International Language Database). Analyzátor takisto využíva pravdepodobnostný prístup, jeho miera úspešnosti je 95%. Autor však túto úspešnosť pokladá za dostačujúcu.
3. **Helsinský analyzátor:** tento analyzátor je založený na tzv. constraint grammars (v preklade obmedzovacie gramatiky) a je 60-krát rýchlejší než ostatné. Predpokladá použitie dvojstupňového morfologického analyzátora Kimmo od Koskenniemiho. Zatiaľ je zo všetkých najúspešnejší. Pokiaľ ide o zvládnutie viacerých jazykov. Momentálne je schopný pracovať s piatimi jazykmi.
4. **Analyzátor D.Cuttinga:** používa skrytý Markov model. Je jazykovo nezávislý, učí sa od začiatku na menších vzorkách, pracuje s váhami pravdepodobnostného výskytu, pracuje iteratívne a vo fáze učenia počíta s 18% vopred označeného textu [20].

2.6 Automatická indexácia

Automatická indexácia je proces redukcie textu pomocou počítačového programu, pri ktorom je dôležitou úlohou získanie pojmov, ktoré výstižne charakterizujú obsah textu. Automatická indexácia sa delí na dva typy: slovná indexácia a pojmová indexácia. Automatická indexácia má významné využitie pri vyhľadávačoch, ako je napríklad vyhľadávač od spoločnosti Google. Musíme myslieť na to, že kvalitné spracovanie textu umožní aj kvalitné vyhľadávanie [26].

2.6.1 Slovná indexácia

Pre slovnú indexáciu sa často používa názov automatická extrakcia. Automatická extrakcia je jednoduchšou a ľahšie programovateľnou metódou, v súčasnosti na jej základe pracuje väčšia časť funkčných systémov automatickej indexácie. Účelom automatické extrakcie je vybrať priamo z textu dokumentu také termíny, ktoré vyjadrujú jeho obsah. Pretože pre indexáciu sú vhodné iba niektoré slová a frázy z textu dokumentu, bola vyvinutá rada predovšetkým štatistických a matematicko-lingvistických metód, ako tieto relevantné termíny v texte identifikovať a ako ich z textu extrahovať. Štandardný postup pri extrakcii termínov sa skladá z niekoľkých procedúr, nižšie uvedená schéma obsahuje základné procedúry a znázorňuje postupnosť krokov [12].



Obrázok 2.4.1: Jednotlivé kroky slovnej indexácie

Slovná indexácia pozostáva z niekoľkých fáz:

1. Lexikálna analýza

2. Odstránenie „stop“ slov
3. Lematizácia slov
4. Porovnávanie slov
5. Váhovanie

2.6.1.1 Lexikálna analýza

Lexikálna analýza identifikuje jednotlivé slová v texte dokumentu. Identifikáciu jednotlivých slov vykonáva počítačový program pomocou algoritmov. Jednotlivé slová sa rozpoznávajú podľa medzier a spojovníkov. Problém nastáva pri rozpoznávaní skratiek, ktoré sú na identifikáciu zložitejšie. Podstatne zložitejšia je identifikácia súsloví, ktoré sú z hľadiska sémantické nosnosti a selektívne sily podstatne významnejšie ako jednotlivé slová. Selektívna sila slúži pre stanovenie miery, do akej miery sú schopné indexačné termíny efektívne vyhľadávať dokumenty [12].

2.6.1.2 Odstránenie „stop“ slov

Po lexikálnej analýze nasleduje odstránenie tzv. „stop“ slov, teda slov, ktoré nemajú žiadnu informačnú hodnotu, napríklad spojky, častice, predložky, citoslovcia. Frekvenčnú analýzu výskytu slov vytvoril americký lingvista a psychológ George Kingsley Zipf. Práve o túto analýzu sa princíp odstránenia „stop“ slov opiera. Na základe Zipsfovho zákona sa skonštruovali slovníky hľadaných slov a nevýznamných slov, ktoré sú základom pri automatickej indexácii. Vylúčením „stop“ slov sa dosahuje zníženie počtu termínov reprezentujúcich dokumenty, čím sa redukuje príznakový priestor a zefektívňuje sa činnosť algoritmov klasifikácie, zhlukovania alebo vyhľadávania informácií.

Nevýhodou je strata určitej časti informácií z textu, napríklad, ak sa odstráni chybne detekované „stop“ slovo. Tiež celkom neplatí, že „stop“ slová nenesú žiadnu informáciu o obsahu textu. Neplnovýznamové (pomocné) slová sa zúčastňujú na jazykovom prejave a často v ňom plnia dôležitú funkciu, napríklad modifikujú význam okolitých plnovýznamových kontextovo zapojených slov. Čiastočne sa táto druhá nevýhoda dá eliminovať použitím vhodného frázového slovníka, v úplnosti sa však tento problém rieši komplexnou syntakticko-sémantickou analýzou [17].

2.6.1.3 Lematizácia slov

Lematizácia je proces, ktorý z tvaru slova v texte určí základný tvar (lemu). Najčastejšie sa tento proces vykonáva odstránením slovotvorných, pádových a iných predpôn a prípon. Tento proces vykonáva počítačový program tzv. lematizátor [17].

2.6.1.4 Porovnávanie slov

Po predchádzajúcich fázach prichádza fáza porovnávania slov. Slová, ktoré prešli cez predchádzajúce fázy sú porovnávané s termínmi, ktoré sa nachádzajú v riadiacich slovníkoch, deskriptoroch, predmetových heslároch alebo tzv. tezauroch. Tezaurus je slovník riadeného selekčného jazyka, ktorý je usporiadaný tak, aby vyjadroval vzťahy medzi jednotlivými pojmami [18].

2.6.1.5 Váhovanie

Poslednou fázou slovnej indexácie je váhovanie. Každé slovo, ktoré sa dostalo až sem, má pre obsah textu rôznu dôležitosť. Váhovanie je úprava termínov dokumentu. Váhovanie prebieha v dvoch základných rovinách.

Prvá rovina váhovania je na základe počtu výskytov v samotnom dokumente, tzv. lokálne váhovanie $L(k_t,d_i)$ a druhou rovinou je globálne váhovanie $G(k_t)$. Váhová frekvencia termínu k_t v dokumente d_i je súčinom lokálnej a globálnej váhy, teda:

$$w = L(k_t,d_i) \times G(k_t)$$

Druhou rovinou je globálne váhovanie $G(k_t)$, ktoré určuje, aký významný je termín v celom korpuse dokumentov. Globálna váha termínu $G(k_t)$ je pre každý dokument rovnaká. Štyri bežne používané globále váhovania sú: *Norm, GfIdf, Idf,* a *Entrópia*.

Tieto globálne váhy zmenšujú dôležitosť termínov, ktoré sa vyskytujú vo väčšej časti dokumentov [2].

2.6.2 Pojmová indexácia

Pojmová indexácia sa niekedy nazýva aj ako automatické priraďovanie. Systémy založené na automatickom priraďovaní používajú podstatne zložitejšie štatistické a matematicko-lingvistické metódy ako systémy založené na slovnej indexácii. Je aplikovaná rada postupov, ktoré sú doménou expertných systémov a systémov založených na umelej inteligencii, ako napríklad rámce, pravidlá, multidimenziálne priestorové modely, samo učiace sa algoritmy, pravdepodobnostné modely (napríklad lineárnej regresnej metódy, bayesovský teorém), neurónové siete a podobne [12].

2.7 Morfológia slovenčiny

Slovenčina spolu s češtinou a poľštinou patria medzi západoslovanské jazyky. Slovenčina je teda slovanský jazyk s bohatou morfológiou.

V slovenčine rozlišujeme 10 slovných druhov: podstatné mená (substantíva), prídavné mená (adjektíva), zámená (pronominá), číslovky (numeráliá), slovesá (verbá), príslovky (adverbiá), predložky (prepozície), spojky (konjunkcie), častice (partikuly) a citoslovcia (interjekcie).

Medzi ohybné slovné druhy patria podstatné mená, prídavné mená, zámená, číslovky, slovesá a medzi neohybné slovné druhy patria príslovky, predložky, spojky, častice a citoslovcia [3].

Tvorením slov sa zaoberá veda – derivatológia. Slovotvorné prostriedky predstavujú materiálnu časť slovotvorného systému, ktorý tvoria slovotvorné základy a slovotvorné formanty. Slovotvorným základom býva zvyčajne koreň slova. Slovotvorný formant je časť slova, ktorá sa pripája k slovotvornému základu. Medzi slovotvorné formanty patria:

1. slovotvorné predpony (prefixy), napríklad: *na-, proti-*
2. slovotvorné prípony (sufixy), napríklad: *-ček, -ka*

3. súbor gramatických morfém, napríklad: *nakupovať – nákup-0, nákup-ca*
4. zvratný komponent, napríklad: rozhodnúť *sa.*

Medzi slovotvorné postupy, ktorými sa v jazyku realizuje onomaziologická štruktúra myšlienkového obsahu patrí:

1. **Odvodzovanie** (derivácia): je základný spôsob tvorenia slov, pri ktorom sa na základe existujúceho motivujúceho slova zmenou jeho morfologickej stavby tvorí nové slovo. Rozlišujeme odvodzovanie preponami, príponami, súborom gramatických morfém, zvratnou derivačnou morfémou a postfixom.
2. **Transfexia: je** slovotvorný postup, v ktorom sa odvodené slovo tvorí gramatickými morfémami.
3. **Reflexivizácia:** je tvorenie slov samostatnou derivačnou morfémov.
4. **Skladanie** (kompozícia): je tvorenie slov spájaním lexikálnych morfém do jednoslovného pomenovacieho útvaru.
5. **Akronymizácia:** je tvorenie slov skracovaním tak, že sa použijú len iniciály, respektíve prvé slabiky viacslovných pomenovaní.
6. **Skracovanie** (abreviácia): je čiastočné vypísanie slova v texte, skrátenie pomenovania.
7. **Univerbizácia:** je proces, pri ktorom z viacslovných pomenovaní vznikajú pomenovania jednoslovné.
8. **Multiverbizácia:** je proces tvorenia viacslovných pomenovaní hlavne tam, kde slovotvorné možnosti jazyka nedovoľujú vytvoriť jednoslovné pomenovanie [10].

2.7.1 Podstatné mená

Podstatné mená sú základným slovným druhom označujúcim nástroje, osoby, zamestnania, predmety a podobne. Substantíva patria medzi ohybné slovné druhy. V morfologickej analýze sa v podstatných menách označuje pád, číslo a rod. podstatné mená exitujú v jednotnom a množnom čísle, ale existuje tu výnimka a tou sú pomnožné podstatné mená, napríklad slovo *ústa*.

Podstatné meno môže mať v slovenčine len jednu rodovú charakteristiku. Môže nadobúdať rod mužský, ženský a stredný. Rovnako ako rod vyplýva z povahy podstatného mena, tak aj číslo z nej vyplýva. Podstatné mená sa v slovenčine vyskytujú v jednotnom (singulári) a množnom (pluráli) čísle. Špeciálnou kategóriou sú pomnožné podstatné mená, ktoré majú pre jednotné a množné číslo rovnaký tvar, napríklad slovo *dvere*.

Na vyjadrenie toho, že pomenovaný jav vstupuje do špecifických vzťahov k iným javom, má podstatné meno pády. V slovenčine rozlišujeme sedem pádov (nominatív, genitív, datív, akuzatív, vokatív, lokál, inštrumentál).

Pri podstatných menách určujeme aj vzor, podľa ktorého sa dané slovo skloňuje. Pri podstatných menách poznáme 16 hlavných (chlap, hrdina, dub, stroj, žena, ulica, dlaň, kosť, gazdiná, idea, pani, mať, mesto, srdce, vysvedčenie, dievča) a 114 vedľajších skloňovacích vzorov [9].

2.7.2 Prídavné mená

Prídavné mená sú ohybné (flektívne) plnovýznamové slová, ktoré pomenúvajú statické príznaky vecí. Majú gramatické kategórie rod, číslo a pád, ktoré sa s hodujú s gramatickými kategóriami podstatným mien. Delia sa na vlastné a privlastňovacie. Prídavné mená majú 7 vzorov (napríklad: pekný, cudzí, otcov, matkin, páví), podľa ktorých sa skloňujú a majú osobitné tvary na vyjadrenie relatívneho stupňa príslušnej vlastnosti. Prvý stupeň (pozitív) má základnú podobu prídavného mena, druhý stupeň (komparatív) sa tvorí pridaním prípony *-ší* alebo *-ejší* a tretí stupeň (superlatív) sa tvorí z druhého stupňa pridaním predpony *naj-* [6].

2.7.3 Zámená

Zámená sú slová so všeobecným významom, ktoré nepomenúvajú javy skutočnosti priamo, ale označujú, naznačujú ich tak, že ukazujú alebo odkazujú na ne z hľadiska podávateľa a pritom vyjadrujú gramatické významy konkrétnych názvov. Zámená sa podľa vecného významu rozdeľujú na 6 druhov: osobné, zvratné, ukazovacie, opytovacie, neurčité a vymedzovacie [9].

2.7.4 Číslovky

Číslovky sú ohybné aj neohybné slová, ktoré pomenúvajú, počet, podielovosť, poradie, násobnosť deja a vecí. Pri gramatickom určovaní kategórií uplatňujú s istými obmedzeniami gramatické kategórie podstatných mien, prídavných mien a prísloviek. Niektoré číslovky majú samostatné skloňovanie [11].

2.7.5 Slovesá

Slovesá patria medzi ohybné slovné druhy, ktoré pomenúvajú dynamické príznaky vecí. Pri slovesách sa určujú gramatické kategórie: osoba, číslo, čas, spôsob, rod a vid. Slovesá majú aj nepravidelné slovesá, napríklad slovo *byť*. Vid je lexikálno-gramatická kategória slovesa. Môže byť dokonavý alebo nedokonavý vid. Dokonavý vid je charakterizovaný faktovosťou a nedokonavý vid je charakterizovaný priebehovosťou. Slovesá majú 24 slovesných vzorov (napríklad: chytať, klaňať, čítať, rozumieť), podľa ktorých sa časujú [9].

2.7.6 Príslovky

Príslovky patria medzi neohybné plnovýznamové slovné druhy, ktoré vyjadrujú okolnosť alebo vlastnosť slovesného deja. Tvoria sa gramatickými morfémami *-e, -y, -o*. Rozlišujeme päť druhov prísloviek: spôsobové, zreteľové, časové, priestorové a stavové [13].

2.7.7 Predložky

Predložky sú neohybné slovná, ktoré v spojení so zámenami, podstatnými menami a číslovkami vyjadrujú okolnostné, predmetové a prívlastkové vzťahy. Predložky sa nevyskytujú samostatne. Delia sa na jednoduché , napríklad *bez cez, do , k* a zložené, napríklad *ponad, popod* [8].

2.7.8 Spojky

Spojky patria medzi neohybné slovné druhy. Ich primárnou funkciou je spájacia, čiže vzťažná funkcia. Delia sa na jednovýznamové a viacvýznamové spojky [9].

2.7.9 Častice

Častice sú pomocné slová, ktorými podávateľ nadväzuje na kontext a pritom vyjadruje rozličné významové odtienky jednotlivých výrazov. Častice paria medzi neohybné slovné druhy, a preto na ich klasifikáciu nemožno použiť morfologické kritériá [9].

2.7.10 Citoslovcia

Citoslovcia sú neohybné slová, ktoré vyjadrujú spontánne, pojmovo nespracované zážitky z oblasti citu, vôle a vnímania. Delia sa na vlastné, napríklad: *ach, jaj* a zvukomalebné (onomatopoje) citoslovcia, napríklad: *kikiriki, čľup* [9].

2.8 Tvorenie slov v iných jazykoch

Každý jazyk má svoje gramatické pravidlá a princíp tvorby slov. Pri niektorých jazykoch je tvorenie slov veľmi podobné, pri iných jazykoch sú v tvorení slov zásadné rozdiely.

Český jazyk je veľmi podobný slovenskému, preto v českom jazyku je slovotvorba (derivatológia) veľmi podobná slovenskému jazyku. Rozdeľuje sa do štyroch kategórií:

1. Odvodzovanie (derivácia)
2. Skladanie (kompozícia)
3. Prechod slova do iného slovného druhu: je spôsob tvorenia slov bez účasti morfologických prostriedkov, mení sa významová zložka počiatočného slova, jeho vetná platnosť a syntaktická funkcia.
4. Skracovanie (abreviácia) [15].

Maďarský jazyk patrí do skupiny ugrofínskych jazykov. Vyskytuje sa v ňom 18 gramatických pádov. Maďarčina sa vyznačuje rozvinutým systémom skloňovania a časovania. Na rozdiel od slovenčiny alebo češtiny, v maďarčine je možné pripojiť viacero rôznych prefixov a sufixov. Preto je maďarský jazyk typickým aglutinačným jazykom. Tvorenie slov prebieha pomocou derivačných

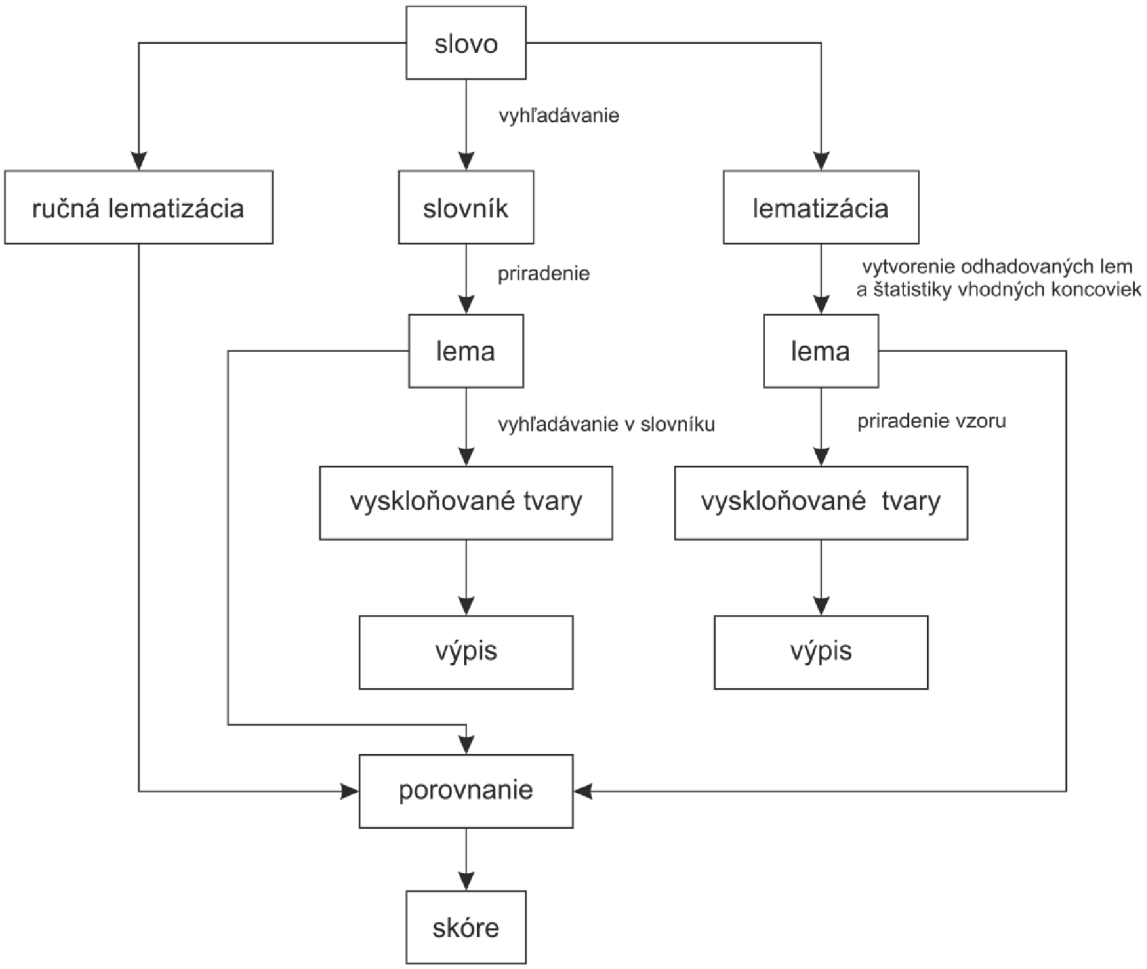
morfém a skladaním, ktoré sa ďalej delí na tvorenie zložených slov podraďovaním a tvorenie zložených slov priraďovaním (zdvojením, napodobňovaním, spájaním slov s podobným významom, spájaním slov s odlišným významom a pravým priraďovaním) [5].

Nemčina patrí medzi západogermánske jazyky. V nemčine existujú iba 4 gramatické pády (nominatív, genitív, datív a akuzatív). Slová v nemčine sa tvoria tromi základnými postupmi: odvodzovaním slov príponami a predponami, skladaním slov a prechodom z jedeného slovného druhu do druhého [4].

3 Postup

V tejto kapitole je vysvetlený celý postup a kroky použité pri vytváraní a hodnotení systému, ktorý lematizuje slová a generuje ich ďalšie tvary podľa pádových prípon jednotlivých vzorov, obsah slovníka, s ktorým systém pracuje, vstupné dáta a popísaný postup pri spracovávaní slovenských dát pre implementáciu stemmingu.

3.1 Systém lematizácie slov



Obrázok 3.1: Postup pri vytváraní a hodnotení systému

3.1.1 Vstupné dáta

Vstupné dáta sú potrebné pri lematizácii, bez nich by sme nedosiahli žiadneho výsledku. Vstupnými dátami rozumieme jednotlivé slová v slovenskom jazyku. Prakticky to môžu byť všetky slovné druhy, od podstatných mien až po citoslovcia. Problém nastáva pri dvojslovných spojeniach. Lematizátor nedokáže tieto slovné spojenia rozoznať, preto vstupnými dátami musí byť práve jedno slovo.

3.1.2 Vytvorenie slovníka

Slovník je dôležitým prvkom systému. Tento slovník sme museli upraviť a vyčistiť od nepotrebných informácií. Aktuálne sa v slovníku nachádza približne 2,7 milióna slovných tvarov.

Slová v slovníku musia dodržovať predpísanú štruktúru, inak by systém zobrazil chybné výsledky. Slová sú usporiadané abecedne a podľa slovných druhov. Na každom riadku slovníka sa musia nachádzať tri stĺpce, ktoré obsahujú:

1. slovný tvar, napríklad: *jedlá*
2. základný tvar – lema, napríklad: *jedlo*
3. morfologický reťazec, napríklad: *k1gNnPc1*

V slovníku je tento príklad reprezentovaný nasledovne: *jedlá jedlo k1gNnPc1* a predstavuje to, že slovný tvar *jedlá* má základný tvar (lemu) *jedlo* a je k nemu priradený morfologický reťazec *k1gNnPc1*, ktorý hovorí, že dané slovo je podstatné meno stredného rodu v nominatíve množného čísla. Medzi každým stĺpcom sa nachádza tabulátor, aby sa stĺpce dali ľahšie rozlíšiť. Každému slovnému tvaru je priradený jeho základný tvar a morfologický reťazec, z ktorého sa môžeme dozvedieť informáciu o tom, akým slovným druhom je dané slovo, aký rod a číslo reprezentuje, v akom páde sa nachádza a podobne.

3.1.3 Morfologické značky v slovníku

Morfologické reťazce, ktoré sa skladajú z morfologických značiek, slúžia pre lepšie orientovanie sa v slovníku a hlavne obsahujú informácie o daných slovách v slovníku. Tieto informácie hovoria o slovnom druhu a gramatických kategóriách slova, ako sú osoba, číslo, čas, rod, vid, či sa jedná o negáciu alebo nesú informáciu pri stupňovaní prídavných mien. Nakoľko je na prvý pohľad nejasné, čo dané značky znamenajú, tak si ich teraz priblížime [22].

3.1.3.1 Slovné druhy

k1 – podstatné mená

k2 – prídavné mená

k3 – zámená

k4 – číslovky

k5 – slovesá

k6 – príslovky

k7 – predložky

k8 – spojky

k9 – častice

k0 – citoslovcia

kA – skratky

kX – nezaradené slová

3.1.3.2 Rod

Rod sa označuje písmenom *g* a za ním nasleduje ďalší znak:

M – mužský životný

I – mužský neživotný

N – stredný

F – ženský

R – rodina (priezviská)

3.1.3.3 Číslo

Číslo sa označuje písmenom *n* a za ním nasleduje znak:

S – singulár

P – plurál

D – duál

R – hromadné označenie členov rodiny (napríklad: Novákovci)

3.1.3.4 Pád

Pád sa v morfologickom reťazci označuje písmenom *c* a za ním nasleduje daný pád:

1 – nominatív

2 – genitív

3 – datív

4 – akuzatív

5 – vokatív

6 – lokál

7 – inštrumentál

3.1.3.5 Negácia

Pri prídavných menách, slovesách a príslovkách sa určuje negácia, označuje sa písmenom *e* a za ním nasleduje znak:

A – afirmácia

N – negácia

3.1.3.6 Stupeň

Prídavné mená sa dajú stupňovať, preto je potrebné túto informáciu zapísať. Stupeň sa označuje písmenom *d* a za ním nasleduje:

1 – pozitív

2 – komparatív

3 – superlatív

3.1.3.7 Osoba

Pri slovesách a pri zámenách sa určuje osoba. Označuje sa písmenom *p* a za ním nasleduje:

1 – prvá osoba

2 – druhá osoba

3 – tretia osoba

X – prvá alebo druhá alebo tretia osoba

3.1.3.8 Vid a typ

Kategórie vid a typ sa určujú pri slovesách. Typ sa označuje písmenom *m* a vid písmenom *a*.

a:

P – perfektum – dokonavý vid

I – imperfektum – nedokonavý vid

B – aj dokonavý aj nedokonavý vid

m:

F – infinitív

I – indikatív prítomný

R – imperatív

A – príčastie minulé

N – príčastie trpné

S – prechodník prítomný

D – prechodník minulý

B – indikatív budúci

3.1.3.9 Druh zámena

Druhy zámen sa označujú písmenami *x* a *y*.

x:

P – osobné

O – privlastňovacie

D – ukazovacie

T – vymedzovacie

y:

P – osobné

F – reflexívne

Q – opytovacie

R – vzťažné

N – záporné

I – neurčité

3.1.3.10 Druh číslovky

Tak isto ako aj pri zámenách sa druhy čísloviek označujú písmenami *x* a *y*.

x:

- *C* – základná
- *O* – radová
- *R* – druhová
- *G* – gramatika

y:

- *N* – záporná
- *I* – neurčitá

3.1.3.11 Druh príslovky

Druhy prísloviek sa označujú ako pri zámenách a číslovkách, tak aj pri príslovkách písmenami *x* a y.

x:

- *D* – ukazovacie
- *I* – vymedzovacie
- *M* – spôsobové
- *S* – stavové

y:

- *Q* – opytovacie
- *R* – vzťažné
- *N* – záporné
- *I* – neurčité

3.1.4 Lematizácia

Keď už máme vytvorený slovník a máme zadané vstupné slovo, tak ďalším krokom je lematizácia. Pri lematizácii je nutné rozlišovať dva faktory a tými sú: slovné tvary, ktoré sa v slovníku nachádzajú a neznáme slová, ktoré sa v slovníku nenachádzajú. Preto sa samotná lematizácia delí na dve časti:

1. Hľadanie základného tvaru zadaného slova v slovníku
2. Lematizácia neznámeho slova

3.1.5 Hľadanie základného tvaru zadaného slova v slovníku

Po zadaní slovného tvaru nasleduje fáza, kedy sa základný tvar hľadá v slovníku. Prehľadáva sa každý riadok slovníka a keď sa nájde dané slovo, vyberie sa z konkrétneho riadku základný tvar slova, čiže lema.

Po tomto kroku nasleduje hľadanie ostatných tvarov slov. Znovu sa prehľadáva celý slovník, riadok po riadku a ukladajú sa nájdené tvary spolu s morfologickým reťazcom do zoznamu ostatných tvarov. Vyhľadávanie prebieha na základe získanej lemy. Tieto informácie (slovný tvar, lema a ostatné tvary) sa následne vypíšu.

3.1.6 Lematizácia neznámeho slova

Problém nastáva, pokiaľ sa v slovníku základný tvar slova nenašiel, preto je nutné prejsť do fázy lematizácie neznámeho slova. Je dôležité vedieť, čo chceme dosiahnuť. Chceme dosiahnuť to, aby nám vznikol koreň slova. Využijeme metódu nazývanú stemming. Zadané slovo prechádzame odzadu znak

po znaku a ukladáme si príponu. Pokiaľ sa nám prípona rovná s niektorou príponou zo vzorov, odstránime príponu, a pokiaľ má dané slovo predponu odstránime ju tiež. Môže sa stať že prípona sa nezhoduje so žiadnou príponou zo vzorov, v tomto prípade sa môže jednať o základný tvar slova alebo ide o nejakú skratku, prípadne sa jedná o nezmyselné slovo.

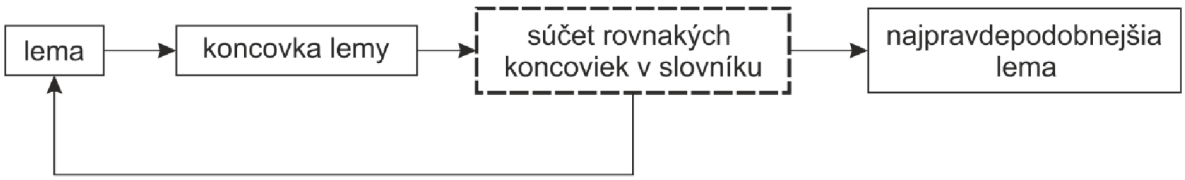
Keď máme vytvorený koreň slova, potrebujeme z neho vytvoriť základný slovníkový tvar, tzv. lemu. Pri stemmingu sme zistili vzor slova na základne prípony a daných pravidiel vzoru. Podľa tohto vzoru doplníme koreň slova príponou príslušnou základnému tvaru slova. Týmto spôsobom vytvoríme základný tvar, čiže lemu.

3.1.7 Vytvorenie zoznamu odhadovaných lem

Pri lematizácii neznámeho slova nepoznáme jeho vzor, poznáme len jeho príponu. Podľa tejto prípony a pravidiel priraďovania slov do vzorov musíme určiť, podľa ktorého vzoru vytvorí systém základný tvar. Nie je to vždy jednoznačné, o ktorý vzor alebo slovný druh ide. Preto je nutné vytvoriť zoznam najpravdepodobnejších lem. Pokiaľ sa prípona daného slova zhoduje s príponami daného vzoru a zároveň dané slovo spĺňa pravidlá daného vzoru, tak sa vytvorí lema, ktorá sa pridá do zoznamu lem. V tomto zozname sa po dokončení lematizácie nachádza niekoľko odhadov, z ktorých sa v ďalšej fáze vyberie najpravdepodobnejší základný tvar slova.

3.1.8 Výber najpravdepodobnejšej lemy

Keď máme vytvorený zoznam lem, musíme z neho vybrať práve jednu najpravdepodobnejšiu lemu. Pre výber tejto lemy je potrebné vytvoriť štatistiku výskytu koncoviek zo zoznamu lem.



Obrázok 3.5: Výber najpravdepodobnejšej lemy

Štatistika sa vytvorí tak, že vyberieme koncovku lemy zo zoznamu lem a porovnávame ju s koncovkami v slovníku, každú zhodu koncoviek pripočítavame. Tento postup použijeme pre každú lemu v zozname. Keď máme vytvorené výsledky, tak vyberieme lemu s najväčším počtom zhôd koncoviek. Týmto dosiahneme štatisticky najpravdepodobnejšiu koncovku lemy. Tento spôsob vytvárania lem nie je dokonalý. Niekedy sa stane, že dané slovo má špecifickú koncovku a preto nemusí byť výsledná lema správna v porovnaní s ručným prekladom.

3.1.9 Priradenie vzoru

Priradenie vzoru leme je dôležitou súčasťou pri vytváraní jej ďalších tvarov pri skloňovaní. Priraďovanie vzorov sa určuje podľa koncoviek lemy a pravidiel, ktoré daný vzor identifikujú. Keďže sa môže stať, že danej leme zároveň vyhovuje niekoľko rôznych vzorov, tak musíme určovanie vzoru doplniť tak, aby záviselo aj na prípone slovného tvaru. Týmto spôsobom zabezpečíme určenie vzoru základného tvaru slova a vylúčime z možností nesprávne vzory.

3.1.10 Skloňovanie

Skloňovanie je záverečná fáza systému, v ktorej už poznáme lemu zadaného slova. V tejto fáze sa vzniknutá lema vyskloňuje podľa priradeného vzoru, ktorý sme získali v predchádzajúcej fáze. Pri skloňovaní musíme použiť koreň slova a nie lemu. Keby sme použili lemu, vznikli by nám nesprávne tvary slov. Skloňovanie prebieha tak, že samotnému koreňu sa pripája postupne prípona, ktorá je pridelená určitému pádu. Po vyskloňovaní nám vzniknú tvary slova vo všetkých pádoch. Nakoniec je každému tvaru pridelený morfologický reťazec, ktorý hovorí o tom, akým slovným druhom je dané slovo, v akom je rode, či je to jednotné (singulár) alebo množné (plurál) číslo a v akom páde sa nachádza. Takto vytvorená lema a ostatné tvary, spolu s morfologickým reťazcom sa následne vypíšu.

3.1.11 Hodnotenie systému

Po vytvorení systému je potrebné tento systém ohodnotiť. Hodnotenie systému je taktiež dôležitou časťou systému. Pri testovaní sa dozvedáme, či je náš systém vytvorený správne alebo úplne nesprávne, prípadne sa v ňom vyskytujú drobné nedostatky, ktoré treba doladiť. Hodnotenie systému lematizácie slov sme vykonali tak, že sme porovnali výsledok systému a výsledok ručnej lematizácie. Ručnú lematizáciu sme si vybrali preto, aby sme zamedzili nedostatkom iných systémov a aby bol výsledok čo najpresnejší. Po prvotnom skúšobnom testovaní sme zistili niekoľko nedostatkov, ktoré bolo treba doladiť. Nedostatky sa týkali ako slovníka, tak aj skriptu. Po dokončení skúšobného hodnotenia sme vytvorili testovacie sady, o ktorých sa viac dozvieme v kapitole 4 [Testovacie sady].

3.1.12 Zhrnutie

Pri vytváraní systémov lematizácie slov použitých v tejto práci sme sa držali vyššie uvedeného postupu a každá verzia systému prešla nasledovnými fázami:

1. Získanie vstupných dát
2. Vyhľadávanie daného slova v slovníku
3. Lematizácia
4. Vytvorenie zoznamu odhadov lem
5. Výber najpravdepodobnejšej lemy

6. Priradenie vzoru
7. Skloňovanie
8. Hodnotenie systému

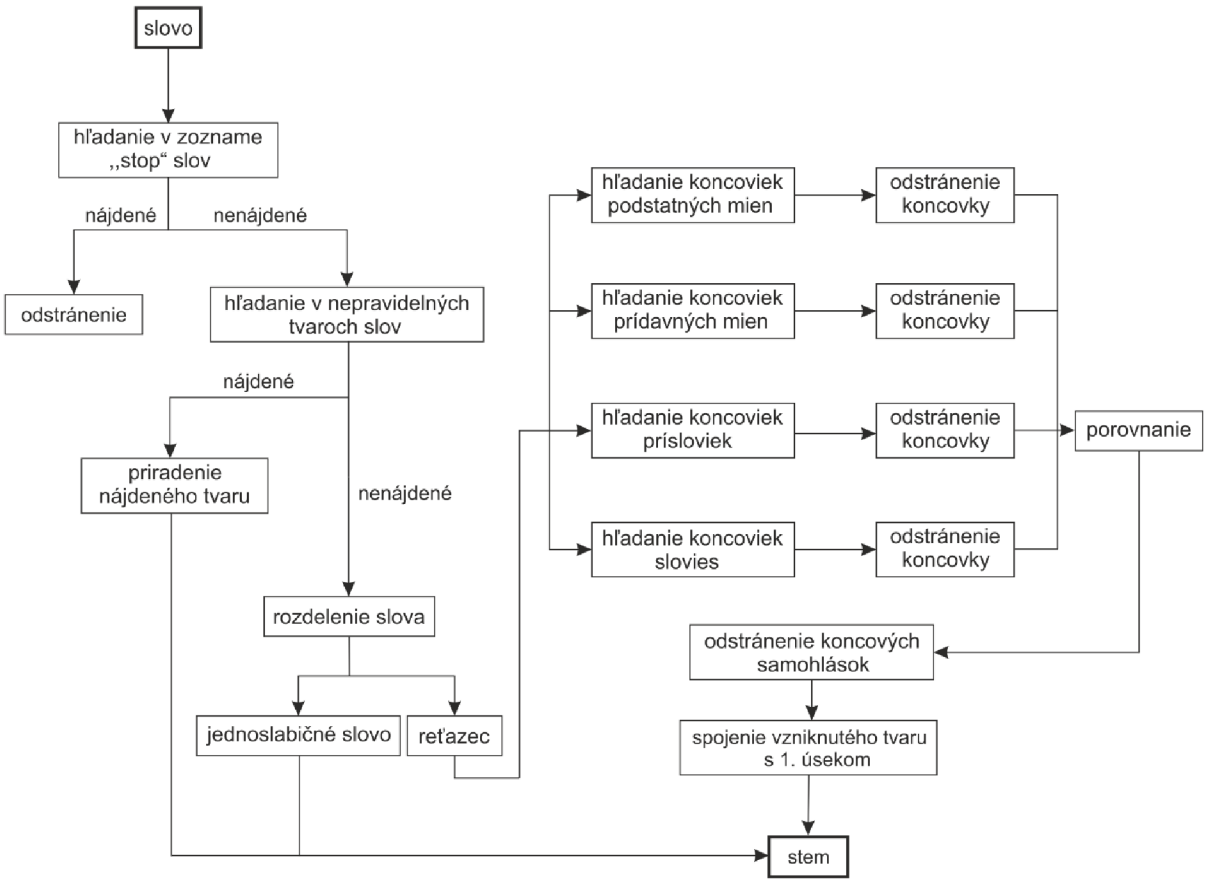
Pre bližšie znázornenie, ako vyzerá výstup zo systému lematizácie slov, je nižšie uvedený príklad, ktorý znázorňuje, že vstupným slovom je slovný tvar *armády*, základným tvarom (lemou) je slovo *armáda* a ďalej sú vypísané ostatné tvary slova:

```
armády
Základný tvar: armáda
Ostatné tvary:
armády:k1gFnPc14
armádam:k1gFnPc3
armádach:k1gFnPc6
armádami:k1gFnPc7
armáda:k1gFnSc1
armády:k1gFnSc2
armáde:k1gFnSc36
armádu:k1gFnSc4
armádou:k1gFnSc7
armád:k1gFnPc2
```

Celý systém lematizácie slov je napísaný v jazyku Python. Pri vytváraní tohto systému bolo pri úprave slovníka vytvorených niekoľko skriptov, ktoré odstránili nepotrebné informácie zo slovníka, prípadne upravili niektoré informácie v slovníku. Tieto skripty boli taktiež napísané v jazyku Python. Pri testovaní systému bolo vytvorených niekoľko skúšobných testovacích sád, pri ktorých sa dolaďovali nedostatky v skripte a v slovníku. Na záver boli vytvorené tri testovacie sady, na ktorých bol systém otestovaný a bola zmeraná úspešnosť systému.

3.2 Systém stematizácie slovenských slov

V slovenskom jazyku sa nachádzajú slová, ktoré sú rozdelené do slovných druhov. Slovné druhy môžu byť ohybné a neohybné alebo iné delenie hovorí o plnovýznamových a neplnovýznamových slovách. V systéme stematizácie slov sa zameriame hlavne na odvodené slová. Tieto slová môžu byť podstatnými menami, prídavnými menami, ale aj slovesami a príslovkami. Je dôležité, aby systém vytváral práve jeden tvar pre súbor odvodených slov, napríklad pre súbor slov *spracovať – spracovávať – spracovaný – spracovane – spracovanie* vytvorí tvar *sprac*.

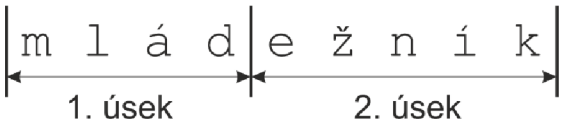


Obrázok 3.2: Postupnosť krokov systému stematizácie slov

3.2.1 Popis algoritmu

Systém stematizácie slov na začiatku načíta zadané slovo. Následne dané slovo porovnáme so slovami v zozname „stop“ slov. Pokiaľ sa v ňom slovo nachádza, tak dané slovo odstránime, nakoľko nemá pre nás žiadnu hodnotu, lebo aj pri indexácií textov sa tieto „stop“ slová vynechávajú. Ak zhoda nenastane porovnáme dané slovo s nepravidelnými tvarmi slov a v prípade zhody mu priradíme správny tvar. V prípade nezhody je nutné spracovať, respektíve rozlíšiť zadané slovo, či sa jedná o jednoslabičné alebo o viacslabičné slovo. Pri jednoslabičnom slove vychádzame z toho, že pravdepodobne nemohlo vzniknúť pridaním rôznych predpôn a prípon. Následne vypíšeme dané jednoslabičné slovo ako výsledok. Naopak, pri viacslabičných slovách je väčšia pravdepodobnosť, že slovo vzniklo z prípon

a predpôn. Takéto slovo rozdelíme na dve časti a to tak, že prvý úsek bude končiť po prvej spoluhláske, pred ktorou sa nachádza samohláska a druhý úsek bude obsahovať zbytok slova, napríklad:



S prvým úsekom slova nepracujeme, ale na reťazec druhého úseku aplikujeme gramatické pravidlá. V reťazci sa hľadajú koncovky podstatných mien, prídavných mien, prísloviek a slovies. Ak sa daná koncovka nájde, tak ju odstránime. Následne prebieha porovnanie výsledkov a vyberieme ten, z ktorého sme odstránili najdlhšiu koncovku. Na záver odstránime prípadné koncové samohlásky a vzniknutý tvar slova spojíme s prvým úsekom slova a vypíšeme vzniknutý tvar slova.

3.2.2 Zoznamy prípon

Systém stematizácie slov využíva súbory, v ktorých sú uložené prípony a predpony jednotlivých slovných druhov. Systém v týchto zoznamoch vyhľadáva prípony alebo predpony a pokiaľ nastane zhoda systém danú príponu alebo predponu odstráni. Jednotlivé koncovky boli vypísané z knihy Onomaziologická štruktúra slovenčiny [1] a z knihy Základy slovenskej lexikológie [10].

Podstatné mená sú väčšinou tvorené príponami. Môžeme ich rozdeliť do niekoľkých skupín:

1. mená nástrojov a strojov: *-ivo, -dlo, ačka, ička, ...*
2. činiteľské mená: *-teľ, -ár, -iar, -ník, -ec, ...*
3. názvy výsledkov deja: *-enka, -ina, -nica, -ivo, ...*
4. pomenovania prostriedkov: *-ie, -nica, -ník, -ka, -ica, ...*
5. pomenovania vlastností: *-oba, -osť, -stvo, -izmus, ...*

Systém pracuje s 1083 tvarmi prípon podstatných mien.

Prídavné mená tak, ako aj podstatné mená, sú prevažne tvorené príponami, napríklad: *-itý, -ov, -in, -ajší, -ný* a podobne. Systém pracuje s 543 tvarmi prípon prídavných mien.

Príslovky majú podstatne menej slovotvorných prípon ako podstatné mená alebo prídavné mená. Systém pracuje s 16 tvarmi prípon prísloviek.

Slovesá sú tvorené predponami aj príponami. Pri slovesách sú predpony dôležitejšie ako pri podstatných a prídavných menách, ale nakoľko by nám pri generovaní stemu vznikali problémy [napríklad], rozhodli sme sa predpony vynechať. Systém pracuje s 164 tvarmi prípon slovies.

3.2.3 Zhrnutie

Pri vytváraní systému stematizácie slovenských slov použitého v tejto práci sme sa držali vyššie uvedeného postupu, ktorý pozostával z fáz:

1. Načítanie slova
2. Rozdelenie slova na dva úseky
3. Hľadanie prípon a predpôn v zoznamoch prípon a predpôn daných slovných druhov

4. Odstránenie prípony alebo predpony
5. Odstránenie koncových samohlások
6. Kontrola nepravidelného tvaru
7. Výpis

Pre bližšie znázornenie výstupu programu je nižšie uvedený príklad, ktorý znázorňuje, že zadaným odvodeným slovám vygeneruje stem (koreň) *sprac* a stem *beh*.

```
spracovanie                    behávať
sprac                          beh

spracovať                      behanie
sprac                          beh

spracovávať                    behane
sprac                          beh

spracovaný                     behať
sprac                          beh

spracovane                     behavý
sprac                          beh
```

Celý systém stematizácie slovenských slov je napísaný v jazyku Python. Systém pracuje s niekoľkými súbormi, v ktorých sú zoznamy prípon a predpôn daných slovných druhov. Na záver bola vytvorená testovacia sada pre podstatné mená, prídavné mená, slovesá a príslovky, podľa ktorej sme vyhodnotili úspešnosť systému.

4 Výsledky a štatistiky

V tejto kapitole je popísaný cieľ práce, ktorý sme sa snažili dosiahnuť. Ďalej sa tu nachádza vyhodnotenie systému, popis testovacích sád, ich vytvorenie, výpočet úspešnosti systému, porovnanie systému s existujúcimi systémami, ktoré lematizujú slová a nedokonalosti, ktoré náš systém obsahuje a popis spracovávania slovenských dát pre implementáciu stemmingu.

4.1 Cieľ práce

Za cieľ práce sme si stanovili vytvoriť systém, ktorý dokáže lematizovať slová, avšak aj tie, ktoré sa v slovníku nenachádzajú a následne dokáže generovať vyskloňované tvary slov, ktorým priradí morfologický reťazec. Pri vytváraní tohto systému sme zistili, že k celkovej úspešnosti je dôležitým faktorom slovník, v ktorom sa nachádzajú rôzne tvary slov. Čím kvalitnejší, teda obsiahlejší slovník máme, tým viac zredukujeme počet slov, ktoré sa v slovníku nenachádzajú a tým dosiahneme väčšej úspešnosti. Nakoľko do slovenského jazyka sa stále pridávajú nové a nové slová, je potrebné mať takýto systém, ktorý vie lematizovať aj slová, ktoré sa nenachádzajú v slovníku. Taktiež je potrebné aktualizovať slovník o tieto nové slová, ktoré sa pridali do slovenčiny. Ďalším cieľom práce je spracovať slovenské dáta pre implementáciu stemmingu.

4.2 Systém lematizácie slovenských slov

Systém lematizácie slov spracováva jednotlivé slová a prevádza ich na základný tvar a následne vypisuje ich ostatné tvary.

4.2.1 Testovacie sady

Testovanie je dôležitou fázou hodnotenia systému. Na základe testovania sa určí celková úspešnosť systému. Aby testovanie bolo čo najobjektívnejšie, je potrebné zabezpečiť to, aby testovacie sady boli vytvorené nezávisle na systéme, čiže aby neobsahovali len slová , ktoré systém stopercentne určí správne. Preto sme vytvorili niekoľko testovacích sád, ktorých obsah sme vybrali náhodne z rôznych článkov, publikácií, odborných prác a kníh.

4.2.1.1 Články z časopisu SME

Pri tvorení tejto testovacej sady sme využili náhodné články z časopisu SME. Články, ktoré sme použili, sa týkali aktuálneho diania na Slovensku, ale aj vo svete. Táto testovacia sada obsahuje 596 slov [24].

4.2.1.2 Odborné články

Pri tvorbe tejto sady sme vychádzali z odborného článku, ktorý napísal M. Paul o deduplikácií kmeňových dát. Táto sada obsahuje 656 slov [19].

4.2.1.3 Úryvok z knihy

V tejto testovacej sade sa nachádza úryvok z knihy Bohovia z Alta, ktorú napísal Matúš Novotňák. Obsah knihy hovorí o katastrofe, ktorá postihla Zem. Konkrétne sa jednalo o výbuch sopky a dôsledky tejto katastrofy poslali ľudstvo do stredoveku. Záleží len na ľuďoch ako sa so situáciou vysporiadajú a akým spôsobom zabezpečia život na Zemi. Testovacia sada obsahuje 724 slov [7].

4.2.2 Existujúce systémy

V súčasnosti existuje niekoľko systémov, ktoré využívajú lematizáciu slov alebo stemming. Rozhodli sme sa bližšie popísať dva existujúce systémy. Prvým z nich je Snowball, ktorý využíva technológiu nazývanú stemming a druhý systém je morfologický analyzátor Majka, ktorý využíva technológiu lematizácie slov.

4.2.2.1 Snowball

Snowball je program, ktorý spracováva reťazce. Tento systém je navrhnutý pre vytváranie koreňov slov. Technológia, ktorú tento systém využíva sa nazýva stemming, čiže prevod slova na jeho koreň. Jeden z hlavných dôvodov vytvorenia stemmeru Snowball bol ten, že bolo nedostatok algoritmov využívajúcich stemming pre iné jazyky ako je angličtina. Tento systém je navrhnutý pre rôzne jazyky, ako sú angličtina, francúzština, španielčina, portugalčina, taliančina, írčina, maďarčina, čeština, rumunčina, nemčina, švédčina, ruština, dánsky jazyk, nórsky a fínsky jazyk. Snowball je založený na gramatických pravidlách ako doplnok k metódam používajúcim slovník pre vyhľadávanie a získavanie textu z dokumentov. Algoritmus, ktorý Snowball využíva je založený na sade prípon, predpôn a koncoviek, ktoré sa môžu v konkrétnom jazyku vyskytovať [25].

4.2.2.2 Morfologický analyzátor Majka

Morfologický analyzátor Majka je program, ktorý k danému slovnému tvaru priradí základný tvar (lemu) a určí slovný druh a ďalšie gramatické kategórie (rod, číslo, pád a podobne). Predchádzajúca verzia morfologického analyzátora Majka sa volala Ajka, ktorá bola napísaná v jazyku C. Analyzátor Majka používa dva systémy, jeden pre lematizáciu slov a druhý pre značkovanie slov, priradenie tzv. tagu (morfologického reťazca). V súčasnosti je Majka navrhnutá pre niekoľko jazykov, ako sú čeština, slovenčina, poľština, švédčina, nemčina, francúzština, taliančina, angličtina, portugalčina, španielčina, kataláncina, ruština, waleský jazyk, ale aj pre menej známe jazyky ako sú astúrčina a galícijčina [21].

4.2.3 Porovnanie úspešnosti s iným systémom

Pre porovnanie úspešnosti nášho systému s iným systémom sme si vybrali morfologický analyzátor Majka. Tak ako aj náš systém, tak aj morfologický analyzátor Majka využíva technológiu lematizácie. Za stopercentný výsledok sme si stanovili výsledok z ručnej lematizácie, aby sme predišli nedostatkom, ktoré obsahuje morfologický analyzátor Majka a náš systém pre lematizáciu slov.

Jednotlivé testovacie sady sme spustili na oboch systémoch. Po získaní výsledkov sme vypočítali úspešnosť nášho systému lematizácie slov a úspešnosť morfologického analyzátora Majka. Výsledky jednotlivých testovacích sád sme zapísali do tabuliek. Následne sme vypočítali priemernú úspešnosť oboch systémov a porovnali ich medzi sebou. Nakoniec sme jednotlivé výsledky testov a priemernú úspešnosť oboch systémov znázornili v grafe, aby sme všetky získané výsledky videli vedľa seba a mohli ich ľahšie porovnať.

4.2.3.1 Prvá testovacia sada

Prvá testovacia sada obsahovala články z časopisu SME. V tomto testovaní bol náš systém oproti Majke lepší o 3,51%.

počet slov	náš systém	Majka
596	98,48%	94,97%

Tabuľka 4.2.3.1: Výsledná úspešnosť pri prvej testovacej sade

4.2.3.2 Druhá testovacia sada

V tejto testovacej sa nachádzal odborný článok. Výsledok testovania úspešnosti bol oproti morfologickému analyzátoru lepší o 0,77%.

počet slov	náš systém	Majka
656	98,63%	97,86%

Tabuľka 4.2.3.2: Výsledok úspešnosti pri druhej testovacej sade

4.2.3.3 Tretia testovacia sada

Tretia testovacia sada obsahovala úryvok z knihy Bohovia z Alta. Zmeraná úspešnosť bola oproti morfologickému analyzátoru Majka lepšia o 0,97%.

počet slov	náš systém	Majka
724	99,17%	98,20%

Tabuľka 4.2.3.3: Výsledok úspešnosti pri tretej testovacej sade

4.2.3.4 Priemerná úspešnosť

Po spustení všetkých testovacích sád a vyhodnotení úspešnosti sme zobrali všetky hodnoty úspešností zvlášť pre náš systém a zvlášť pre morfologický analyzátor Majka. Pre každý systém sme vyhodnotili priemernú úspešnosť, ktorú sme vypočítali nasledovne:

$$avgu = \frac{\sum u}{n}, \text{ kde}$$

$avgu$ – priemerná úspešnosť

u – úspešnosti

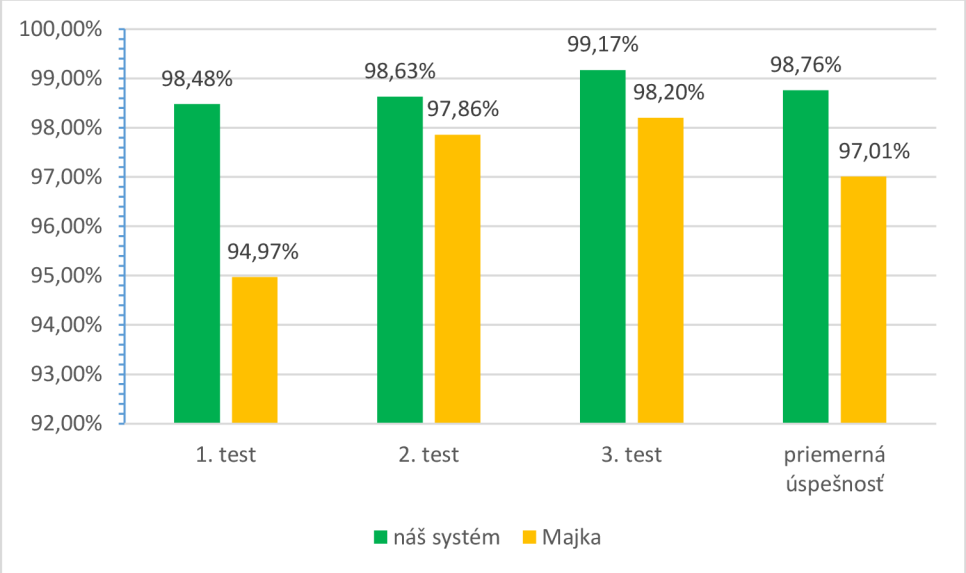
n – počet testovacích sád.

Priemerná úspešnosť morfologického analyzátora Majka je 97,01% a priemerná úspešnosť nášho systému je 98,76%.

náš systém	Majka
98,76%	97,01%

Tabuľka 4.2.3.4: Priemerné úspešnosti systémov

Získané úspešnosti z jednotlivých testovacích sád a vypočítaná priemerná úspešnosť oboch systémov bola zaznamenaná do grafu pre lepšiu prehľadnosť.



Graf 4.2.3: Znázornenie jednotlivých testov a priemernej úspešnosti

4.2.4 Výhody a nevýhody systému

Nami vytvorený systém lematizácie slov sa snaží odhadovať všetky slovné druhy. Systém pracuje na základe slovníka a na základe pravidiel odhaduje lemy pre neznáme slova. Výhodou systému je, že dokáže odhadovať aj skratkové slová (napríklad: *IS, IKEA, SME*).

Náš systém lematizácie slov nie je dokonalý, nevie určiť 100% všetkých slov, ktoré sa v slovenskom jazyku nachádzajú, preto sa tu vyskytujú nedokonalosti. Pokiaľ systém zistí, že hľadané

slovo sa v slovníku nenachádza, nastáva problém a prechádza sa do fáze lematizácie neznámeho slova. Toto neznáme slovo môže byť nejaké meno, priezvisko, podstatné meno, prídavné meno, skratka, citoslovce, teda slovo, ktoré nesie nejakú informáciu. Ale toto slovo môže byť aj nezmyselné slovo. Systém tieto nezmyselné slová nevie rozlíšiť a pracuje s nimi, ako by boli obyčajné slová nesúce nejakú informáciu, nejaký význam. Ďalšou chybou vyskytujúcou sa vo výsledkoch boli mená, priezviská a činné príčastia slovies, systém niekedy tieto slová určil nesprávne.

Čas potrebný na získanie výstupu zo systému lematizácie slov je ďalším problémom, respektíve nepríjemnosťou. Pri lematizácii slova a následnom hľadaní ostatných tvarov v slovníku systém prehľadáva niekoľkokrát slovník, ktorý obsahuje veľký počet slovných tvarov a to je príčinou, že systému trvá niekoľko desiatok sekúnd, kým zobrazí výsledok. Tento problém by sa dal odstrániť napríklad lepším návrhom slovníka.

4.3 Systém stematizácie slovenských slov

Systém stematizácie slovenských slov nemusí vytvoriť presný koreň slova, ale mal by pre všetky možné tvary daného slova generovať rovnaký stem (koreň).

4.3.1 Testovacie sady

Pre otestovanie systému boli vytvorené 4 testovacie sady, jedna sada pre podstatné mená, jedna pre prídavné mená, jedna sada pre príslovky a jedna sada pre slovesá.

1. **Podstatné mená:** sú prevažne tvorené príponami, ale v slovenskom jazyku sa vyskytujú aj podstatné mená tvorené predponami, napríklad slovo *pradedko*. Nakoľko náš systém nedokáže rozlíšiť kedy sa jedná o predponu, napríklad pri slovách *pradedko*, *Praha*, tak sme sa rozhodli odstraňovanie predpôn pri podstatných menách vynechať. Pri podstatných mená vznikajú alternácie, kde sa mení napríklad *ch* na *š* alebo *h* na *ž* (*breh – briežok*). Testovacia sada obsahuje 4000 slov a dosahuje úspešnosť 76,52%.

počet slov	náš systém
4000	76,52%

Tabuľka 4.3.1.1: Výsledok úspešnosti pri testovacej sade podstatných mien

2. **Prídavné mená:** sú tak isto ako podstatné mená tvorené väčšinou príponami. Kvôli komplikáciám s predponami, ktoré sa pri niektorých slovách nedali rozlíšiť, sme sa rozhodli ich vynechať. Testovacia sada obsahuje 1316 slov a dosahuje úspešnosť 81,23%.

počet slov	náš systém
1316	81,23%

Tabuľka 4.3.1.2: Výsledok úspešnosti pri testovacej sade prídavných mien

3. **Príslovky:** testovacia sada prísloviek obsahuje 1251 slov a úspešnosť dosahuje úspešnosť 85,85%.

počet slov	náš systém
1251	85,85%

Tabuľka 4.3.1.3: Výsledok úspešnosti pri testovacej sade prísloviek

4. **Slovesá:** sú tvorené príponami aj predponami. Nakoľko pri podstatných aj prídavných menách sme predpony neodstraňovali, tak sme sa to rozhodli aj pri slovesách z dôvodu, že pri odvodených slovách by nám nevznikal ten istý tvar slova, napríklad pri slovách *zapracovať*, *zapracovaný* by nám vzniklo *zaprac* a *prac*. Testovacia sada obsahuje 1180 slov a dosahuje úspešnosť 76,35%.

počet slov	náš systém
1180	76,35%

Tabuľka 4.3.1.4: Výsledok úspešnosti pri testovacej sade slovies

4.3.2 Chyby

Pri hodnotení systému sme zistili, že nami vytvorený systém nie je stopercentný. Najčastejšie chyby, ktorých sa systém dopúšťal sú, že systém odtrhol buď príliš veľkú časť slova alebo naopak príliš malú časť slova. Menej častými chybami bolo to, že systém niektoré dvojslabičné slová vyhodnotil ako jednoslabičné. Dôvodom tejto chyby je to, že systém rozpoznáva slabiku na základe prvej spoluhlásky, ktorej predchádza samohláska, napríklad slovo *krtko* systém vyhodnotí ako jednoslabičné slovo.

5 Záver

Zadaním tejto bakalárskej práce bolo zoznámiť sa s metódami pre morfologickú analýzu a lematizáciu v jazykoch s bohatou morfológiou, s dostupnými slovníkmi a metódami pre výkonné spracovanie textov, napríklad pre indexovanie.

V tejto práci je popísaný spôsob tvorby systému pre lematizáciu slov, slovník, s ktorým systém pracuje, vstupné dáta, na záver bola vyhodnotená úspešnosť tohto systému a systém bol porovnaný s iným systémom. Ďalej je popísaný systém stematizácie slovenských slov, súbory, s ktorými pracuje a na záver vyhodnotená úspešnosť systému

V rámci tejto bakalárskej práce bol vytvorený systém pre lematizáciu slov, ktorý bol v závere porovnaný s už existujúcim morfologickým analyzátorom Majka, ktorý tiež využíva metódu lematizácie. Pri vytváraní systému bolo vytvorených niekoľko pomocných skriptov, ktoré slúžili na úpravu dát v slovníku. Celý systém je napísaný v jazyku Python. Systém pozostáva z niekoľkých fáz: načítanie slova, načítanie slovníka, prehľadávanie slovníka, lematizácia, výber najpravdepodobnejšej lemy, priradenie vzoru, skloňovanie a výpis výsledku.

Pri hodnotení výsledkov z prvého testu, ktorý obsahoval články z časopisu SME, bol náš systém lepší o 3,51%, pri druhom teste, ktorého obsahom boli odborné články, bol náš systém lepší o 0,77% a pri treťom teste, ktorý obsahoval úryvok z knihy, bol náš systém lepší o 0,97%. Nakoniec pri vypočítaní celkovej priemernej úspešnosti bol náš systém v porovnaní s morfologickým analyzátorom Majka lepší o 1,75%. Náš systém dokázal správne určiť väčší počet slov ako morfologický analyzátor Majka. Dôvodom väčšej úspešnosti je to, že náš systém lematizácie slov dokáže lematizovať aj neznáme slová, ktoré sa nenachádzajú v slovníku. Morfologický analyzátor Majka nedokázal určiť prídavné mená podľa vzorov otcov a matkin, skratky a v niekoľkých prípadoch mal problém určiť mená, prípadne priezviská. Týmto nedostatkom bolo to, že morfologický analyzátor Majka sa nesnaží lematizovať neznáme slová. V takomto prípade by sa mohol použiť odhadovač vzorov (napríklad, ktorým disponuje VUT FIT, súbor *lntrf2lpn.py*) pre slová, ktoré Majka nemá v slovníku. Na základe získaného vzoru by sa dal určiť základný tvar slova.

Vytvorený systém lematizácie slov bol v porovnaní s morfologickým analyzátorom Majka výrazne pomalší. Tento problém je spôsobený tým, že slovník, s ktorým systém pracuje je obyčajný textový súbor. Výraznému zrýchleniu lematizácie slov by pomohlo uložiť slovník do iného formátu, prípadne navrhnúť inú štruktúru obsahu slovníka. Taktiež k urýchleniu systému by pomohlo navrhnúť rýchlejší spôsob prehľadávania slovníka, napríklad vynechať niektoré časti slovníka, v ktorých sa nebude hľadať.

Nami vytvorený systém lematizácie slov, ktorý dokáže lematizovať aj slová, ktoré sa v slovníku nenachádzajú a generovať vyskloňované tvary, by mohol byť po menšej úprave nadstavbou morfologického analyzátora Majka. Týmto spôsobom by sme docielili ešte väčšej úspešnosti, pretože

morfologický analyzátor Majka disponuje slovníkom, v ktorom sa nachádza viac slovných tvarov než v našom slovníku pre náš systém lematizácie slov a náš systém dokáže lematizovať slová, s ktorými mal morfologický analyzátor Majka problémy.

Ďalším systémom, ktorý bol vytvorený v tejto bakalárskej práci, bol systém stematizácie slovenských slov. Celý systém je napísaný v jazyku Python a používa niekoľko súborov, v ktorých sú umiestnené konkrétne prípony alebo predpony daných slovných druhov. Systém stematizácie slov sa skladá z niekoľkých fáz: načítanie slova, rozdelenie slova na dva úseky, hľadanie prípon a predpôn v zoznamoch prípon a predpôn jednotlivých slovných druhov, odstránenie prípony a predpony, odstránenie koncových samohlások, kontrola nepravidelného tvaru a výpis.

Pri hodnotení systému, ktoré pozostávalo zo štyroch testovacích sád, sme vypočítali úspešnosť pri jednotlivých slovných druhoch. Pri podstatných menách sme dosiahli úspešnosť 76,52%, pri prídavných menách 81,23%, pri slovesách 76,35% a pri príslovkách 85,85%.

Nami vytvorený systém stematizácie slovenských slov dokáže generovať stemy slov, taktiež dokáže generovať jeden tvar pre odvodené slová. Po menšej úprave by sa mohol systém začleniť do systému Snowball, ktorý dokáže stamatizovať slová niekoľkých jazykov, ale zatiaľ sa slovenčina medzi nimi nenachádza.

Literatúra

[1] HORECKÝ, Ján. *Onomaziologická štruktúra slovenčiny*. Spisy SJS, 2003. ISBN 80-967574-9-0.

[2] Kolektív autorov. *Dolovanie znalostí z textov.* Košice: Equilibria, 2010. ISBN 978-80-89284-62-7.

[3] Kolektív autorov. *Morfológia slovenského jazyka*. Bratislava: Vydavateľstvo Slovenskej akadémie vied, 1966. ISBN 71-024-66.

[4] Kolektív autorov. *Praktická mluvnice nemčiny.* Praha: Státní pedagogické nakladatelství, 1983. ISBN 14-195-83.

[5] MISADOVÁ, Katarína. *Kapitoly z morfológie maďarského jazyka.* [online]. Bratislava: Univerzita Komenského, 2011. ISBN 978-80-223-2984-2.
Dostupné z: https://fphil.uniba.sk/fileadmin/fif/katedry_pracoviska/kmjl/Misad_Kapitoly_z_morfologie_madarskeho_jazyka.pdf

[6] NÁBĚLKOVÁ, Mira. *Vzťahové adjektíva v slovenčine.* Bratislava: Vydavateľstvo Slovenskej akadémie vied, 1993. ISBN 80.224-0352-0.

[7] NOVOTŇÁK, Matúš. *Bohovia z Alta.* [online], 2015. ISBN 978-80-8171-070-4.
Dostupné z: http://knihy.rs-design.sk/assets/Knihy/html/bohovia-z-alta.html

[8] ORAVEC, Ján. *Slovenské predložky v praxi.* Bratislava: Slovenské pedagogické nakladateľstvo, 1968.

[9] PÁLEŠ, Emil. *Parafrázovač slovenčiny.* Bratislava: Vydavateľstvo Slovenskej akadémie vied, 1994. ISBN 80-224-0109-9.

[10] RIPKA, Ivor; IMRICHOVÁ, Mária. *Základy slovenskej lexikológie.* Prešov, 2003. ISBN 80-8068-207-0.

[11] SABOL, Filip. *Nové vymedzenie čísloviek.* Slovenská reč [online]. 1993, roč.58, č.5.
Dostupné z: http://www.juls.savba.sk/ediela/sr/1993/5/sr1993-5-lq.pdf

[12] SCHWARZ, Jozef. *Současný stav a trendy automatické indexace dokumentů* [online]. Praha, 2002.
Dostupné z: http://full.nkp.cz/nkdb/docs/studie/typy.html

[13] ŠIKRA, Juraj. *Sémantika slovenských prísloviek*. Bratislava: Veda, 1991. ISBN 80-224-0322-9.

[14] WEISHEITELOVÁ, Jana. *K některým problémům automatické morfologické analýzy a lemmatizace.* [online], 2011.
Dostupné z: http://sas.ujc.cas.cz/archiv.php?art=2413

[15] MRÁZKOVÁ, Zuzana. *Slovotvorba v učebnicích českého jazyka pro střední školy.* Olomouc, 2012. Bakalárska práca. Univerzita Palackého v Olomouvi, Filozofická fakulta. Vedúci práce Božena Bednářiková.
Dostupné z:
http://theses.cz/id/rqthyo/SLOVOTVORBA_V_UEBNICCH_ESKHO_JAZYKA_PRO_STEDN_KOLY.pdf

[16] SEDLÁČEK, Radek. *Morfologický analyzátor čestiny.* Brno, 1999. Diplomová práca. Masarykova univerzita, Fakulta informatiky. Vedúci práce Pavel Rychlý.
Dostupné z: https://nlp.fi.muni.cz/projekty/ajka/ajka.pdf

[17] web: Automatická morfologická analýza a strojový slovník češtiny.
Dostupné z:
https://digilib.phil.muni.cz/bitstream/handle/11222.digilib/124477/SpisyFF_401-2011-1_5.pdf?sequence=1

[18] web: Český pedagogický tezaurus.
Dostupné z: http://www.npmk.cz/knihovna/cesky-pedagogicky-tezaurus

[19] web: Deduplikácia kmeňových dát.
Dostupné z: http://www.softec.sk/files/clanky/28-29-paul-infoware.pdf

[20] web: Informační technologie a korpusová ligvistika.
Dostupné z: http://webserver.ics.muni.cz/bulletin/articles/67.html

[21] web: Majka.
Dostupné z: https://nlp.fi.muni.cz/ma/

[22] web: Morfologický slovník a morfologický analyzátor pro češtinu.
Dostupné z:
https://knot.fit.vutbr.cz/wiki/index.php/Morfologick%C3%BD_slovn%C3%ADk_a_morfologick%C3%BD_analyz%C3%A1tor_pro_%C4%8De%C5%A1tinu#k2_-_p.C5.99.C3.ADdavn.C3.A9_jm.C3.A9no

[23] web: MorphoDiTa
Dostupné z: http://ufal.mff.cuni.cz/morphodita

[24] web: SME.
Dostupné z: http://www.sme.sk/

[25] web: Snowball.
Dostupné z: http://snowball.tartarus.org/texts/introduction.html

[26] wikipedia: Automatická indexace.
Dostupné z: https://cs.wikipedia.org/wiki/Automatick%C3%A1_indexace

[27] wikipedia: Morfologické značky.
Dostupné z: https://wiki.korpus.cz/doku.php/seznamy:tagy

Zoznam príloh

Manuál

V priloženom CD sa nachádzajú dva programy: program pre lematizáciu slovenských slov a program pre stematizáciu slovenských slov.

1. Program na lematizáciu slovenských slov

Tento program sa nachádza v priečinku *lemmatizator*. Tento priečinok obsahuje 2 súbory:

1. *lemmatizator.py* – spúšťací skript
2. *ma.txt* – slovník, s ktorým program pracuje

Spustenie programu:
Program sa spúšťa zadaním jedeného príkazu v termináli: *python3 lemmatizator.py*

Po spustení skriptu sa objaví hlavička:

```
Lemmatizátor
Author: Šimon Lipták, xlipta02@stud.fit.vutbr.cz
*********************************************/
```

Po zobrazení tejto hlavičky zadáme slovo, z ktorého program vytvorí základný tvar a vypíše ostatné tvary slova.

Ukážka príkladu z programu:

```
chlapom
Základný tvar: chlap
Ostatné tvary:
chlapi:k1gMnPc1
chlapov:k1gMnPc24
chlapom:k1gMnPc3
chlapoch:k1gMnPc6
chlapmi:k1gMnPc7
chlap:k1gMnSc1
chlapa:k1gMnSc24
chlapovi:k1gMnSc36
chlapom:k1gMnSc7
```

Program ukončíme pomocou klávesovej skratky *Ctrl + C*.

2. Program na stemizáciu slovenských slov

Tento program sa nachádza v priečinku *stemmer*. Tento priečinok obsahuje 9 súborov:

1. *stemmer.py* – spúšťací skript
2. *nepravidelnosti.txt* – zoznam slov s nepravidelným skloňovaním alebo časovaním
3. *stop_words.txt* – zoznam „stop“ slov
4. *slovot_prip_pod_m.txt* – zoznam prípon podstatných mien
5. *pod_konc.txt* – zoznam pádových koncoviek podstatných mien
6. *slovot_prip_prid_m.txt* – zoznam prípon prídavných mien
7. *prid_konc.txt* – zoznam pádových koncoviek prídavných mien
8. *slovesa_konc.txt* – zoznam prípon slovies
9. *prisl_konc.txt* – zoznam prípon prísloviek

Spustenie programu:
Program sa spúšťa zadaním jedeného príkazu v termináli: *python3 stemmer.py*

Po spustení skriptu sa objaví hlavička:

```
Stemmer
Author: Šimon Lipták, xlipta02@stud.fit.vutbr.cz
*********************************************/
```

Po zobrazení tejto hlavičky zadáme slovo, z ktorého program vytvorí stem slova.

Ukážka príkladu z programu:

```
spracovať
sprac
spracovávať
sprac
spracovaný
sprac
spracovanie
sprac
spracovane
sprac
```

Program ukončíme pomocou klávesovej skratky *Ctrl + C*.