Katedra Informatiky
Fakulta Matematiky, Fyziky a Informatiky
Univerzita Komenského, Bratislava

Podobnosť slov

(Diplomová práca)

Martin Vlčák

Vedúci: RNDr. Michal Forišek Phd.

Bratislava, 2011

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Martin Vlčák
Študijný program: informatika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: 9.2.1. informatika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský

Názov : Podobnosť slov

Cieľ : Cieľom práce je preskúmať možnosti a aplikácie definovania významovej podobnosti slov na základe štatistických ukazovateľov, ako je napríklad počet webstránok nájdených vyhľadávačom, resp. počet a miesta výskytov v korpuse jazyka.

Vedúci : RNDr. Michal Forišek, PhD.

Dátum zadania: 13.11.2009

Dátum schválenia: 18.02.2011

prof. RNDr. Branislav Rovan, PhD.
garant študijného programu

student

vedúci práce

Dátum potvrdenia finálnej verzie práce, súhlas s jej odovzdaním (vrátane spôsobu sprístupnenia)

5.5.2011 Forišek

vedúci práce

Martin Vlčák

Podobnosť slov

Diplomová práca

Katedra informatiky
Fakulta matematiky, fyziky a informatiky
Univerzita Komenského, Bratislava

2011

Čestne vyhlasujem, že som túto diplomovú prácu vypracoval samostatne,
s použitím uvedených zdrojov.

. . . . . .

Abstrakt

Autor: Martin Vlčák
Názov diplomovej práce: Podobnosť slov
Škola: Univerzita Komenského
Fakulta: Fakulta Matematiky, Fyziky a informatiky
Katedra: Katedra informatiky
Vedúci diplomovej práce: Michal Forišek

Táto práca skúma zisťovanie sémantickej podobnosti slov, na základe výsledkov www vyhľadávačov. Obsah www stránok je tvorený ľudskými používateľmi a preto je veľmi dobrým zdrojom výrazov, na základe ktorých, možno určiť sémantickú podobnosť slov. Práca sa venuje hľadaniu vhodnej metódy na vyjadrenie podobnosti slov a spôsobu ktorým možno na základe nameranej podobnosti slov, zaradiť slovo na jeho miesto do korpusu jazyka.

Kľúčové slová: Podobnosť slov, semantika, www, vyhľadávač, kontext, metrika, vzdialenosť

Obsah

Kapitola 1

Úvod

Určovanie podobnosti slov, je už dlhodobo výzvou pri spracovaní prirodzeného jazyka. V mnohých prípadoch, je pre človeka veľmi jednoduché z reči, alebo písaného textu určiť význam jednotlivých slov. Na druhej strane, je tento problém náročný z hľadiska vypočítateľnosti. Pri spracovaní prirodzeného jazyka, je potrebné poznať jednotlivé slová, ich význam a tiež skúmať súvislosti medzi slovami.

Spracovanie slov prirodzeného jazyka možno pre menšiu náročnosť výpočtu rozdeliť na viacero podúloh. Jednou z nich je vedieť označkovať jednotlivé slová tak, aby niesli aj pridanú informáciu, ktorá uľahčí ich spracovanie. V prípade Slovenského jazyka, by mohlo byť cieľom tejto úlohy vedieť okrem iného pre ľubovoľné podstatné meno vedieť povedať jeho rod, číslo a pád.

Problematikou skúmania podobnosti v slovenčine sa už zaoberal Emil Páleš vo svojej kandidátskej. Vytvoril model na skúmanie vlastností slovenského jazyka a tiež sa zaoberal problémom nejasného významu jednotlivých slov jazyka. Jeho práca slúži zatiaľ ako jediný známy komplexný výpočtový model na určovanie podobnosti v slovenskom jazyku.

Inou úlohou pri spracovaní prirodzeného jazyka je vedieť zadeliť slová do skupín. To znamená, vedieť vytvoriť množiny slov - kontexty, pričom slová v každej množine majú určité spoločné vlastnosti. Ďalšou úlohou spracovania jazyka je určovanie sémantickej podobnosti medzi ľubovoľnými dvoma slovami jazyka. To znamená vedieť presne popísať reláciu medzi dvoma slovami a zaviesť metriku na základe ktorej možno kvantifikovať podobnosť slov.

Existujúce systémy riešiace problém spracovania prirodzeného jazyka, riešia

tieto spomínané úlohy oddelene, pričom po ich vyriešení sa výsledky spájajú do uceleného celku, ktorý umožňuje poskytovať informácie o jazyku.

V našej diplomovej práci sa zaoberáme sémantickou podobnosťou slov. Skúmame, ktoré slová z prirodzeného jazyka sa môžu vyskytovať v rovnakom kontexte a skúmame mieru podobnosti určenú na základe výsledkov www vyhľadávača. Ešte predtým než začneme skúmať podobnosť slov predstavíme počítačovú lingvistiku ako ucelenú vednú disciplínu v oblasti informatiky a prezentujeme niekoľko zaujímavých výsledkov dosiahnutých v tejto oblasti.

1.1 Počítačová lingvistika

Emil Páleš vo svojej práci definuje počítačovú lingvistiku ako interdisciplinárnu vedu na rozhraní kybernetiky, jazykovedy a kognitívnej psychológie, ktorej cieľom je vytvoriť formálny model jazykového vedomia.[Pal94]

Výpočtová lingvistika sa radí pod oblasť umelej inteligencie. Umelá inteligencia je veda, ktorá študuje inteligentné správanie živých organizmov. Zaoberá sa hľadaním hraníc a možností symbolickej, znakovej reprezentácie poznatkov a procesov ich nadobúdania, udržovania a využívania.

Využiteľnou prednosťou počítača je jeho schopnosť spracúvať veľké kvantá informácie, aké by jazykovedec-človek nikdy nemohol obsiahnuť. Jazykový prejav je fenomén taký zložitý, živý a rozmanitý, že žiaden ľudský subjekt nemôže formulovať svoje hypotézy o jazyku pri plnom vedomí všetkých vzťahov, ktoré môžu s jeho objektom záujmu súvisieť a ovplyvniť ho. U človeka vždy musí nastať zúženie vedomia koncentráciou. To znamená, že ak sa výskumník sústredí na niektorú podoblasť, ostatné oblasti musí čiastočne pustiť z hlavy. Tieto upadajú do mierneho polotieňa a vzťahy v nich sa musia v tej chvíli chápať zjednodušene, schematicky. Jedine stroj si dokáže naraz pamätať a brať do úvahy všetky zákonitosti všetkých podsystémov jazyka a testovať ich správanie v plnej vzájomnej interakcii. Len počítač dokáže overiť každú hypotézu na mnohomiliónových korpusoch a oslobodiť jazykovedca od sugescie špeciálneho príkladu. [Pal94]

Každý model rečovej komunikácie je zároveň modelom ľudskej mysle. Pri tvorbe kognitívno-lingvistického modelu sa nevyhneme otázke, ako súvisí jazyk a reč s myslením, cítením, vnímaním a konaním človeka. Počítačová lingvistika

nám teda v neposlednom rade pomáha - pochopiť nás samých ! Každý, aj neúspešný experiment v tomto smere prispieva k prehodnoteniu našich postojov a preformulovaniu otázok zmysluplnejším spôsobom. [Pal94]

1.2 Zaujímavé systémy na prácu s jazykom

Počiatky práce s jazykom v informatike siahajú veľmi hlboko. Už v 60. rokoch minulého storočia vznikol na Masachussettskom technologickom inštitúte program zvaný Eliza. Bol to program, ktorý bol schopný komunikovať s človekom formou otázok a odpovedí. Bol schopný komunikovať s človekom tak prirodzene, že sa mnohí odborníci dlhšiu dobu domnievali, že tento program uspel v Turingovom teste inteligencie.

Iný nástro,j ktorý bol schopný komunikovať s človekom na vysokej úrovni predstavil Emil Páleš. Tento stroj nazval papagájovač. Princíp je taký, že papagájovač a človek sa striedajú vždy po jednej fráze v komunikácií. Papagájovač pritom zovšeobecňuje typy otázok a odpovedí a spájal ich asociatívnymi väzbami. Snaží sa zistiť, akým spôsobom sa reaguje na rozličné druhy viet.

Ak narazí na nový typ otázky, odpovedá neviem a otázku si zapamätá. Po nejakom čase sa opýta podobným spôsobom a zapamätá si, akým spôsobom na ňu človek odpovedal. Na oznamovacie vety najprv odpovedá "to je zaujímavéä neskôr pri nejakej príležitosti iniciuje rozhovor analogickým spôsobom. Opäť si zapamätá, ako sa na takú vetu reaguje

1.3 Podobnosť slov a webové stránky

Pretože množstvo www stránok sa neustále rozrastá, je jednou zo základných vlastností sémantického webu vedieť rozdeľovať slová do skupín. Napomáha to jednoduchšiemu a hlavne rýchlejšiemu vyhľadávaniu jednotlivých fráz na internete. Pre človeka je veľmi užitočné, aby bol obsah na internete rozkategorizovaný. V súčasnosti vyhľadávacie stroje podporujú do určitej miery kategorizáciu obsahu. Momentálne najväčšie vyhľadávače kategorizujú obsah pomocou značkovania a indexovania stránok a ich obsahu.

Značky na jednotlivých www stránkach sú jedným možným spôsobom, ako možno rozdeľovať obsah na stránkach do skupín. Slúžia prevažne pre vyhľadávacie stroje a na ich základe ľudský používateľ dostáva finálnu stránku, bez potreby poznať tieto značky.

Značkovaním obsahu nemožno riešiť problém viacvýznamovosti obsahu internetu. Momentálne najväčší vyhľadávač google síce nápovedy ponúka, ale tie málokedy umožňujú nejaké presné sémantické utriedenie. Väčšina ľudí pri vyhľadávaní zadáva 1 až 2 kľúčové slová. Stroj google následne len vyberie najviac zodpovedajúci obsah danej fráze. Zadaný počet slov teda vo väčšine prípadov nepostačuje na to, aby bolo možné jednoznačne zaradiť hľadanú frázu do nejakej sémantickej skupiny.

Obsah možno kategorizovať tak, že každý autor www stránky musí na nej označkovať jednotlivé časti obsahu, poprípade označiť celú stránku, čo konkrétne obsahuje. Ale keďže značky na stránky sú pridávané tiež len človekom, môžu sa v nich tiež vyskytovať nepresnosti. Pre človeka intuitívnejší spôsob získavania obsahu ako je značkovanie stránok je prezeranie obsahu vlastnými očami. Internet postupne nahrádza ostatné zdroje informáciami astáva sa najväčším zdrojov textov.

Do príchodu internetu boli najväčším zdrojom informácií pre človeka knihy. Príchodom internetu sa ale mnohé zmenilo. Internetové stránky ponúkajú nepreberné množstvo obsahu, či už v podobe elektronických kníh, článkov, alebo diskusií. Všetky tieto informácie na www sieti majú ale prvotného pôvodcu. Človeka. Preto www stránky predstavujú obrovskú databázu slov prirodzeného jazyka.

Na www stránkach sa neustále obajvujú nové slovné tvary a nové slovné spojenia. Každý jeden príspevok pridaný na internet jeho používateľom nám napovedá viacej o štruktúre jazyka. Preto sme sa rozhodli v tejto práci venovať sa sémantickej podobnosti slov na základe www vyhľadávačov. Postupne budeme v práci uvažovať nad rôznymi spôsobmi určovania podobnosti slov na základe ich výskytu na www stránkach.

1.4 Motivácia pre určovanie sémantickej podobnosti slov

Určovanie sémanitckej podobnosti je legitímny problém počítačovej lingvistiky. Automatické určenie sémantickej podobnosti znamená, že počítač bude vedieť z textu odhaliť v akom význame sa dané slovo používa. Určenie významu slova posunie počitače o výrazný krok bližšie k pochopeniu ľudskej reči.

Význam sémantickej podobnosti slov je zjavný hlavne na príkladoch z oblasti umelej inteligencie. V oblasti umelej inteligencie je snaha o to, aby stroje dokázali pochopiť jednotlivé príkazy podávané v ľudskom jazyku. Na to, aby to bolo možné, potrebujú vedieť poznať kotext v ktorom sa práve daný príkaz vyskytuje. To sa dá docieliť práve pomocou rozoznávania sémantického významu.

Ďalšou motiváciou pre poznávanie sémantickej podobnosti je možnosť vyhľadávania so sémantickým nápovedami, kedy si uživateľ bude môcť zvoliť kontext v ktorom chce vyhľadávať. Inou aplikáciou je tiež kontextový prekladač. Je to prekladač textov, ktorý by vedel na základe výskytu slov v texte vybrať zo slovníka pri jeho preklade správny ekvivalent.

Kapitola 2

Problematika

V tejto kapitole si predstavíme spôsoby na určovanie podobnosti slov a na meranie podobnosti slov. Na začiatku tejto kapitoly je potrebné formulovať problém podobnosti slov. Na jednej strane je potrebné vedieť povedať o dvoch slovách či sú navzájom podobné a na strane druhej je potrebné zmerať ich podobnosť. Na začiatku tejto kapitolu uvedieme, že v práci často používame pojem metrika a vzdialenosť slov. Treba upozorniť, že nie vždy sa musí nutne jednať o metriku a vzdialenosť v zmysle korektnej matematickej definície. Toto názvoslovie používame preto, že je zaužívané aj v iných článkoch publikovaných v tejto oblasti. Na začiatok tejto kapitoly uvedieme definície, ktoré čitateľa lepšie uvedú do problematiky sémantickej podobnosti.

Definícia 2.0.1 *Koncept je množina súvislých n-týt, ktoré spolu úzko súvisia.*

Definícia 2.0.2 *Binárna relácia byť(sémanticky) podobný je relácia, ktorá slová z prirodzeného jazyka prehlási za podobné ak sú podobné aj v ľudskej reči.*

Keď sa pozrieme na podobnosť ako na operáciu "byť podobný", je dobré si uvedomiť, že táto operácia musí spĺňať isté vlastnosti. Podobný v tomto zmysle znamená častý výskyt slov v prirodzenom jazyku v tom istom kontexte. Od operácie podobnosti očakávame komutatívnu vlastnosť. Teda ak prehlásime jedno slovo za podobné z druhým, očakávame, že to platí aj naopak.

Ako teda dosiahnuť výslednú operáciu podobnosti ? Ako prehlásiť slová za sémanticky podobné ? Toto sa realizuje použitím meraní podobností slov. Po

zmeraní podobnosti je možné nastaviť prahovú hodnotu na základe ktorej prehlásime slová za podobné. V nasledujúcich častiach tejto kapitoly, si postupne predstavíme existujúce prístupy, ktorými možno merať podobnosť slov a určiť, či sú si slová podobné.

2.1 Určenie podobnosti slov pomocou www vyhľadávača

S narastajúcim počtom www stránok a ich obsahu možno stále čoraz presnejšie určovať podobnosť slov pomocou vyhľadávacích strojov. Je dobré si uvedomiť, že ak chceme určovať podobnosť slov, je potrebné mať k dispozícií slovník slov daného jazyka a tiež súvislé texty, v ktorých sa dané slová vyskytujú v rôznom kontexte. V súčasnosti www sieť spĺňa obe tieto podmienky. Vďaka tomu je možné pomocou výsledkov vyhľadávacích strojov určovať vzdialenosti medzi jednotlivými slovami jazyka. Nevýhodou určovania podobnosti slov pomocou www vyhľadávačov je fakt, že tieto techniky sú silne závislé od pripojenia na internet, ako aj to, že nedovoľujú zisťovať priamo podobnosť celých textov. Môže sa stať teda, že pri použití tohto prístupu, bude okrem zložitosti na porovnanie podobnosti slov nezanedbateľne vysoký aj čas potrebný na odoslanie na vyhľadávací server a takisto prijatie odpovede.

Vzdialenosť medzi slovami možno pomocou www vyhľadávačov určovať viacerými spôsobmi. Jeden z prvých prístupov ako možno z výsledkov www vyhľadávača vyjadriť podobnosť slov predstavili vo svojej práci Rudi L. Cilibrasi a Paul M.B. Vitányi. [CV04] Keďže naša diplomová práca sa do veľkej miery opiera o podobnosť založenú na www vyhľadávačoch, budeme sa jej podrobnejšie venovať v nasledujúcej kapitole. V nasledujúcich sekciách tejto kapitoly opíšeme iné metódy pre prácu so sémantickou podobnosťou. K týmto metódam sa následne vrátime v závere, pri porovnávaní s našimi výsledkami.

2.2 WordNet

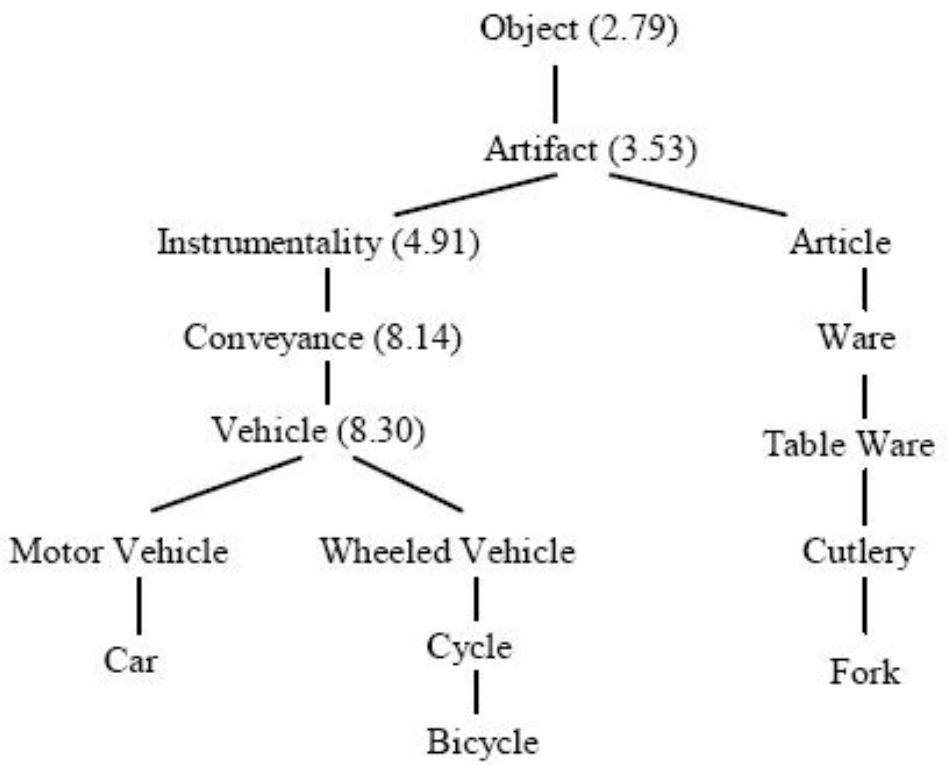
Wordnet je robustná databáza slov v anglickom jazyku, vytvorená George A. Millerom a jeho spolupracovníkmi. Účelom wordnetu je kombinovať obyčajný slovník jazyka, ale tiež aj grupovať slová jazyka podľa ich významu. Wordnet umožňuje rozlišovať medzi štyrmi slovnými druhmi. Rozlišuje podstatné mená, prídavné mená, slovesá a príslovky.

Takéto členenie umožňuje lepšie vyjadriť význam jednotlivých slovných druhov. Wordnet umožňuje viaceré relácie medzi jednotlivými slovami. Možno definovať či sa slovo nachádza v skupine, ktorá môže byť určená iným slovom. Wordnet je reprezentovaný taxonomickým n-árnym stromom. Slová jazyka predstavujú jednotlivé vrcholy stromu.

Kontext v prirodzenom jazyku určuje v akom zmysle sú dané slová chápané (vo všeobecnosti možno kontext chápať ako konkrétny pohľad na nejakú špeciálnu oblasť) V sieti Wordnet je kontext reprezentovaný podstromom nejakého vrcholu v taxonomickom strome. Na základe taxonómie určenej wordnetom vznikali následne ďalšie štúdie poukazujúce na možné zisťovanie sémantickej podobnosti a porovnávaní textov na základe sémenatiky. [Fel98]

Medzi hlavné prínosy WordNetu možno zaradiť to, že umožňuje celkom efektívne vyhľadávať slová z jazyka podľa ich významu. Taxonomický strom wordnetu je v súčasnosti reprezentovaný databázou, pričom sú implementované viaceré užívateľské rozhrania umožňujúce užívateľom jednoduché vyhľadávanie slov v konceptoch.

Je snaha tiež prepájať zásobu slov WordNetu so sémantickým webom. Takýmto prepojením by bolo možné vyhľadávať medzi slovami na webe aj na základe sémantiky, pretože každé slovo má v štruktúre WordNetu svoju pozíciu, ktorá presne definuje jeho význam. V tejto práci ukazujeme iný spôsob založený na vytvorení grafových clusterov reprezentujúci jednotlivé koncepty. Nevýhodou používania wordnetu na zisťovanie podobnosti je jeho statickosť. Slová prirodzeného jazyka sa totiž neustále menia a wordnet nezaručuje ich aktuálnosť. Na druhej strane veľkou výhodou je efektívne prehľadávanie stromov, ktoré umožňuje rýchlu prácu so slovnou zásobou uloženou v tejto štruktúre.



Obr. 2.1: Ukážka časti štruktúry WordNetu

2.3 Metriky podobnosti založené na Wordnet a analýze textov

V predchádzajúcej sekcií sme predstavili sieť slov Wordnet, vytvorenú za účelom kategorizácie slov. V tejto časti predstavíme existujúce metriky na meranie podobnosti. Tieto metriky predpokladajú existenciu slovníka, veľkého množstva textu, alebo sa priamo opierajú o taxonomickú štruktúru wordnet.

Jednou z metrík, ktorá sa opiera priamo o wordnet štruktúru je metrika predstavená Wu a Palmerom. Táto metrika meria vzdialenosť konceptov. Je založená na hĺbke konceptov v taxonomickom strome

$$\frac{Sim = 2 * depth(LCS)}{depth(concept_1) + depth(concept_2)}$$

Kde $depth(LCS)$ je hĺbka najbližšieho spoločného uzla v taxonomickom strome a v menovateli je hĺbka porovnávaných konceptov. Viacej o tejto metrika sa môže

čitateľ dočítať v [WP94]

Iný spôsob merania podobnosti konceptov tiež založený na taxonomickej štruktúre predstavil vo svojej práci Resnik. Ten navyše predpokladá existenciu obsiahlejšieho korpusu jazyka. Vo svojej práci predstavil informáciu o obsahu textu, ktorú definoval nasledovne:

$$IC(c) = -\log P(c)$$

$P(c)$ predstavuje pravdepodobnosť výskytu konceptu c v obsiahlom korpuse. Následne metriku podobnosti definoval ako :

$$Sim = IC(LCS)$$

Postupne na základe tejto práce vznikali iné, ktoré rôznym spôsobom upravovali túto prezentovanú metriku. Viacej o takýchto metrikách sa čitateľ môže dočítať v [JC97], [Res95], [CM05].

2.4 Iné spôsoby merania podobnosti slov a vlastností jazyka

Medzi ďalšie spôsoby skúmania prirodzeného jazyka patrí Latentná sémantická analýza(LSA). Pri používaní LSA sa predpokladá, že poznáme namerané hodnoty podobnosti, na základe ktorých potom prebiehajú ďalšie výpočty. LSA sme sa v našej práci nezaoberali.

Iný spôsob ako určovať podobnosť je zamerať sa na jednotlivé slovné tvary a ich možnú štruktúru (pády, rody, spojenie s predložkami) a následne na základe takéhoto modelu porovnávať celé texty. Výhodou takéhoto modelu je určite jeho komplexnosť a to, že umožňuje porovnávať priamo celé texty na základe tvarov jednotlivých slovných druhov spomenutých v texte. Jeho nevýhodou je jeho obsiahlosť. Ak chceme zadefinovať vo výpočtovom modely takúto štruktúru, tak musíme počítať s väčším časom potrebným na výpočet.

V tejto práci sa zaoberáme výpočtovým modelom menej komplexnejším, ktorý dokáže riešiť problém sémantickej podobnosti prípadne kontextu, v akom sa zvolené slovo z jazyka vyskytuje, ale nezaoberáme sa možnými tvarmi daného slova

spôsobeným skloňovaním ani jeho možným postavením vo vernej forme. O komplexnom modely sa čitateľ viacej dočíta v knihe Emila Páleša[Pal94].

2.5 Výskyt kontextu v texte

Ako sme už spomínali v úvodnej definícií tejto kapitoly, koncept môžme definovať, ako množinu n-týt, v našom prípade slov z prirodzeného jazyka. Slová patriace do jedného spoločného konceptu, možno získať viacerými spôsobmi. Jedným je zistenie konceptu z nejakej štruktúry, napríklad wordnetu. Iným použiteľným spôsobom získania konceptu, je jeho získanie z textu. Na základe ich spoločného výskytu v rôznych slovných spojeniach a frázach.

Predpokladajme, že máme n slov. Teraz môžme vyjadriť, pravdepodobnosť toho, že táto množina je skutočne kontextom nasledovne :

$$P(C) = \frac{freq(C)}{N},$$

kde N vyjadruje počet všetkých prezeraných textov, $freq(C)$ hovorí o tom, ako je kontext frekventovaný. Frekvenciu možno získať z textov spôsobom, pri ktorom si zoberieme všetky frázy obsahujúce slová z C. Takýmto spôsobom vieme popísať, aká je pravdepodobnosť súvislosti slov. Ak je dostatočná, prehlásime, že tvoria kontext. V inakšom prípade kontext tvoriť nebudú. K takto definovanej pravdepodobnosti konceptu sa vrátime ešte pri prezentácií našej metriky pre určovanie podobnosti slov.

Kapitola 3

Vzdialenosť slov

V tejto kapitole sme z okrem už zmienenej literatúry čerpali z výskumov v oblasti sémantiky, uvedených v [RTS08], [Tur05], [SH06], [mss07]

V tejto kapitole sa budeme venovať prevažne spôsobom merania podobnosti slov a možným problémom, ktoré môžu nastať pri meraní podobnosti slov. Ukazujeme tu postupne metriky, ktoré vychádzajú z výsledkov vyhľadávačov pre dané slová a postupne sa dostávame k metrike, ktorú neskôr využijeme na zostrojenie modelu podobnosti slov, ktorý v tejto práci prezentujeme.

Keďže táto práca sa pri hľadaní relácie podobnosti do veľkej miery opiera o určovanie vzdialenosti slov pomocou vyhľadávacích strojov, (konkrétne google) tak v tejto kapitole budeme väčšinou pojednávať o vzdialenosti slov určenej www vyhľadávačmi. Keďže práca bola založená na určovaní podobnosti slov pomocou www vyhľadávačov, tak považujeme za potrebné na začiatku tejto kapitoly ozrejmiť fungovanie vyhľadávacieho stroja google.

3.1 Vyhľadávací stroj google

Google je momentálne v Európe najpoužívanejším www vyhľadávačom. Počet stránok oindexovaných vyhľadávacím strojom google neustále narastá. Preto vhodným zdrojom odkiaľ čerpať výskyt slov prirodzeného jazyka, je práve z tohto www vyhľadávača.

Od www vyhľadávača pri zisťovaní podobnosti očakávame isté vlastnosti, ktoré musí zaručovať. Je to predovšetkým zvolenie jazyka, v ktorom sú napísané

hľadané slová a tiež schopnosť vedieť vyhľadávať podľa zvoleného výrazu.

Potreba voľby jazyka pri vyhľadávaní vyplýva z toho, že hľadáme sémantickú podobnosť slov konkrétneho prirodzeného jazyka. Keby sme neboli schopný zvoliť jazyk dosiahnuté výsledky by mohli byť skreslené. Ako príklad stačí zvoliť slovo zo slovenčiny les.

Význam tohto slova v slovenčine je jasný, ale keby sme nemali možnosť zvoliť si jazyk v akom chceme vyhľadávať obsah, tak okrem výsledkov v slovenčine (a samozrejme iných) by sme dostali aj výsledky výskytu tohto slova na francúzskych stránkach, kde toto slovo predstavuje veľmi frekventovaný člen. To by mohlo naše výsledky skresliť až do takej miery, že by sa stali nepoužiteľnými.

Ako teda docieliť spomínané vlastnosti ? Asi by nebolo veľmi použiteľné pri vyhľadávaní slov každé slovo zadávať na stránke google vyhľadávača a hlavne by sme tak neboli schopní v reálnom čase pracovať s už len trochu väčšou množinou slov. Riešenie je v uvedomení si, že google je svojím spôsobom len obyčajný vyhľadávací server. To nám umožňuje prenášať na neho parametre a tiež spracovávať odpoveď od servera.

Parametre sa posielajú použitím HTTP protokolu v poli GET parametrov. Jazyk, pre ktorý chceme vidieť zobrazené výsledky vieme poslať ako parameter clang_xx, pričom výraz xx predstavuje kód jazyka (pre slovenčinu sk, angličtinu en). Tiež vieme docieliť hľadanie konkrétnych slov aj s použitím logických výrazov. Stačí nastaviť q parameter poľa get na hodnotu zodpovedajúcemu výrazu.

Nakoľko nie je cieľom tejto časti úplne dopodrobna vysvetliť ako funguje vyhľadávanie na www sieti, ale len presnejšie ozrejmiť jednotlivé základné funkcie potrebné pre pochopenie nasledujúcich častí, tak viacej o vyhľadávačoch môže čitateľ nájsť napríklad v dokumentácií k vyhľadávača google na internete.

Na záver tejto časti ešte pripomenieme že GET parametre možno preniesť na server priamo v url adrese, čo umožňuje jednoduché vybudovávanie query dotazov na vyhľadávací server. Spracovávanie výsledkov z vyhľadávacieho servera ako aj ich presné odosielanie na server popíšeme v kapitole Implementácia.

3.2 Zisťovanie podobnosti na báze www vyhľadávačov

Pri zisťovaní podobnosti pomocou www vyhľadávačov vystávajú 2 základné problémy. Prvým problémom je nameranie vzdialenosti medzi slovami. Len ale ťažko možno predpokladať, že len podľa nameranej hodnoty podobnosti budeme vedieť určiť podobnosť slov. Pre človeka je veľmi ťažké označiť nejakú hodnotu za dostatočnú na vyjadrenie podobnosti slov. Od nameranej hodnoty očakávame skôr, že na základe nej je možné, prehlásiť slová za podobné v určitom kontexte.

Druhým problémom je rozhodnutie o podobnosti medzi slovami na základe nameranej vzdialenosti medzi nimi. Je vidieť, že druhý problém je do úzkej miery spojený s tým prvým. Potrebujeme vedieť na základe nameraných vzdialeností medzi slovami vedieť povedať, či sú aj podobné alebo nie. Na to je potrebné si zadefinovať pojem prahovej hodnoty a následne aj pojem dobrej metriky.

Definícia 3.2.1 *Prahovou hodnotou budeme nazývať hodnotu, pre ktorú platí, že ak je vzdialenosť medzi slovami menšia ako táto hodnota, tak slová vyhlásime za sémanticky rôzne. V opačnom prípade ich prehlásime za sémanticky podobné.*

Definícia 3.2.2 *Dobrou metrikou budeme nazývať metriku, na základe ktorej vieme určiť korektne konkrétnu prahovú hodnotu.*

Dobrou metrikou na základe spomínaných definícií nazveme takú metriku, ktorá postupne pre skupinu slov nameria také vzdialenosti, že na základe týchto výsledkov vieme korektne pri porovnaní s ľudskou rečou prehlásiť slová za podobné alebo nie (Napríklad najjednoduchšie prahovú hodnotu zvolíme ako priemer nameraných hodnôt). V nasledujúcich častiach sa pozrieme na metriky, ktoré sme používali pri testovaní sémantickej podobnosti slov a povieme niečo o ich vlastnostiach. Tiež sa pozrieme na to, v ktorých prípadoch sú tieto metriky úspešné pri zisťovaní podobnosti a naopak kedy ich použitie nie je úplne vhodné. Prahovú hodnotu sme získavali na základe vykonaných pokusov, preto sa jej budeme venovať v nasledujúcej kapitole venovanej zhrnutiu výsledkov.

3.2.1 Normalized Google Distance(NGD)

V prvej časti tejto kapitoly si predstavíme metriku, založenú na výsledkoch vrátených www vyhľadávačom. Predstavujeme tu metriku prezentovanú v [CV04]. Tiež poukazujeme na niektoré jej problémy a hľadáme riešenia pre ich odstránenie.

Táto metrika bola založená na základe kolmogorovskej zložitosti na slovách. Kompletné odvodenie tejto metriky s úplnými dôkazmi nájde čitateľ v spomínanej literatúre. NGD je teda definovaná ako

$$NGD = \frac{\max\{\log(f(x)), log(f(y))\} - \log f(x,y)}{\log N - \min\{\log(f(x)), log(f(y))\}}$$

V tejto metrike x, y predstavujú porovnávané slová. Funkcia $f(x)$ predstavuje počet výsledkov vrátených google vyhľadávačom pre dané slovo x. Funkcia f(y) počet výsledkov pre y a f(x,y) vráti počet výskytov x a y na rovnakej stránke.

Táto metrika funguje dobre pre slová s malým počtom iných kontextov. Teda pre slová, ktoré sa nevyskytujú príliš často v iných kontextoch, ako práve v kontexte v ktorom spolu súvisia. To ale nastáva v zriedkavých prípadoch, keďže obsah na internete nie je rozdelený rovnomerne. Namerané podobnosti za použitia NGD možno vidieť na obrázku 3.1

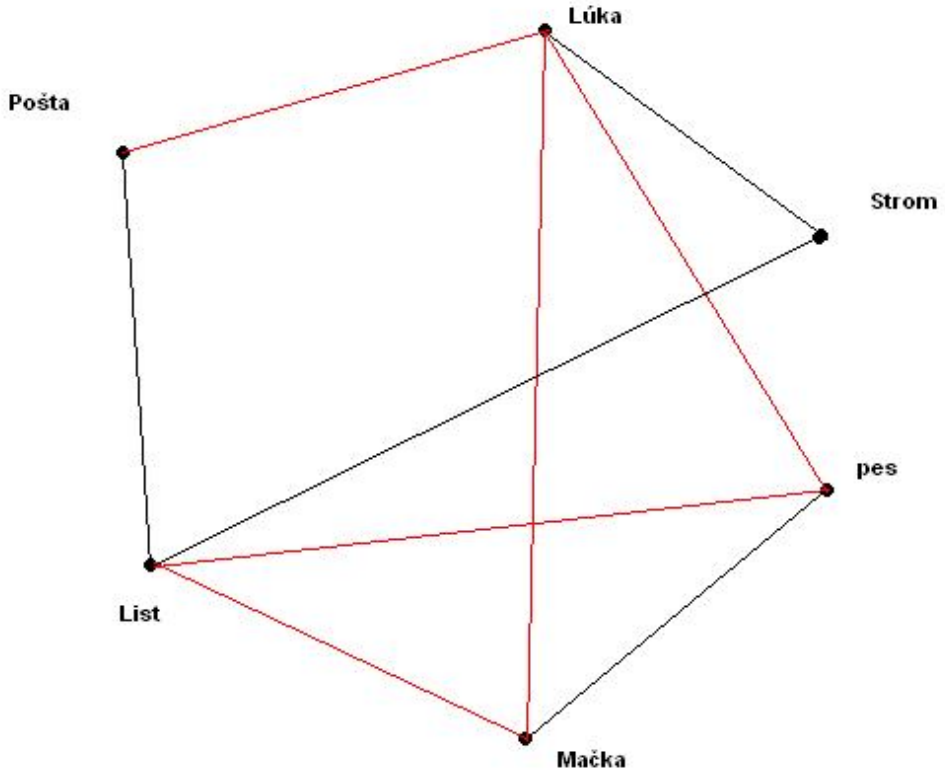
Slovo:	pošta	strom	list	les	lúka	macka
pošta	-	0.34	1.0	0.83	0.40	0.42
strom	0.34	-	0.88	0.52	0.39	0.40
list	1.0	0.88	-	1.42	1.02	0.95
les	0.71	0.57	1.42	-	0.71	0.70
lúka	0.44	0.38	1.02	0.85	-	1.0
macka	0.42	0.40	0.95	0.72	1.0	-

Obr. 3.1: Hodnoty získané použitím NGD

Existujú slová a frázy, ktoré majú väčšie zastúpenie na www stránkach a na druhej strane sú slová, ktoré majú omnoho menší výskyt. Ak sa slovo vyskytuje

na veľkom počte stránok, je viacej pravdepodobné, že na nejakej stránke sa objaví spoločne so slovom, s ktorým sémanticky obzvlášť nesúvisí.

Aj napriek rozdielu vo výskyte môžu byť dve slová sémanticky blízke. Naopak pri slovách s vysokým výskytom, je vysoká pravdepodobnosť, že sa budú vyskytovať na spoločných stránkach aj spolu so slovami, s ktorými na počutie nemajú nijakú podobnosť.



Obr. 3.2: Graf sémantických podobností získaný pomocou NGD

Iný problém, súvisí so spomínaným algoritmom vyhľadávania google je, že občas pri slove s nízkym výskytom na www stránkach sa pri porovnávaní so slovom podobným a viacej frekventovaným vyskytne prípad, že počet stránok obsahujúcich obidve slová je väčší ako počet stránok vrátených pre menej frekventované slovo. Tento problém vyplýva z nejasných vlastností vyhľadávacieho algoritmu používaného google. Dá riešiť tak, že ak je počet výsledkov vrátených pre priemer väčší ako hodnota, vrátená pre nejaké konkrétne slovo, tak automaticky sa dá prehlásiť o slovách, že sú podobné. (V zobrazenej tabuľke bola takouto hodnotou

jednotka)

Najväčším nedostatkom merania podobnosti slov pomocou metriky NGD, je ale fakt, že nespĺňa definíciu byť dobrou metrikou. Ako možno vidieť aj na obrázku 3.2, tak na základe nameraných vzdialeností medzi slovami mohli vzniknúť aj hrany pre slová, s ktorým na počutie nie sú nijak podobné. Navyše, pre to, aby takýto graf nebol kompletný, sme aplikovali prístup, na základe ktorého sme hrany dávali len medzi tie vrcholy, kde nameraná hodnota podobnosti presiahla vopred zvolenú hodnotu. Takéto skórovanie, hrán ale nie je možné vo všeobecnosti.

Výhodou prezentovanej metriky je jej relatívne nízka náročnosť. Podobnosť pre n slov pri porovnávaní po dvojiciach vieme získať v čase $O(n^2)4\delta$, kde δ predstavuje konštantu, vyjadrujúcu čas potrebný na komunikáciu s vyhľadávacím serverom. Tiež je predpoklad, že takéto meranie sémantickej vzdialenosti, bude úspešné pri meraní vzdialenosti kratších slovných spojení.

3.2.2 Vzdialenosti na základe rozdielu výskytu

Na základe toho, že sa NGD neukázalo ako vhodný spôsob na overovanie sémantickej podobnosti, kvôli kolísavosti výsledkov, sme uvažovali o iných možných spôsoboch merania sémantickej podobnosti založených na počte výsledkov vrátených www vyhľadávačom. Snažili sme sa potvrdiť, alebo vyvrátiť to, že sémantickú podobnosť nemožno určovať len na základe výsledkov vrátených www vyhľadávačmi. Uvažovali sme dve metriky merania vzdialenosti. Jednu sme definovali ako $\frac{|A \cap B|}{|A \cup B|}$ a následne druhú ako $\frac{|A \cap B|}{|\min A, B|}$.

A predstavuje množinu stránok, vrátených pre prvé slovo, B zas pre druhé. Prvou spomenutou metrikou sme chceli preskúmať možnosť určovania podobnosti na základe toho, že sme dali do pomeru spoločný výskyt oboch slov na internete k všetkým stránkam, na ktorých sa aspoň jedna vyskytuje. V druhej metrike sme skúmali vzdialenosť slova s menším výskytom od prieniku oboch slov.

Obe tieto metriky sú relatívne jednoduché na vypočítanie. Obe obsahujú iba 2 krát komunikáciu potrebnú s vyhľadávacím strojom. Pri spomínaných metrikách sa ale definitívne potvrdilo, že nemožno vychádzať pri meraní podobností slov iba z vzdialenosti získanej na základe spoločných výskytov na stránkach.

Testy vykonané na takýchto metrikách potvrdili, že ak chceme získať objektívnu metriku, ktorá bude spĺňať aj definíciu dobrej metriky, tak je potrebné brať do úvahy aj iné faktory ako spoločný výskyt slov na stránkach. Ako definitívny kontra príklad pre takéto metódy môžme použiť vlastnosť, že prirodzená ľudská reč sa neustále vyvíja. Len čo sa objaví v jazyku slovo, ktoré je veľmi blízke inému slovu, ktoré sa už nezanedbateľný čas používa, tak na základe spomínaných výsledkov dostaneme, že dané slová spolu súvisia len minimálne, aj keď to nemusí byť v konečnom dôsledku pravda. Je preto potrebné uvažovať nad iným spôsobom určovania podobnosti slov.

3.3 Vzdialenosť na základe blízkosti slov v spoločných textoch

Položme si na začiatku tejto časti otázku. Ako je možné, že pre človeka v prirodzenej reči nie je žiadny problém zisťovanie významu slov ? Človek sa učí jeho rodný jazyk od narodenia. Používaním jazyka získava postupne jeho základné znalosti. Kontext jednotlivých slov získava z viet a tento kontext je určený ostatnými slovami spomenutými v ľudskej reči.

Nastáva teraz otázka, akým spôsobom, by na základe výskytu jednotlivých slov na internete bolo možné získať význam týchto slov a definovať sémantickú podobnosť medzi slovami. Je dobré si uvedomiť, že tak ako aj v ľudskej reči, tak isto aj na internete sa jednotlivé slová vyskytujú rôzne v spojení s inými slovami. Význam slova určuje do veľkej miery ostatná časť vety.

Keď sa pozrieme na obsah stránok, vidíme, že z ich textového obsahu sa dá určiť význam slova. Frázy ako "auto na ceste", "strom v záhrade" celkom dostatočne určujú sémantickú podobnosť týchto slov. Podobnosť slov možno teda určovať na základe ich spoločného výskytu v jednotlivých slovných spojeniach.

Pre zostrojenie algoritmu nájdenia sémantickej podobnosti slov sa pozrime ešte raz na pravdepodobnosť, že nejaká skupina slov naozaj tvorí kontext v ľudskej reči. Už v kapitole, kde sme predstavili problematiku, sme povedali, že

$$P(C) = \frac{freq(C)}{N}$$

pre nejaký kontext C a jeho frekvenciu. Teraz na základe spomínaných faktov vyslovíme nasledujúcu definíciu.

Definícia 3.3.1 *Web Kontextom veľkosti 2 (W2C) budeme nazývať taký kontext, ktorý pozostáva z 2 slov.*

Na základe tejto definície vieme vyjadriť pravdepodobnosť výskytu 2 slov v jednom kontexte nasledovným spôsobom.

$$P(W2C) = \frac{freq(W2C)}{N}$$

Frekvenciu kontextu W2C vieme vyjadriť pomocou funkcie, ktorej hodnoty závisia od výsledkov www vyhľadávača ako

$$freq(W2C) = g(w_1, w_2),$$

kde w_1 a w_2 predstavujú slová z kontextu $W2C$, a funkcia g vracia počet stránok vrátených www vyhľadávačom, kde sa w_1 a w_2 vyskytujú dostatočne blízko seba. Za hodnotu N, možno teraz dosadiť počet všetkých stránok, ktoré ľubovoľne ďaleko od seba obsahujú slová w_1, w_2.

Keďže sme si povedali, že W2C tvoria dve slová, môžeme teraz konečne ukázať pravdepodobnosť podobnosti 2 slov založenú na www vyhľadávači. Znakom $\circ$ budeme označovať reláciu byť podobný.

$$P(w_1 \circ w_2) = \frac{g(w_1, w_2)}{N},$$

Táto pravdepodobnosť už vyjadruje vzdialenosť slov, získanú pomocou www vyhľadávača. Toto bude naša metrika na meranie podobnosti slov. Od tejto časti práce, ak spomíname metriku merania podobnosti slov, tak máme na mysli práve túto metriku, pokým nie je povedané inak.

Hodnoty podobnosti 2 slov získané takouto metrikou pre dvojice slov sa môžu líšiť podľa toho v akom veľkom priestore N hľadáme frázy.

Algoritmus na zistenie podobnosti založený na vyhľadávaní v krátkych frázach je na obrázku 3.3

Na základe uvedeného algoritmu môžme vysloviť ešte nasledovnú definíciu

VSTUP: Slová w_1, w_2 z prirodzeného jazyka, presnosť d
VÝSTUP: Číselne vyjadrená podobnosť medzi danými slovami
1 Nájdi obsah n stránok obsahujúcich obe slová
2 Zisti počet z n stránok, kde existuje výskyt slov vo vzdialenosti d
3 Vráť vyrátaný pomer všetkých nájdených stránok
a stránok z výskytom slov vo vzdialenosti d

Obr. 3.3: Algoritmus nájdenia vzdialenosti slov, založený na slovných spojeniach

Definícia 3.3.2 *Hovoríme, že dve slová sú d-podobné, ak algoritmus 3.3 vráti pre nejakú hodnotu, na základe ktorej ich môžme prehlásiť za podobné.*

Metrika založená na tomto algoritme je na rozdiel od predošlých metrík založená len na spoločných stránkach, kde sa obe slová vyskytujú. Neuvažujeme pri nej vôbec o ostatných stránkach, kde sa slová nevyskytujú. Presnosť vzdialenosti medzi slovami, ktorú vypočíta spomenutý algoritmus vyplýva z dostatočného počtu stránok, kde sa obe tieto slová vyskytujú.

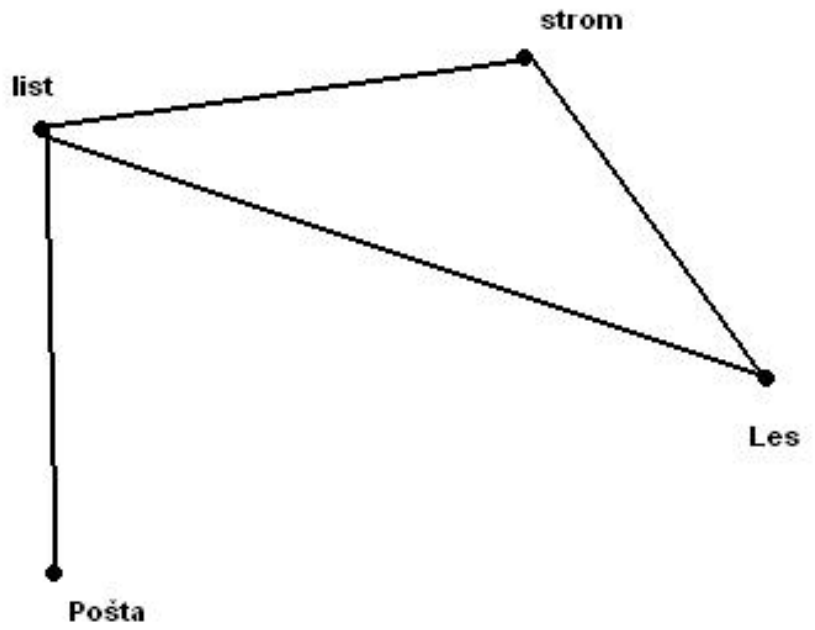
Spoločný výskyt 2 slov na relevantnom počte www stránok sa potvrdil na neúspechu predošlých metrík. V uvádzanom algoritme testujeme, či sa pri spoločných výskytoch slová vyskytujú dostatočne blízko pri sebe viackrát na základe čoho môžme hovoriť o ich podobnosti. Pri spomínanom algoritme treba uvažovať o podstatne dlhšom čase potrebnom na zistenie podobnosti slov.

Na zmeranie podobnosti medzi dvoma slovami v tomto prípade nestačí už len jedným požiadavkám na vyhľadávací server získať počet záznamov pre daný dotaz, ale treba takisto pozerať obsah konkrétnych stránok vrátených vyhľadávačom. Pri zisťovaní podobnosti využívame aj fakt, že vyhľadávací stroj google vracia najskôr záznamy s vyšším rankingom na základe čoho sa môžme spoľahnúť, že tieto stránky obsahujú relevantný obsah. V prípade, že by to tak nebolo, bolo by nutné skontrolovať krátke frázy na všetkých stránkach.

V našej práci sme uvažovali o n v rozmedzí od 50 - 500 podľa toho koľko slov sme navzájom porovnávali. Keďže vykonať 500 vyhľadávacích dotazov na rôzne stránky, by bolo neúnosné, pri našich meraniach sme uvažovali, že slová 2 slová budeme prehlasovať za podobné, ak sú maximálne vzdialené od seba tak, ako je

dĺžka popisu stránky vráteného google vyhľadávačom. To nám umožní na jeden dotaz pre vyhľadávač skontrolovať 10 stránok, čiže potrebný počet dotazov na google server bude 50.

Ukazuje sa, že metrikou na základe algoritmu 3.3 môžme nazvať dobrou metrikou. V kapitole venovanej testovaniu uvedieme výsledky, ktoré sa nám podarilo na základe tejto metriky dosiahnuť. Hlavným prínosom takéhoto spôsobu merania vzdialenosti medzi slovami je, že vieme na jeho základe zostrojiť sémantický model, ktorý umožní zisťovať ďalšie vlastnosti o jazyku. V nasledujúcej časti opíšeme tento model a ukážeme, ktoré problémy ohľadom sémantickej podobnosti vieme na ňom správne riešiť a ktoré naopak nie.



Obr. 3.4: Graf sémantických podobností získaný algoritmom 3.3

3.4 Grafový model podobnosti slov

Na základe definovanej metriky sémantickej vzdialenosti predstavenej v predchádzajúcej je možné slová na základe ich sémantického významu uložiť do zvolenej štruktúry. Na tejto štruktúre je potom následne možné overovať dodatočne ešte iné vlastnosti vyplývajúce zo sémantickej podobnosti slov. V tejto časti ukážeme,

ako sme reprezentovali model sémantickej podobnosti. V tejto časti pojem vzdialenosti medzi slovami vyjadruje vzdialenosť, získanú použitím algoritmu 3.3.

My sme v našej práci vyjadrili namerané hodnoty neorientovaným grafom $G = (V, E)$ s ohodnotenými hranami tak, že množina vrcholov grafu predstavovala všetky slová, ktorých možný význam sme chceli získať.

Váha hrán grafu predstavovala nameranú mieru podobnosti a samotná hrana medzi slovami vyjadrovala, že na základe zvolenej prahovej hodnoty sme prehlásili slová za podobné. Váha hrán predstavovala nameranú vzdialenosť medzi slovami. Vo vrcholoch a hranách je možné mať uložené aj iné informácie.

V nasledujúcej časti popíšeme, aké dodatkové informácie je užitočné ukladať vo vrcholoch takto vytvoreného grafu.

3.4.1 Možnosti informácie vo vrchole

Vo vrcholoch takéhoto grafu je možné zapamätávať si okrem slova aj iné informácie potrebné pre určenie sémantickej podobnosti.

Užitočnou informáciu, ktorú si možno ukladať vo vrchole je, slovný druh slova. V našej práci sme uvažovali o 3 slovných druhoch. Podstatných menách, prídavných menách a slovesách. Na týchto slovných druhoch boli vykonávané aj všetky testy, ktorých výsledky popisujeme v nasledujúcej časti práce.

Definícia 3.4.1 *Nech slová v_1 a v_2 sú prídavné mená z prirodzeného jazyka a w nech je podstatné meno. Majme binárnu reláciu $\sim$ vyjadrujúcu podobnosť slov. Potom ak pre prídavné mená v_1 a $_2$ platí, že $v_1 \sim w$ a zároveň aj $v_2 \sim w$ tak potom platí aj $v_1 \sim v_2$*

Predchádzajúca definícia vraví o tom, že relácia byť podobný pre prídavné mená spĺňa vlastnosť tranzitívnej relácie. V našej práci sme sa venovali pri implementácii vyriešeniu problému podobnosti prídavných mien na základe hrán v uvedenom grafe.

Na algoritme 3.3 možno vidieť, ako sme pridávali podobnosť medzi adjektíva. Nepodarilo sa nám ale zistiť na základe akej hodnoty hrany by sme mohli považovať, adjektíva za synonymá. Pri iných slovných druhoch nebola takáto konštrukcia možná. Na nasledujúcom obrázku je ešte zobrazenie novovytvoreného

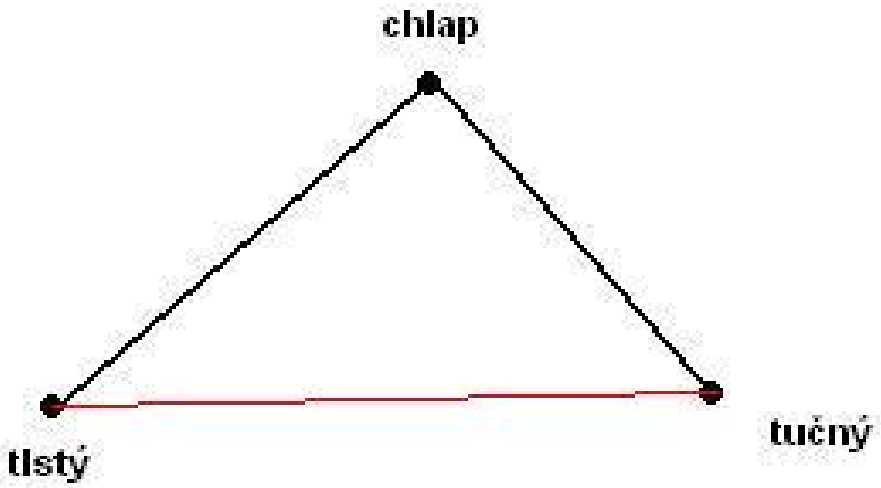
VSTUP: Grafová štruktúra opisovaná v tejto časti a podmnožina A vrcholov grafu, obsahujúca vrcholy reprezentujúce adjektíva.

VÝSTUP: Graf s pridanou hranou medzi vrcholy reprezentujúce adjektíva s hranou do rovnakého podstatného mena

```
for v ∈ A
    for w ∈ a
        if v.to == w.to
            pridajHranu(v,w)
```

Obr. 3.5: Algoritmus pridania podobnosti adjektívam

vzťahu podobnosti medzi slovami. Červená hrana zobrazuje novú hranu pridanú prezentovaným spôsobom.



Obr. 3.6: Príklad vzniknutej štruktúry adjektív v grafe 3.3

3.4.2 Hľadanie kontextu a jeho význam pri zisťovaní podobnosti

Na základe zistení z predošlej časti sme dospeli k tomu, že pri meraní podobnosti slov na základe www vyhľadávačov sa môžeme stretnúť hlavne s problémom, že

nameraná hodnota podobnosti neodráža úplne skutočnú sémantickú podobnosť slov v ľudskom ponímaní. V ľudskej reči je význam slova veľmi jednoduché zistiť, na základe slov, spolu s ktorými sa vyskytuje pri hovorení. Aby sme vedeli skutočne poznať sémantický význam slova, potrebujeme poznať ostatné slová, spolu s ktorými sa vyskytuje v rôznych kontextoch.

Nasledovná úloha je nájsť všetky kontexty daného slova. Na to, aby sme vyriešili problém nájdenia kontextu, potrebujeme poznať namerané hodnoty medzi jednotlivými slovami a mať zvolenú prahovú hodnotu, na základe ktorej budeme slová z prirodzeného jazyka prehlasovať za podobné.

Na vyriešenie tohto problému a vytvorenia modelu, v ktorom je možné definovať vlastnosti vyplývajúce z podobnosti slov sme použili spomínaný grafový model. Kontexty, v ktorých sa môžu jednotlivé slová vyskytovať sme reprezentovali podgrafmi v okolí vrcholu, ktorý reprezentoval dané slovo(clustermi).

Od jednotlivých podgrafov sme vyžadovali, aby spĺňali podmienku, že sú úplné grafy. Našli sme teda všetky úplné podgrafy, v okolí vrcholu v reprezentujúceho dané slovo. Následne sme hľadali maximálne nezávislé podgrafy(také, že pre ne platí, že nájdený podgraf nie je podgrafom žiadneho iného). Všetky slová reprezentované takýmito podgrafmi sme už vyhlásili za kontexty daného slova. Algoritmus nájdenia kontextov je na obrázku 3.7

Poznanie kontextu nám umožňuje zaradiť slovo na určité miesto v korpuse jazyka. Tiež nám umožňuje, vyhľadávanie slov v určitom kontexte. Príklad takéhoto vyhľadávania uvádzame v nasledujúcej časti.

3.4.3 Porovnávanie textov

Keďže pomocou www vyhľadávačov nie je možné porovnávať celé texty, tak sme sa zaoberali v našej práci možnosťou porovnávania vybraných textov na základe výsledkov pre jednotlivé slová vrátené www vyhľadávačmi. Vo viacerých súčasných prácach je snaha číselnou hodnotou vyjadriť podobnosť textov. Asi najčastejší prístup je definovať vzdialenosť textov t_1 a t_2 nasledovne

$$sim(t_1, t_2) = \sum_{w_i \in t_1} \sum_{w_j \in t_2} sim(w_i, w_j)$$

Pri takto zvolenom prístupe sa ale znova dostávame ku problému, ktorý v tejto práci riešime pre slová. To je potreba zvolenia prahovej hodnoty, od ktorej možno

VSTUP: Slová z prirodzeného jazyka, prahová hodnota,
namerané hodnoty podobnosti medzi slovami
VÝSTUP: Podgrafy reprezentujúce kontexty
1 Zostroj vrcholy reprezentujúce slová z prirodzeného jazyka
2 Na základe prahovej hodnoty zostroj hrany medzi vrcholmi.
3 Pre každý vrchol :
3.1 Nájdi všetky kompletné grafy, ktoré vrchol vytvára spolu
s grafmi vo vzdialenosti 1 od neho.
3.2 Odstráň zbytočné podgrafy
3.3 Zapamätaj si ostávajúce podgrafy
ako kontexty slova reprezentovaného vrcholom.

Obr. 3.7: Algoritmus vytvorenia modelu reprezentujúceho podobnosť a nájdenie kontextov slova

texty prehlasovať za podobné. Preto sme v našej práci volili prístup založený na porovnávaní jednotlivých textov na základe kontextov.

Ako ukážku sme implementovali jednoduché vybranie kontextu, ktorý najlepšie pasuje pre zadaný text. Implementovali sme to nasledovne

1. Pre každý kontext nájdi počet výskytov v texte.
2. Vráť kontext s najväčším výskytom

Uvažovali sme aj nad zohľadňovaním vzdialenosti slov medzi sebou, ale nakoniec sa ukázalo, že pre jeden text to nemá veľký štatistický význam a preto prezentujeme iba tento spôsob.

Kapitola 4

Výsledky a testovanie

V tejto kapitole postupne popíšeme výsledky, ktoré sme dosiahli v oblasti zisťovania sémantickej podobnosti slov našim spôsobom. Na začiatku uvedieme porovnania s existujúcimi spôsobmi zisťovania sémantickej podobnosti slov. Následne sa pozrieme na výsledky, ktoré sme dostali aplikovaním jednotlivých metrík.

4.1 Porovnanie s existujúcimi metrikami

V tejto časti uvedieme stručné porovnanie nami prezentovaného spôsobu určovania podobnosti slov a spôsobmi prezentovanými v úvode tejto práce a poukážeme na niektoré vlastnosti našej prezentovanej schémy v porovnaní s taxonomickou štruktúrou wordnet.

Všetky v úvode zmieňované metódy vyžadovali analýzu veľkého množstva textov a značné informácie o jazyku. Výhodou nášho prístupu je, že v súčasnosti existuje na internete nepreberne veľa textov, ktoré dokážu nahradiť rôzne statické texty. Ako výhoda meranie podobnosti pomocou www vyhľadávačov sa ukazuje aj fakt, že výsledky dokážu relatívne rýchlo odraziť novovzniknuté tvary slov jazyka.

Úplne odlišný od siete wordnet je náš prezentovaný spôsob na určovanie sémantickej podobnosti medzi slovami. Kým štruktúra worndet vznikla na základe spolupráce matematikov a jazykovedcov, tak nami prezentovaná štruktúra vznikala automatizovane. Oba prístupy majú svoje negatíva. Hlavným negatívom nášho prístupu z pohľadu zložitosti je, že na vytvorenie všetkých kontextov slova

potrebujeme, exponenciálny čas vzhľadom na počet susedov daného vrcholu. Ale výhodou je, že ak vieme približne aké texty ideme porovnávať, nemusíme vytvárať takúto štruktúru z celého jazyka, ale stačí ak budeme pracovať len z jeho podmnožinou.

4.2 Voľba prahovej funkcie

Táto časť tejto kapitoly sa zaoberá voľbou prahovej funkcie. Vedieť zvoliť správne prahovú funkciu sa ukázalo ako silne nedeterministické. Iné je, keď porovnávame všetky slová z jazyka, alebo aspoň slová rovnomerne rozložené tak, že pre každé slovo je približne taký istý počet slov, ktoré s ním sémanticky nesúvisia a taký istý počet slov ,ktoré s ním naopak sémanticky úzko súvisia.

Na druhej strane je tu prípad, keď porovnávame len slová, ktoré spolu väčšinou úzko sémanticky súvisia. Vždy je možné voliť hodnotu prahovej funkcie v parametri (na základe toho aké máme slová), alebo je možné použiť nasledujúci postup k jej voľbe.

VSTUP: Presnosť n = počet stránok z ktorých
berieme výsledky, dvojice slov rozdielnych
a aj dobrých na ľudské počutie
VÝSTUP: Prahová hodnota
1 Zrátaj podobnosť dvojíc zo vstupu
2 Vráť priemernú hodnotu z takto zmeraných hodnôt

Obr. 4.1: Heuristika pre hľadanie

Hore uvedená heuristika popisuje voľbu prahovej hodnoty. Veľmi dôležitým pre jej fungovanie je parameter n, ktorý hovorí o počte stránok vrátených www vyhľadávačom. Je dôležitý preto, aby sme správne zvolili prahovú hodnotu. Totiž, relevantnosť výsledkov vrátených na neskorších pozíciách klesá. Preto je nutné, aby parameter n odpovedal skutočnému počtu stránok vrátených www vyhľadávačom, z ktorých sme brali výsledky.

Tento algoritmus dostáva na vstup dvojice slov. Polovicu dvojíc takých, že sú

sémanticky podobné na počutie pre človeka a druhú polovicu zase takých, ktoré podobné nie sú. Toto nám umožní objektívne zmerať prahovú hodnotu. Viacej o implementácií merania tejto hodnoty sa čitateľ dozvie v nasledujúcej kapitole venovanej implementácií.

4.3 Výsledky nameraných hodnôt na jednotlivých slovách

V tejto časti postupne ukážeme ako merajú podobnosť prezentované metódy v tejto práci. Postupne porovnáme jednotlivé metriky na zisťovanie podobnosti slov.

Ako prvé si porovnáme metriku NGD s nami prezentovaným algoritmom na zisťovanie podobnosti slov. Uvidíme na príklade niektoré jeho problémy. Výsledky sú v nasledujúcej tabuľke. V druhom stĺpci je algoritmus, ktorý prezentujeme v tejto práci.

Dvojica:	NGD	algoritmus 3.3
graf : les	0.9022182018115772	0.55
graf : príroda	0.6846934459877567	0.29
graf : smrad	0.7947269028096201	0.09
les : príroda	0.7451024698411022	0.78
les : smrad	0.9627718306644065	0.081
príroda : smrad	0.3870362754184061	0.17

V predchádzajúcej tabuľke je možné vidieť výsledky 2 porovnávaných algoritmov. Keď si zoradíme hodnoty, ktoré sme dostali pomocou nami prezentovaného jednoduchého algoritmu od najmenšej po najväčšiu, uvidíme, že namerané hodnoty súhlasia s podobnosťou ako ju vníma človek. Najviac je sémanticky podobný les s prírodou. Na druhom mieste je graf s lesom. Nameraná hodnota možno niekomu nezbehlému v matematických vedách nepríde ako korektná, ale predsa len možno na základe nej určiť, že v istom kontexte sú si tieto 2 slová podobné.

Na rozdiel od toho, ak meriame za použitia NGD, tak dostávame hodnoty veľmi rôzne, z ktorých na pohľad len veľmi ťažko určiť do akej miery sú si dané slová podobné. Asi najväčším problémom v prezentovaných výsledkoch je name-

ranie väčšej hodnoty pre úplne nezmyselnú dvojicu les-smrad ako pre zdanlivo sémanticky veľmi blízku dvojicu les-príroda.

V predchádzajúcich úvahách sa potvrdilo, že porovnávanie slov na základe textov na internete funguje lepšie ako porovnávanie založené na NGD. V predchádzajúcej časti sme, ale vyslovili dohad, že pridanie informácie v podobe vyhľadávania konkrétnych fráz do značnej miery pomôže pri presnosti NGD. V nasledujúcej krátkej tabuľke ukážeme či sa tento dohad potvrdil alebo nie.

Dvojica:	nameraná hodnota podobnosti s NGD
bill gates : zakladatel microsoft	0.31
van gogh : impresionisticky maliari	0.54
iveta radicova : impresionisticky maliari	0.94
iveta radicova : premierka slovenska	0.32

V uvedených výsledkoch vidíme, že sa nepotvrdil dohad, že keď vezmeme konkrétne frázy, tak dostaneme výsledky, ktoré by lepšie vedeli napovedať o sémantickej podobnosti 2 slov.

Prezentovali sme aj 2 iné metriky. Ich vlastnosti oli vo výsledkoch veľmi podobné metrike NGD. Vyznačovali sa hlavne značnou kolísavosťou pre jednotlivé slová a tiež sa ukázalo, že určiť z nich kedy sú slová podobné a kedy nie je veľmi ťažké.

Tiež sa ukázalo, že pre testovanie slov v slovenčine je naša metóda vhodná, ale pre slová z väčšieho priestoru slov, ako sú slová v angličtine je pre nás ťažšie testovateľná. Vo väčšom priestore slov dochádza totiž k problému, že na väčšom počte stránok, ktoré sú vrátene ako prvé sa nachádzajú aj menej podobné slová blízko seba.

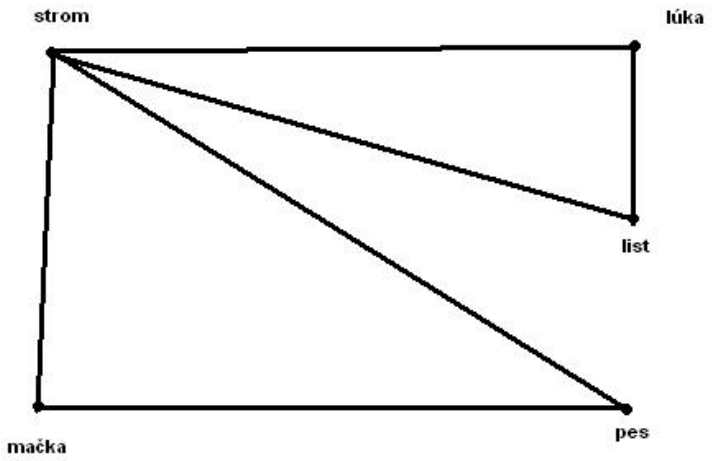
Na spustení testov sa nám tiež podarilo ukázať že na základe podobností podstatných mien s adjektívom, možno považovať adjektíva za podobné.

V nasledujúcej tabule uvádzame stručný príklad nameraných hodnôt.

Dvojica:	nameraná hodnota podobnosti s NGD
krásna : žena	0.71
pekná : žena	0.73
krásna : pekná	0.19

Na základe vyššie uvedených výsledkov, môžeme takto spájať hranami adjektíva, ktoré sú v kontexte s rovnakým podstatným menom.

Na záver ešte uvedieme obrázok s hranami vyjadrujúcimi reláciu podobnosti, ktorú sa nám podarilo získať za použitia našej metriky. Na základe metriky sa nám podarilo zvoliť hrany tak, aby sme použitím prahovej hodnoty mohli dobre vytvoriť štruktúru grafu.



Obr. 4.2: Graf vyjadrujúci podobnosť slov.

Na testoch sa potvrdilo, že naša metrika na meranie podobnosti slov fungovala dobre. Ukázalo sa, že v prípade rozhodovania podobnosti, je v porovnaní s metrikami, založenými iba na spoločnom výskyte slov úspešnejšia nami prezentovaná metóda, založená na výskyte slov vo frázových spojeniach, ktorou sa dokážeme lepšie priblížiť ľudskej mysli.

Kapitola 5

Implementácia

V tejto kapitole popíšeme zisťovanie podobnosti z hľadiska implementácie. Popíšeme tiež niektoré problémy, s ktorými sme sa stretli pri implementácií. Budeme sa venovať architektúre, ktorú sme pri implementácií použili.

5.1 Technológie

Grafový model sme celý implementovali v jazyku C++. C++ je objektovo orientovaný programovací jazyk. Jeho výhoda oproti novším programovacím jazykom je jeho rýchlosť. Keďže C++ už existuje viacero rokov, existuje veľa veľmi kvalitných a odladených kompilátorov tohto jazyka.

Ďalším dôvodom prečo programovať programy v C++ a nie napríklad v C je väčšia ponuka rôznych knižníc jazyka C++, čo umožňuje pohodlnejšie programovanie konkrétnych aplikácií. Ako príklad takýchto knižníc možno uviesť STL (Standard Template Library). Je to knižnica, ktorá zahŕňa v sebe prvky potrebné pre kvalitnú a hlavne pohodlnú implementáciu. Patria sem rôzne štruktúry a algoritmy, pomáhajúce pri implementácií programov.

Na druhej strane ani v C++ nie je obsiahnuté úplne všetko. Preto sme zvolili na implementáciu sýtemu pre komunikáciu s vyhľadávacím serverom google programovací jazyk java. Bolo to hlavne z dôvodu, že existuje voľná knižnica od spoločnosti Apache, ktorá umožňuje jednoduchý spôsobom implementovať HTTP klienta a komunikovať so serverom.

Na vyhľadávanie pomocou google sme využívali jednoduché požiadavky, pri-

čom parametre čo vyhľadávame sme si prenášali v url adrese. Podstatné parametre, ktoré sme potrebovali prenášať boli jazyk, v ktorom sa majú výsledky zobrazovať, ktorá stránka výsledkov vyhľadávania sa má zobraziť a samozrejme dotaz pre vyhľadávané slovo. Viacej o vyhľadávaní pomocou google sme už popísali v tretej kapitole.

5.2 Architektúra

Celý systém metód, ktoré mali pomôcť k zisteniu sémantickej podobnosti slov sme implementovali ako sme už spomenuli v 2 programovacích jazykoch. Jedným bola java a druhým C++. Predávanie nameraných hodnôt z programu v jave do programu v C++ sme realizovali prostredníctvom jednoduchých textových súborov.

V jave sme implementovali spomínané metriky medzi dvoma slovami. Skonštruovali sme dve triedy, pričom v jednej sme implementovali samotné počítanie metrík a prahovej hodnoty. Potrebné slová sme si brali z textového súboru. V druhej sme implementovali spracovávanie HTML odpovede od vyhľadávacieho servera. Pri komunikácií s vyhľadávacím serverom sme využívali už zmienenú knižnicu od Apache na hitovanie servera google.

V C++ sme implementovali grafovú štruktúru, ktorá vyjadruje model podobnosti slov. Základné triedy takejto implementácie boli SimGraph, Node a Edge. Pomocou tried Node a Edge sme reprezentovali vrcholy a hrany grafu. Slová uložené vo vrchole boli vyjadrená prostredníctvom členskej premennej typu string triedy Node. Hrany reprezentované triedou Edge 2 obsahovali členské premenné typu smerník na Node vyjadrujúce koncové vrcholy hrany. Každá hrana mala tiež ako členskú premennú typu double, ktorá vyjadrovala hodnotu hrany.

Celý graf sme reprezentovali triedou SimGraph. Tá mala členskú premennú typu vector<Node> vyjadrujúcu zoznam vrcholov grafu. V tejto triede sme ako metódy publikovali všetky spomínané algoritmy.

5.3 Problémy

V tejto časti popíšeme technologické problémy s ktorými sme sa stretli pri našej implementácií. Táto časť má pomôcť jej čitateľovi, aby sa jednoduchšie preniesol cez technologické problémy pri takomto zisťovaní sémantickej podobnosti.

Ako prvý problém spomenieme fakt, že pri implementácií treba počítať s problémom pri vyhľadávaní. Google sa bráni proti tomu aby bol jeho server zahlcovaný nadmerným množstvom requestov z rôznych automatov. Preto pri zisťovaní podobnosti väčšej množiny slov, nastáva problém, že po čase začne vyhľadávací stroj google vracať chybové stránky, ktoré neobsahujú dáta potrebné pre určovanie podobnosti slov. Tento problém sme riešili tak, že sme do zdrojového kódu vkladali bloky, ktoré program na chvíľu uspali. Tiež sme na testovanie využívali viacej sád slov, pričom sady boli menšie. Výsledky sme až potom spájali.

Druhý problém súvisel so samotným spracovávaním stránky, ktorú vrátil google vyhľadávač. Tu už bolo samotný problém to, že v čase písania tejto práce neexistovala knižnica, ktorá by umožňovala pracovať s odpoveďami z google servera. Na zistenie toho, či sú slová blízko v texte vedľa seba sme mohli využiť len zoznam vrátených stránok. Ten obsahuje aj krátky text obsahujúci hľadané slová. Či sú slová blízko seba v texte sme museli rozhodovať tak, že sme zistili či sú medzi danými slovami obsiahnuté tri bodky, čo znamená, že je medzi nimi viacej znakov. Pri takomto prístupe sme, ale nemohli aplikovať parametrizovatelnú vzdialenosť slov. Na druhej strane ale výhoda tohto prístupu je, že vyžaduje 10 násobne menej requestov na nejakú www stránku.

Najzávažnejším problémom, ale pri písaní tejto práce bolo riešenie kódovania slovenského textu vráteného vyhľadávačom. Nepodarilo sa nám zistiť, aké vhodné kódovanie použiť, aby sme mohli korektne určiť podobnosť dvoch slov obsahujúcich diakritiku. Dochádzalo k tomu, že odpoveď prijatá od vyhľadávacieho servera s diakritikou bola nečitateľná.

Tento problém nás ale nakoniec doviedol k čiastočnému riešeniu, ktoré v konečnom dôsledku, zefektívňuje určovanie podobnosti slov. Ako náhle zadáme vyhľadávaču google nejaké slovo, automaticky zobrazí na svojej stránke všetky tvary tohto slova. Všetky výskyty toho slova sa na google stránke nachádzajú v tagoch. Pri hľadaní slov na stránke preto nebolo vždy potrebné vyhľadávať sa-

motné slová, ale v prípade menších sád slov sa stačilo pozrieť na výskyt em tagov. My sme pri menších setoch slov vyhľadávali aj s použitím časti slova.

Obr. 5.1: Ukážka krátkeho textu, ktorý sme museli spracovať aby sme zistili vzdialenosť slov

Posledným, ale nie príliš závažným nedostatkom, s ktorým sme sa stretli, bolo, že stránka ktorú sme dostali od vyhľadávacieho servera nepredstavovala koketné html a preto sme ju museli spracovávať ako reťazec a nebolo možné použiť iné kvalitnejšie spracovávanie html objektu.

Kapitola 6

Záver

Problematika skúmania sémantickej podobnosti slov spadá pod oblasť informatiky zaoberajúcou sa spracovaním prirodzeného jazyka. Tejto problematike sa informatika venuje už viacero rokov. Momentálny prístup k riešeniu problémov v oblasti spracovania prirodzeného jazyka je skúmať zvlášť niektoré podproblémy v tejto oblasti a postupne ich spájať do celku. Momentálne je vyriešená len malá časť tejto problematiky.

Hlavnou príčinou zložitosti výpočtového spracovania prirodzeného jazyka je nedeterminizmus tejto úlohy. Reč človeka je riadená mozgom, o ktorom vieme v súčasnosti stále veľmi málo. Ako sa aj ukazuje na výsledkoch našej práce k dobrému riešeniu v úlohách nemusia nutne viesť tie najlepšie algoritmy, ale mnohokrát je riešenie založené na heuristickom prístupe kedy sa snažíme preniesť správanie človeka na počítač.

V našej práci najskôr predstavujeme existujúce metódy merania podobnosti slov, založené prevažne na taxonomických štruktúrach vybudovaných človekom. Následne na to prechádzame k automatizovanému určovaniu sémantické podobnosti slov na základe www vyhľadávačov. Predstavujeme jednoduchú metriku na meranie podobnosti medzi slovami na základe ktorej, je možné zostrojiť jednoduchý model, v ktorom sú slová uložené a umožňuje následne dodefinovať niektoré vzťahy medzi slovami a nájsť kontexty pre slová. Nájdením konceptov slov sa nám podarilo začleniť slovo na jeho miesto v korpuse jazyka.

Na koniec práce uvádzame výsledky ktoré sme dosiahli a porovnávame niektoré naše výsledky z oblasti skúmania sémantickej podobnosti. Rozoberáme prí-

pady pre ktoré je vhodné použiť konkrétnu metriku na zistenie podobnosti.

Treba spomenúť, že nami prezentovaný model nemôže ani v najmenšom konkurovať veľkým taxonomickým štruktúram ani iným modelom, ktoré sa podrobne zaoberajú štruktúrou jazyka.

Možností na pokračovanie v tejto práci je vzhľadom na problémy v danej oblasti veľmi veľa. Bolo by užitočné do budúcna vybudovať aplikáciu s vlastnou databázou, ktorá by vedela na základe výsledkov www vyhľadávačov rozhodovať problém podobnosti, ale zároveň by disponovala aj úložným priestorom pre dáta. Veľmi zaujímavým a ambicióznym pokračovaním je aj pokúsenie sa zostrojiť komunikátor s človekom, ktorý by sémantický význam jednotlivých fráz človeka čerpla práve z www stránok.

Dodatok A

Príloha

Prikladáme CD, na ktorom je napálený java program pre zisťovanie podobnosti. CD tiež obsahuje program napísaný v C++, ktorý slúži na zisťovanie niektorých slov jazyka.

Literatúra

[CM05] Courtney Corley and Rada Mihalcea. Measuring the semantic similarity of texts. pages 13–18, 2005.

[CV04] Rudi Cilibrasi and Paul M. B. Vitányi. The google similarity distance. *CoRR*, abs/cs/0412098, 2004.

[Fel98] Christiane Fellbaum. *Wordnet: An Electronic Lexical Database*. Bradford Books, 1998.

[JC97] Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *CoRR*, cmp-lg/9709008, 1997.

[mss07] Measuring semantic similarity between words using web search engines. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 757–766, New York, NY, USA, 2007. ACM.

[Pal94] Emil Pales. *Sapfo Parafrázovac slocenciny*. Slovenská akademia vied, 1994.

[Res95] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. 1995.

[RTS08] Geetha Manjunath Rajesh Thiagarajan and Markus Stumptner. Computing semantic similarity using ontologies. 2008.

[SH06] Mehran Sahami and Timothy D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 377–386, New York, NY, USA, 2006. ACM.

[Tur05] Peter D. Turney. Measuring semantic similarity by latent relational analysis. *CoRR*, abs/cs/0508053, 2005.

[WP94] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.