Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

FIIT-0000-00000

Martin Nemček

Spracovanie učebných textov

Bakalárska práca

Vedúci práce: Ing. Miroslav Blšták

December 2015

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

FIIT-0000-00000

Martin Nemček

Spracovanie učebných textov

Bakalárska práca

Študijný program: Informatika
Študijný odbor: 9.2.1 Informatika
Miesto vypracovania: FIIT STU BA
Supervisor: Ing. Miroslav Blšták

December 2015

Anotácia

Slovenská technická univerzita v Bratislave
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ
Študijný program: Informatika
Autor: Martin Nemček
Bakalárska práca: Spracovanie učebných textov
Vedúci práce: Ing. Miroslav Blšták
December 2015

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum....

Annotation

Slovak University of Technology Bratislava
FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES
Degree Course: Informatika
Author: Martin Nemček
Bachelor thesis: Spracovanie učebných textov
Supervisor: Ing. Miroslav Blšták
December 2015

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum....

ACKNOWLEDGMENTS

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.....

DECLARATION

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua....

.....

Martin Nemček

Obsah

Zoznam obrázkov

Zoznam tabuliek

Zoznam ukážok

1 Úvod

Internet je v dnešných dňoch zaplnený obrovským množstvom dát a informácií. Mnohé z týchto dát sa na internete vyskytujú mnohonásobne, či už v identickej podobe alebo s úpravou. Avšak, čím ďalej tým viac, z týchto informácií vyskytujúcich sa na internete sú informácie irelevantné.

Stáva sa to až príliš často a myslím, že každý už zažil situáciu, kedy hľadal informácie na nejakú konkrétnu tému a musel sa „prehrabať" kopou nepodstatných informácií, ktoré mu boli ponúkané. Stáva sa to medzi všetkými kategóriami používateľov na internete.

Jednou z majoritných skupín používateľov, ktorý sa s takouto situáciou stretávajú denno denne sú študenti. Študenti všetkých škôl, od základných až po univerzity, získavajú informácie na učenie, projekty alebo zadania primárne z internetu alebo učebných textov kníh. Keď musia prechádzať obrovské množstvá dát z rôznych zdrojov, je to náročne, často až frustrujúce a berie im to veľmi veľa času. Tento čas by mohli využiť efektívnejšie, napríklad na rozvoj svojich vedomostí.

Učebné texty sú často písané v neštruktúrovanej forme a prirodzenom jazyku. Pre stroje je mnohokrát náročné správne interpretovať tieto informácie. Jedným z hlavných dôvodov je fakt, že každý jazyk je odlišný a obsahuje špecifické charakteristiky, ktoré môžu byť napríklad slovosled vety, gramatické kategórie slov, ale aj vetné členy a vzťahy medzi nimi.

Tieto, ale aj mnohé iné charakteristiky jazyka sa dajú využiť pri jeho spracovaní na takú podobu, aby s ním vedeli aj stroje jednoducho narábať. Takýto proces sa nazýva *spracovanie prirodzeného jazyka* (angl. Natural Language Processing - NLP). Spracovanie jazyka ma viacero aplikácií, z ktorých sú to napríklad preklad z jedného jazyka do druhého, vytiahnutie najpodstatnejších entít z textu, prípadne aj štatistika ich výskytu a mnohé ďalšie.

My posunieme spracovanie prirodzeného jazyka ešte o kúsok ďalej a budeme sa zaoberať ako dopomôcť študentom so spracovaním veľkého množstva informácií, hlavne z učebných textov. Študentom najviac pomôže, ak dokážu rýchlo z textu vytiahnuť tie najpodstatnejšie, najdôležitejšie informácie a údaje, ktoré sa im ďalej budú omnoho ľahšie spracovávať. Proces určovania a extrakcie najpodstatnejších

informácií z učebného textu môžme nazvať spoznamkovávanie.

Zameriame sa hlavne na využitie vetných členov a vzťahov medzi nimi, na určenie najpodstatnejšej informácie z vety. Tieto extrahované informácie následne ponúkneme používateľovi (študentovi).

2 Analýza

V tejto kapitole priblížím a rozoberiem čo je spracovanie prirodzeného jazyka, jeho využitie v aplikáciach a systémoch a jeho hlavné úlohy. Ďalej zanalyzujem nástroje, ktoré sa dajú využiť na spracovanie prirodzeného jazyka a tak isto sa pozriem na aplikácie a systémy, ktorých základom je spracovanie textu.

2.1 Spracovanie prirodzeného jazyka

Spracovanie prirodzeného jazyka (angl. Natural Language Processing - NLP) môže byť definované ako automatické alebo poloautomatické spracovanie ľudského jazyka [2]. Počítače doposiaľ nedokážu plne porozumieť ľudskému jazyku, či už sa jedná o písaný alebo hovorový. Preto hlavným hlavným cieľom NLP je vybudovať výpočtové modely prirodzeného jazyka pre jeho analýzu a generovanie [3].

Porozumenie ľudskej reči je mnohokrát náročne aj pre samotných ľudí a nie to ešte pre počítače. Na svete je veľké množstvo jazykov, ktoré sa od seba líšia charakteristikami typickými pre konkrétny jazyk. Taktiež každý človek sa líši a preto výslovnosť rovnakého slova viacerými ľuďmi môže byť odlišná. Ďalej máme slangové slová a slová typické len pre určité územie. Pri spracovávaní prirodzeného jazyka treba vziať do úvahy tieto a aj ďalšie premenné. Dosiahnutie tohto cieľa je preto často veľmi náročné.

V súčastnosti najpoužívanejšie algoritmy na NLP využívajú strojové učenie. Dosiahnutie úplného porozumenia a spracovania ľudského prirodzeného jazyka by znamenalo vyriešiť *AI-complete* problém, čo znamená, že obtiažnosť tohto problému je ekvivalentné s obtiažnosťou problému vytvorenia počítaču inteligentného ako človek, takzvané „true AI".

2.2 Využitie spracovania prirodzeného jazyka

V súčastnosti má NLP niekoľko hlavných hlavných využití v aplikáciach a systémoch. Z hľadiska spracovania učebných textov je pre nás najdôležitejšie využitie z pohľadu *extrakcie informácií*, ktoré je podrobnejšie popísané v sekcii 2.2.1 Extrakcia informácií Ďalšie využitia NLP sú napríklad [4]:

- Strojový preklad (angl. Machine Translation)
- Rozpoznávanie reči (angl. Speech Recognition)
- Sumarizáciu textu (angl. Text Summarization)
- Dialógové systémy (angl. Dialogue Systems)
- Výber informácií (angl. Information Retrieval)

2.2.1 Extrakcia informácií

Systémy a aplikácie zamerané na extrakciu informácií vyhľadávajú a extrahujú informácie z textov, článkov a dokumentov, pričom reagujú na používateľove informačne potreby. Výstup z takýchto systémov a aplikácií nepozostáva iba zo zoznamu kľúčových slov, ktoré by sa dali pokladať za extrahované informácie, ale naopak sú v tvare preddefinovaných šablón [4].

Extrakcia informácií využíva niekoľko z hlavných úloh spracovania prirodzeného textu. Sú to *Značkovanie slovných druhov*, *Rozpoznávanie názvoslovných entít*, a ďalšie [4]. Tieto a aj ostatné úlohy spracovania prirodzeného jazyka sú podrobnejšie opísané v sekcii 2.3 Úlohy spracovania prirodzeného jazyka.

Výber informácií a extrakcia informácií spolu úzko súvisia, ale sú to dve rozdielne využitia NLP. Prvé spomínané využitie slúži na vyhľadávanie relevantných zdrojov informácií v databázach textov, článkov a dokumentov podľa používateľových potrieb. Na vyhľadaných zdrojoch následne prebehne extrakcia informácií.

My sa pri spracovaní textov zameriame hlavne na extrakciu informácií, aby sme dokázali z učebného textu extrahovať relevantné informácie pre študenta, a tým získali poznámky.

2.3 Úlohy spracovania prirodzeného jazyka

NLP ma niekoľko hlavných úloh. Podrobnejšie si priblížime tie, ktoré sú relevantné vzhľadom na implementáciu spracovania učebných textov. Úlohy spracovania prirodzeného textu: [1]

- Značkovanie slovných druhov (angl. Part-of-speech tagging) 2.3.1
- Rozdelenie vety na menšie časti (angl. Chunking)
- Rozpoznávanie názvoslovných entít (angl. Named Entity Recognition) 2.3.2
- Označovanie sémantického postavenie (angl. Semantic Role Labeling)
- Rozpoznanie koreferencií (angl. Coreference resolution) 2.3.3
- Morfologické segmentovanie (angl. Morphological Segmentation)
- Generovanie prirodzeného jazyka (angl. Natural Language Generation)
- Optické rozoznávanie textu (angl. Optical Character Recognition)
- Rozloženie vzťahov (angl. Dependency parsing) 2.3.4
- a mnoho ďalších

2.3.1 Značkovanie slovných druhov

Hlavnou úlohou značkovania slovných druhom (angl. Part-of-speech tagging) je každému slovu vo vete priradiť unikátnu značku, ktorá odrážať jeho syntaktickú úlohu vo vete [1]. Sú to, napríklad v slovenskom jazyku podmet, prísudok, príslovkové určenie alebo v anglickom jazyku noun, adverb, verb, atď. Tak isto to môže byť označenie určujúce množné číslo, napríklad signulár alebo plurál.

Problémom pri značkovaní slovných druhov je mnohoznačnosť. Znamená to, že slovo môže mať viacero významov a môže byť viacerými slovnými druhmi. Napríklad v slovenskom jazyku slovo *kry* môže predstavovať sloveso s významom rozkazu *prikry!*, ale taktiež môže predstavovať podstatné meno s významom *kríky*. V anglickom jazyku to je napríklad slovo *book*, ktoré môže predstavovať podstatné meno (angl. noun) *kniha* alebo sloveso (angl. verb) vo význame *rezervovať*.

2.3.2 Rozpoznávanie názvoslovných entít

Rozpoznávanie názvoslovných entít (angl. Named Entity Recognition) označuje mená a názvy, ktoré sa vyskytujú v texte. Rozdeľuje tieto entity do kategórií, ako sú napríklad *osoby*, *organizácie* alebo *lokácie* [1].

Ťažkosti pri rozpoznávaní názvoslovných entít spôsobuje kapitalizácia slov, takzvané písanie entít s veľkým začiatočným písmenom. V anglickom jazyku to jednoduché, kedže v angličtine sa entity píšu s veľkým začiatočným písmenom. Príklad je *Slovak University of Technology*. Avšak v iných jazykoch to neplatí a entity sa nemusia písať s veľkým začiatočným písmenom.

2.3.3 Rozpoznanie koreferencií

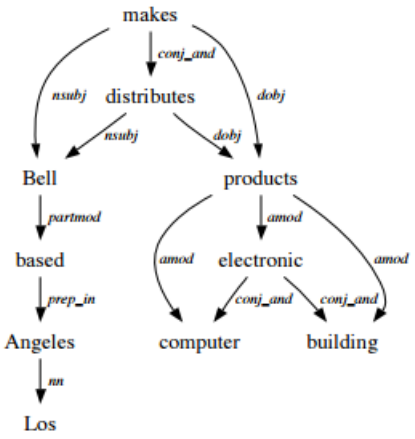
Nájdenie, identifikácia a rozpoznanie koreferencií v texte je úlohou rozpoznávania koreferencií (angl. Coreference resolution). V texte sa často používajú zámena (angl. pronouns) *to*, *tí*, *on*, anglicky *it*, *those*, *he* alebo menné frázy (angl. noun phrase). Tieto zámena a menné frázy sa odkazujú na iné podstatné mená alebo mená a názvy a je úlohou rozpoznávania koreferencií identifikovať referenciu na podstatné meno alebo meno, alebo názov, väčšinou entity z reálneho sveta, na ktoré sa odkazujú. Táto úloha spracovania prirodzeného textu sa využíva v aplikáciách (NLP) ako sú extrakcia informácií (viď. 2.2.1 Extrakcia informácií) a odpovedanie na otázky [5].

Príklad: **Martin Nemček** napísal túto bakalársku prácu. **On** študuje na FIIT STU BA.

Tu je vidno, že zámeno *on* sa odkazuje na meno *Martin Nemček*.

2.3.4 Rozloženie vzťahov

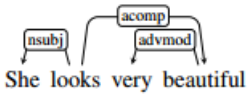
Rozloženie na vzťahy nám poskytuje jednoduchý opis gramatických vzťahov slov vo vete. Aplikovaním rozloženia vzťahov na vetu *Bell, based in Los Angeles, makes and distributes electronic, computer and building products.* vznikne strom vzťahov (angl. dependency tree) (viď. obrázok 1 Strom vzťahov) [6].



Obr. 1: *Strom vzťahov*

V tomto orientovanom stromovom grafe slová vety predstavujú vrcholy, pričom prechody medzi vrcholmi, hrany, sú reprezentované vzťahmi medzi nimi.

Ďalšia reprezentácia vzťahov zapisuje vzťahy priamo do vety. Na obrázku 2 Vzťahy vo vete vidíme, že medzi slovami *She* a *looks* je vzťah **nsubj** - nominal subject, medzi *looks* a *beautiful* je vzťah **acomp** - adjevtival complement, a v neposlednom rade medzi slovami *very* a *beautiful* je vzťah **advmod** - adverb modifier [6].



Obr. 2: *Vzťahy vo vete*

2.4 Nástroje na spracovanie prirodzeného jazyka

V súčastnosti je vyvinutých alebo sú vo vývoji viacero nástrojov, ktoré sa dajú použiť pri spracovávaní prirodzeného jazyka. Vývoj takýchto nástrojov je podporovaný na známych univerzitách ako sú napríklad Princeton, Stanford alebo Camridge ale samozrejme svoje slovo tu ma aj velikán Google. Pozrieme sa bli-

žšie na niektoré z týchto nástrojov, čo ponúkajú a ako sa dajú využiť.

2.4.1 WordNet

WordNet je databáza anglických slov vyvíjaná na Princetonskej univerzite. Databáza obsahuje podstatné mena, prídavné mená, slovesá a príslovky, ktoré sú zatriedené do synonymických sád, synsetov.

Tento nástroj je dostupný vo webovej verzií (viď. Obr. 3), ale ponúka stiahnutie aj jeho databázových súborov, ktoré, po splnení licenčných požiadaviek, sa dajú využívať v projektoch.

Slová do synetov sú zaraďované podľa významu. To znamená, že slová auto a automobil, ktoré sú zameniteľné vo vete, sú zaradené do rovnakého synsetu. WordNet v súčastnosti (r. 2015) obsahuje 117 000 synsetov. Každý z týchto sysnsetov taktiež obsahuje krátku ukážku použitia slova.

Vo WordNet-e sa nachádzajú aj vzťahy medzi slovami v zmysle nadradenosti. Tým sa mysli to, že stolička je nábytok a nábytok je fyzická vec a takto to pokračuje až po najvyššie slovo, od ktorého „dedia" všetky - entita (viď. Obr. 4. Okrem vzťahu nadradenosti WordNet obsahuje aj vzťah zloženia. Stolička sa skladá z operadla a nôh. Toto zloženie je typické len konkrétne slovo a neprenáša sa hore stromom nadradenosti, lebo pre stoličku je typické, že sa skladá z operadla a nôh, ale to už nie je typické pre nábytok. Prídavné mená obsahujú aj vzťah antonymity, takže slovo *suchý* bude prepojené so slovom *mokrý* ako so svojím antonymom.

Obr. 3: *Webové rozhranie*

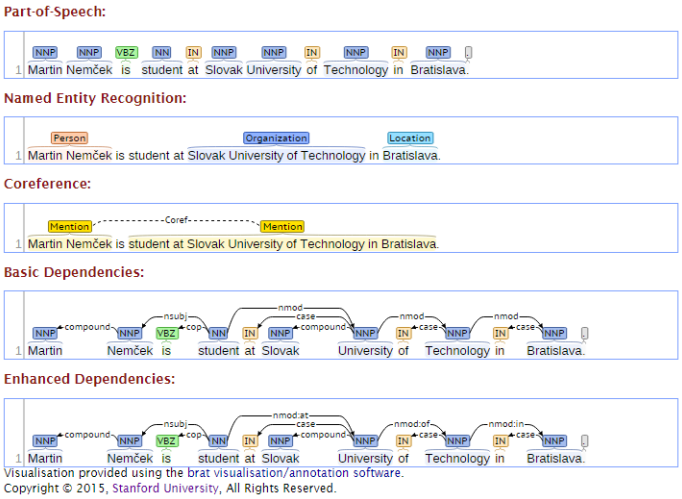
Obr. 4: *Nadradenost'slov*

2.4.2 StanfordNLP

Nástroj StanfordNLP je vyvíjaný na Stanfordskej univerzite. Skladá sa z niekoľkých softvérov, ktoré sa zameriavajú na úlohy spracovania prirodzeného jazyka popísané v sekcií 2.1 Spracovanie prirodzeného jazyka. Sú to softvéry *Stanford Parser*, *Stanford POS Tagger*, *Stanford EnglishTokenizer*, *Stanford Relation Extractor* a mnoho ďalších. *Stanford CoreNLP* zahŕňa viacero z týchto softvérov, a práve tento nástroj budeme používať pri spracovaní učebných textov.

Nástroje StanfordNLP sú implementované v Jave, ale sú dostupné aj v iných programovacích jazykoch ako C#, PHP alebo Python.

Dostupne je aj online webové demo. Na obrázku 5 StanfordNLP online demo vidíme výstupy z nástrojov ponúkaných balíkom StanfordNLP pre jednoduchý vstupný text skladajúci sa z jednej vety „Martin Nemček is student at Slovak University of Technologies in Bratislava.”.



Obr. 5: *StanfordNLP online demo*

2.4.3 CambridgeAPI

CambridgeAPI je vytvorený na Cambridge univerzite. Umožňuje prístup k viacerým rôznym slovníkom. Momentálne tento nástroj ponúka prístup k pätnástim prekladovým slovníkom ako napríklad anglicko-čínsky, anglicko-ruský, anglicko-

arabský, anglicko-japonský a ďalšie. Všetky prekladové slovníky majú primárny jazyk angličtinu. Slovenčinu v súčastnosti nepodporuje.

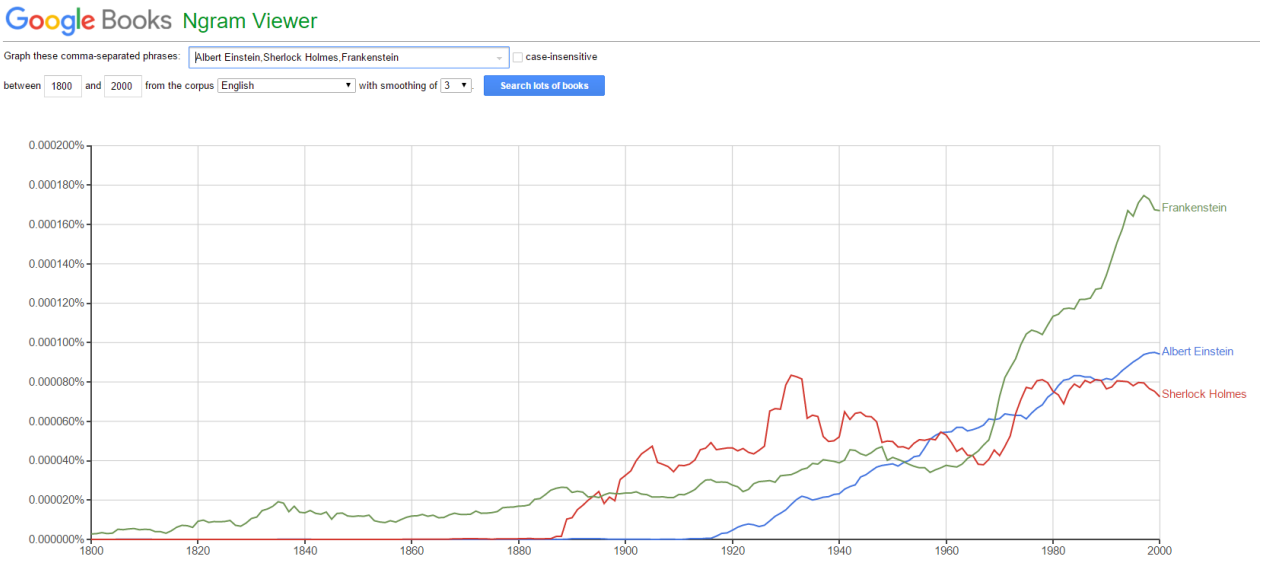
Spomínaný nástroj funguje na princípe dopytovania pomocou HTTP protokolu. Na obdržanie korektnej odpovede je potrebné mať osobný API kľúč. Ten sa dá získať kontaktovaním správcov CambridgeAPI.

2.4.4 Google Ngram

Google Ngram je postavený na ďalšom softvéry tohto giganta, Google Books. V knihách, napísané od roku 1500 až do súčastnosti, vyhľadáva výskyty n-gramov. Podporuje len niektoré jazyky, ako angličtina, francúzština, ruština, čínština a ďalšie. Na vyhľadávanie v knihách využíva optické rozoznávanie textu, pričom dokáže spracovať regulárne výrazy, pričom tie môžu byť použité iba ako náhrada celého slova, ale nie uprostred slova. Slovné spojenie „* Einstein" spracuje, pričom „Albert Einste*n" nie.

N-gram je postupnosť *n* za sebou idúcich slov alebo častí textu. *Martin* je n-gram veľkosti jedna, 1-gram alebo unigram. *Martin Nemček* je n-gram veľkosti dva, 2-gram alebo bigram a tak ďalej, pričom *n* môže byť ľubovolné kladné, celé číslo.

Google Ngram Viewer poskytuje vizualizáciu vyhľadaných dát. Je dostupný vo webovom rozhraní. Na obrázku 6 Google Ngram Viewer vidno vizualizáciu výskytu mien *Albert Einstein,Sherlock Holmes,Frankenstein* v knihách od roku 1800 do roku 2000.



Obr. 6: *Google Ngram Viewer*

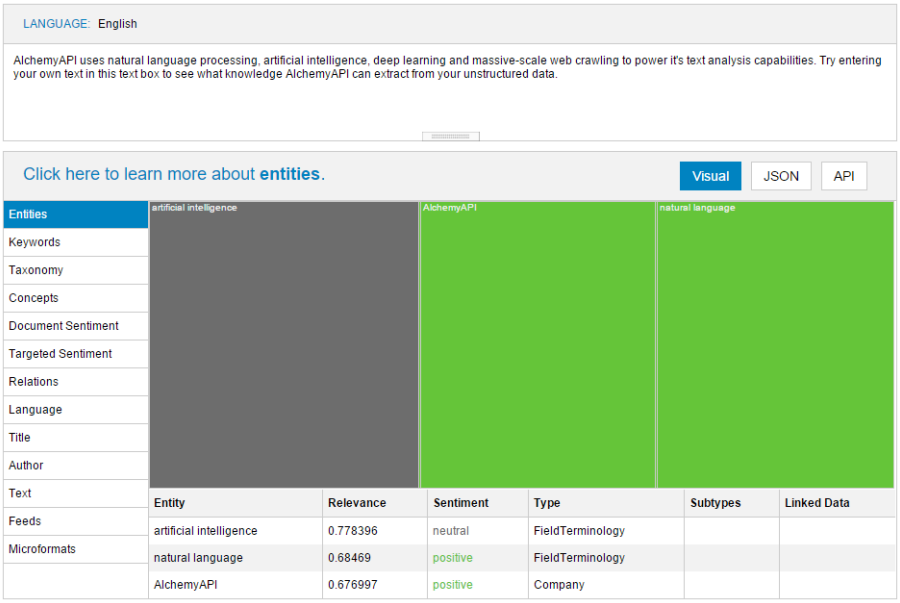
Tento nástroj okrem iného ponúka aj surové (angl. raw) dáta na stiahnutie.

2.4.5 AlchemyAPI

AlchemiAPI dvanásť funkcií, z ktorých sú niektoré zamerané na úlohy spracovania prirodzeného jazyka popísané v sekcií 2.1 Spracovanie prirodzeného jazyka, ako napríklad extrakcia entít, extrakcia kľúčových slov, extrakcia vzťahov, ale aj iné zaujímave funkcie, napríklad extrakcia autora z textu.

Na používanie tohto nástroja je potrebné sa zaregistrovať pre obdržanie API kľúču. S týmto kľúčom je tisíc dopytov denne zdarma. Dostupnosť v programovacích jazykoch je široká, kedže ponúka knižnicu v deviatich najpoužívanejších programovacích jazykoch.

Pre AlchemyAPI je dostupné aj online webové demo, viď obrázok 7 AlchemyAPI online demo, kde je vidno širokú ponuku, ktorú tento nástroj ponúka.



Obr. 7: *AlchemyAPI online demo*

Dáta sú vo formáte JSON a okrem spracovania prirodzeného jazyka AlchemyAPI ponúka aj nástroje na extrahovanie obsahu z obrázku alebo rozpoznávanie tvári na obrázkoch.

3 Smerovanie práce

V letnom semestri plánujem dokončiť prototyp. To znamená, spraviť používateľské rozhranie, ktoré bude umožňovať vložiť text na spracovanie, zobrazí poznámky a tak isto umožní používateľovi pre ľubovolnú vetu pozmeniť tvar poznámky poznámky. Tieto zmeny sa uložia do databázy a zohľadnia pri ďalšom použití.

Okrem dokončenia prototypu plánujem napísať všetky potrebné kapitoly a dokončiť tým celú prácu.

4 Opis prototypu

V zimnom semestri som implementoval prototyp aplikácie na spoznámkovanie učebného textu.

4.1 Notenizer

Notenizer je prototyp aplikácie na extrahovanie relevantných informácií z učebných textov. Využíva nástroj Stanford CoreNLP, ktorý je implementovaný v Jave, ale cez IKVM je portnutý aj na C#. Na ukážke 1 Spustenie StanfordCoreNLP je ukázané prepojenie nástroja StanfordCoreNLP s aplikáciou Notenizer.

```
String jarRoot = @"stanford-corenlp-3.5.2-models";

Properties properties = new Properties();
// Zvolime, ktore nastroje chceme pouzit.
// pos = part-of-speech tagger
// ssplit = sentence split
// atd.
properties.setProperty("annotators", "tokenize, ssplit, pos, parse");
properties.setProperty("sutime.binders", "0");
properties.setProperty("ner.useSUTime", "false");

// Nastavenie aktualneho priecinku, aby StanfordCoreNLP vedel najst
// vsetky potrebne subory
String currentDirectory = Environment.CurrentDirectory;
Directory.SetCurrentDirectory(jarRoot);
StanfordCoreNLP pipeline = new StanfordCoreNLP(properties);
Directory.SetCurrentDirectory(currentDirectory);

// Vytvorenie anotacie z textu
Annotation annotation = new Annotation(text);

// Spustenie
pipeline.annotate(annotation);
```

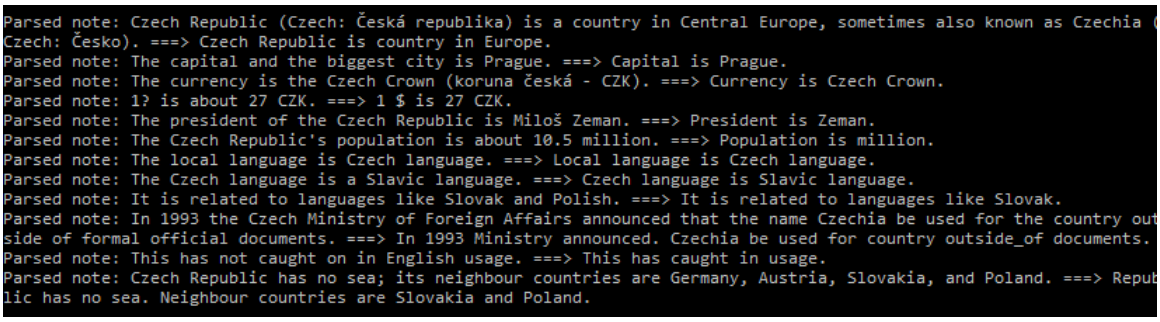
Ukážka 1: *Spustenie StanfordCoreNLP*

Údaje získane z tohto nástroja, napríklad POS značky, vzťahy medzi slovami, pozície slov a mnoho ďalších, Notenizer ďalej spracováva. Najdôležitejšie vlastnosti, ktoré sa využívajú v najväčšej miere pri spracovávaní sú závislosti (angl. dependency) medzi slovami vo vete.

Spracovávaný text sa postupne spracováva po vetách. Každá veta sa samostatne „rozparsuje", spoznámkuje. Vety sa parsujú na základe pravidiel. Na začiatku je daná statická sada pravidiel na spracovanie viet a textov. Po tom, ako sa celý text spracuje, tak sa použité pravidlá uložia do databázy aj s informáciami o pôvodnej vete a novo vytvorenej, zjednodušenej vete. Následne pri opätovnom používaní aplikácie, keď sa začne spracovávať text, tak sa vyhľadajú pre každú vetu pravidlá v databáze, vyberú sa tie s najväčšou zhodou a podľa toho sa spracuje daná veta. Statické pravidla na spracovanie vety sa v tomto prípade použijú len v prípade, ak v sa v databáze nenašli žiadne pravidlá na spracovanie vety, ktoré by pre danú vetu vyhovovali, to znamená, že takúto alebo podobnú vetu zatiaľ Notenizer nespracovával.

Na obrázku 8 Ukážkový výstup prototypu je ukázaný ukážkový výstup prototypu pre vstupný text z wikipédie: *Czech Republic (Czech: Česká republika) is a country in Central Europe, sometimes also known as Czechia (Czech: Česko). The capital and the biggest city is Prague. The currency is the Czech Crown (koruna česká - CZK). 1€ is about 27 CZK. The president of the Czech Republic is Miloš Zeman. The Czech Republic's population is about 10.5 million. The local language is Czech language. The Czech language is a Slavic language. It is related to languages like Slovak and Polish. In 1993 the Czech Ministry of Foreign Affairs announced that the name Czechia be used for the country outside of formal official documents. This has not caught on in English usage. Czech Republic has no sea; its neighbour countries are Germany, Austria, Slovakia, and Poland.*

Výstup je v tvare [pôvodná veta] ===» [poznámka z pôvodnej vety].



Obr. 8: *Ukážkový výstup prototypu*

4.1.1 Pravidlá

Pri spracovaní pôvodnej vety sa na túto vetu aplikuje *pravidlo na spracovanie*. Toto pravidlo obsahuje okrem iného zoznam závislostí slov vo vete. Podľa týchto závislostí slov vo vete sa spracovávanej vete vyhľadajú slová, ktoré majú byť použité v poznámke. Vyhľadávajú sa podľa okrem iného podľa POS značiek a indexov vo vete.

4.1.2 Ukladanie a hodnoty pravidiel

Po spracovávaní sa do databázy uložia použité pravidla aj s ostatnými informáciami o pôvodnej a novej vete. Ukladá sa

- Hodnota pôvodnej a novej vety
- Zoznam indexov slov za ktorými bola v poznámke ukončená veta (ak pôvodná veta je súvetie, tak z nej môže vzniknúť viacero poznámok)
- Všetky závislostí slov v pôvodnej vete
- Všetky závislosti slov v poznámke

pričom každá závislosť slov vo vete sa skladá z

- Názvu závislosti
- Hodnota governora závislosti

- Hodnota dependenta závislosti
- Pozícia závislosti vo vete
- POS značka a index slova pre governora aj dependenta

4.1.3 Vyhľadanie pravidla

Pri spracovávaní vety sa v prvok kroku pozrie do databázy a vyhľadá sa pravidlo na spracovanie tejto vety.

Pravidlo sa v databáze vyhľadáva podľa nasledovných podmienok.

1. Počet závislostí vety
2. Názvy závislostí vety

Veta v databáze, ktorej pravidlo chceme použiť, musí spĺňať tieto dve pravidlá. Musí mať rovnaký počet závislostí vo vete ako aktuálne spracovávaná veta a taktiež názvy všetkých závislostí musia sedieť.

Avšak pri týchto podmienkach môže nastať situácia, kedy pre aktuálne spracovávanú vetu bude vyhovovať viacero viet z databázy. V tomto prípade určujeme pravidlo vety s najväčšou zhodou a to sa následne aplikuje.

Zistenie najväčšej zhody ma viacero krokov. Najskôr sa spočítavajú zhody POS značiek governerov a dependentov a indexy slov nezávisle od seba, čiže sa len zisťuje, či spracovávaná veta obsahuje závislosť s nejakou hodnotou indexu, alebo governora, atď. V druhom kroku sa spočítavajú zhody POS značiek a zároveň aj indexov slov vo vete. To znamená, že sa zisťuje, či sa v spracovávanej vete nachádza napríklad závislosť, ktorej governor ma hodnotu *car* a zároveň má index hodnotu 3. V poslednom, treťom, kroku sa zisťuje, či sa závislosti zhodujú na všetkých hodnotách, čiže governor a jeho hodnota a index a dependent a jeho hodnota a index *súčasne*.

Tieto tri hodnoty sa na záver spočítajú a tým získame percentuálne ohodnotenie zhody viet.

5 Design

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat.

5.1 Subsection

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Nam liber tempor cum soluta nobis eleifend option congue nihil imperdiet doming id quod mazim placerat facer possim assum. Typi non habent claritatem insitam; est usus legentis in iis qui facit eorum claritatem. Investigationes demonstraverunt lectores legere me lius quod ii legunt saepius. Claritas est etiam processus dynamicus, qui sequitur mutationem consuetudium lectorum. Mirum est notare quam littera gothica, quam nunc putamus parum claram, anteposuerit litterarum formas humanitatis per seacula quarta decima et quinta decima. Eodem modo typi, qui nunc nobis videntur parum clari, fiant sollemnes in futurum.

6 Results

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat.

6.1 Subsection

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Nam liber tempor cum soluta nobis eleifend option congue nihil imperdiet doming id quod mazim placerat facer possim assum. Typi non habent claritatem insitam; est usus legentis in iis qui facit eorum claritatem. Investigationes demonstraverunt lectores legere me lius quod ii legunt saepius. Claritas est etiam processus dynamicus, qui sequitur mutationem consuetudium lectorum. Mirum est notare quam littera gothica, quam nunc putamus parum claram, anteposuerit litterarum formas humanitatis per seacula quarta decima et quinta decima. Eodem modo typi, qui nunc nobis videntur parum clari, fiant sollemnes in futurum.

7 Conslusions

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Morbi sit amet arcu. Fusce pharetra dapibus elit. Duis malesuada. Proin at elit vitae quam cursus tristique. Quisque fermentum. Praesent dictum. Nullam vehicula. Nunc pharetra dolor ut velit. Sed pulvinar, est sed congue tempor, nibh arcu cursus enim, quis consequat magna lacus sed pede. In sagittis. Etiam volutpat, velit id tincidunt egestas, augue ligula auctor eros, sit amet viverra sapien tortor at odio. In diam libero, fringilla ut, adipiscing condimentum, ultricies at, dui. Phasellus vitae risus.

Pellentesque vulputate ante ut diam. Sed adipiscing malesuada odio. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Nam a leo. Praesent velit. Aenean vehicula accumsan quam. Nulla dolor lorem, imperdiet a, ullamcorper hendrerit, ultrices at, urna. Integer placerat ligula id purus. Sed id nisl. Pellentesque tincidunt neque in lacus. In non quam et felis suscipit viverra.

Literatúra

[1] Collobert, Ronan and Weston, Jason and Bottou, Léon and Karlen, Michael and Kavukcuoglu, Koray and Kuksa, Pavel. *Natural Language Processing (Almost) from Scratch*. JMLR.org, 2011.

[2] Ann Copestake *8 Lectures, Natural Language Processing* 2004.

[3] Akshar Bharati, Vineet Chaitanya, and Rajeev Sangal *Natural language processing: a Paninian perspective* MIT Press Cambridge, MA, USA, 1995.

[4] Preeti and BrahmaleenKaurSidhu *Natural Language Processing* A Vinitha et al, Int.J.Computer Technology & Applications,Vol 4 (5),751-758

[5] Volha Bryl, Claudio Giuliano, Luciano Serafini, and Kateryna Tymoshenko *Supporting natural language processing with background knowledge: coreference resolution case* Trento, Italy.

[6] Marie-Catherine de Marneffe and Christopher D. Manning *Stanford typed dependencies manual* September 2008

A Technical documentation

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat.

A.1 Implementation

Modul abc

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Nam liber tempor cum soluta nobis eleifend option congue nihil imperdiet doming id quod mazim placerat facer possim assum.

Modul def

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Nam liber tempor cum soluta nobis eleifend option congue nihil imperdiet doming id quod mazim placerat facer possim assum. Typi non habent claritatem insitam; est usus legentis in iis qui facit eorum claritatem. Investigationes demonstraverunt lectores legere me lius quod ii legunt saepius. Claritas est etiam processus dynamicus, qui sequitur mutationem

consuetudium lectorum. Mirum est notare quam littera gothica, quam nunc putamus parum claram, anteposuerit litterarum formas humanitatis per seacula quarta decima et quinta decima. Eodem modo typi, qui nunc nobis videntur parum clari, fiant sollemnes in futurum.

B User documentation

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat.

B.1 Instalation

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi.

B.2 Run the application

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi.

C Electronic medium

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat:

/Application

- implementácia opisovaného riešenia

/Documentation

- bakalárska práca spolu s anotáciami v slovenskom a anglickom jazyku

/Documentation/Latex

- latex zdrojové súbory dokumentácie

/Documentation/BibTeX

- BibTeX súbor s použitými referenciami

/Documentation/Resources

- dostupné použité zdroje

/Resources

- vstupne/testovacie dáta opisované v dokumente

/Source/Dependencies

- inštalačné súbory pre knižnice, ktoré potrebuje aplikácia

read.me - popis obsahu média v slovenskom a anglickom jazyku