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

1 Úvod **1**

2 Analýza **3**

2.1 Spracovanie prirodzeného jazyka 3

2.1.1 Značkovanie slovných druhov 4

2.1.2 Rozpoznávanie názvoslovných entít 4

2.1.3 Rozpoznanie koreferencií 5

2.1.4 Rozloženie vzťahov 5

2.2 Enumeration 6

2.3 Itemization 6

2.4 Citation 6

2.5 Labesl & References 6

2.6 Examples 6

3 Design **8**

3.1 Subsection 8

4 Results **9**

4.1 Subsection 9

5 Conslusions **10**

Literatúra **11**

A Technical documentation **12**

A.1 Implementation 12

B User documentation **14**

B.1 Instalation 14

B.2 Run the application 14

C Electronic medium **15**

Zoznam obrázkov

1 Name figure 5

Zoznam tabuliek

Zoznam ukážok

1 Example 1 6
2 Názov 7

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, kde a ako sa dá využiť. Ďalej zanalyzujem existujúce aplikácie a systému, ktorých základom je spracovanie textu.

2.1 Spracovanie prirodzeného jazyka

Spracovanie prirodzeného jazyka (angl. Natural Language Processing - NLP) je oblasť vedného oboru, ktorý sa zaoberá počítačovými technológiami. Hlavným cieľom NLP je dosiahnuť aby počítač dokázal porozumieť ľudskej, či už písanej alebo hovorovej, reči.

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 z obtiažnosti problému vytvorenia počítača inteligentného ako človek, takzvané „true AI".

NLP ma niekoľko hlavných úloh. Podrobnejšie si priblížime tie, ktoré sú relevantné vzhľadom na implementáciu spracovania učebných textov.

Úlohy spracovania prirodzeného textu: [Natural Language Processing (Almost) from Scratch, Ronan Collobert]

- Značkovanie slovných druhov (angl. Part-of-speech tagging) 2.1.1
- Rozdelenie vety na menšie časti (angl. Chunking)
- Rozpoznávanie názvoslovných entít (angl. Named Entity Recognition) 2.1.2

- Označovanie sémantického postavenie (angl. Semantic Role Labeling)
- Rozpoznanie koreferencií (angl. Coreference resolution) 2.1.3
- Morfologické segmentovanie (angl. Morphological Segmentation)
- Generovanie prirodzeného jazyka (angl. Natural Language Generation)
- Optické rozoznávanie textu (angl. Optical Character Recognition)
- Rozloženie vzťahov (angl. Dependency parsing) 2.1.3
- a mnoho ďalších

2.1.1 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. 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*.

2.1.2 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*.

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.

2.1.3 Rozpoznanie koreferencií

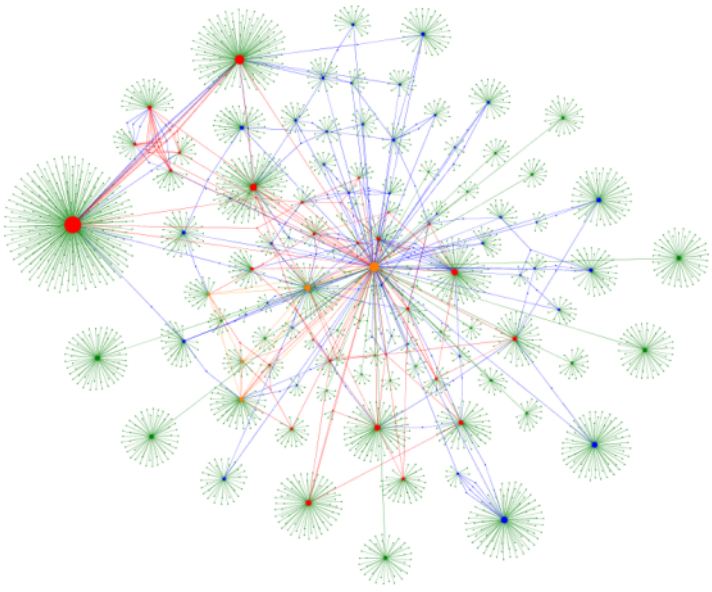
Nájdenie, identifikácia a rozpoznanie koreferencií v texte je úlohou rozpoznávania koreferencií (angl. Coreference resolution). V texte sa často používajú zámena (angl. pronouns) *to*, *tí*, *on* alebo anglicky *it*, *those*, *he* a mnoho ďalších. Tieto zámena sa odkazujú na iné podstatné mená alebo mená a názvy a je úlohou rozpoznávania koreferencií určiť, na ktoré podstatné meno alebo meno, alebo názov sa konkrétne zámeno odkazuje.

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*.

2.1.4 Rozloženie vzťahov

Dummy text..



Obr. 1: *Name figure*

2.2 Enumeration

1. goal 1
   (a) goal 1.a
   (b) goal 1.b
2. goal 2
3. goal 3

2.3 Itemization

- item 1
  - item 1.1
  - item 1.2
- item 2
- item 3

2.4 Citation

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat [1].

2.5 Labesl & References

See Section 2.6.

2.6 Examples

```
<table class="metric_index">
  <tr>
    <th>Lines of code</th>
    <th>Value</th>
  </tr>
  <% if (filenum and modulenum) then %>
    <tr>
      <td class="name">Number of files</td>
      <td class="value"><%=filenum%></td>
    </tr>
```

```
    <tr>
      <td class="name">Number of modules</td>
      <td class="value"><%=modulenum%></td>
    </tr>
    <tr>
  <% end %>
  <tr>
    <td class="name">Lines Total</td>
    <td class="value"><%=LOC.lines%></td>
  </tr>
  <!--
              skryty zdrojovy kod
    podobne zobrazenie ostatnych metrik riadkov
  -->
</table>
```

Ukážka 1: *Example 1*

```
local parser  = require 'leg.parser'
local rules = require 'metrics.rules'
-- << skryty zdrojovy kod >> --
local capture_table = {}
grammar.pipe(LOC_capt.captures, AST_capt.captures)
grammar.pipe(block_capt.captures, LOC_capt.captures)
-- << viacero rovnakych volani s tabulkami captures inych modulov >> --
grammar.pipe(capture_table, cyclo_capt.captures)
local lua = lpeg.P(grammar.apply(parser.rules, rules.rules, capture_table))
local patt = lua / function(...)
  return {...}
end
local result = patt:match(code)[1]
```

Ukážka 2: *Názov*

3 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.

3.1 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.

4 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.

4.1 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.

5 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] Roberto Lerusalimschy, Luiz Henrique de Figueiredo, and Waldemar Celes. *Lua 5.1 Reference Manual*. Lua.Org, 2006.

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