Prezentácia projektových zadaní

Možnosť vlasnej témy, môže súvisieť s diplomkou alebo bakalárkou

Zameranie: spracovanie textových dát za účelom vyhľadávania alebo extrakcie informácií

DBPedia spootlight

naprogramovat NE extraction pomocou DBPedia spootlight

Regex editor

vytvorenie jednoducheho editora na regularne vyrazy ked bude moct uzivatel cez Java aplikaciu testovat regularne vyrazy na nacitanom txt subore.

Vypisanie extrahovanych udajov aj skupin (groups)

Podla moznosti integracia z Ontea nastrojom.

Integracia makier z ontea aj s prikladmi textov a spustenim na nich v description. Vytvaranie a editacia makier pomocou nastroja.

Štatistický prekladač

Na základe rôznych jazykových verzií stránok vytvoriť prekladač na základe štatistických údajov.

Existuje napr9klad package Moses.

Alebo vytvorenie prekladača slov na základe spracovania anchor textov.

Je možné riešiť nasledovné projekty:

- prekladač slovenčina <=> angličtina na základe anchor textov
- prekladač slovenčina <=> čeština na základ hociakých textov. Pri podobných jazykoch by mohol byť menší problém s tým že nevieme rozpoznať slovné druhy (POS taging)

Gazetteer

Vytvorenie podobného gazeteera (slovnika) pre information extraction ako ma GATE.

Vlastnosti:

- lineárna zlozitosť (iba jeden prechod textom)
- definovanie oddeľovača slov (tokenizatora), najlepšie po znakoch, pričom hľadanie zhody začína vždy za white space znakom
- generovanie aliasov pomocou regularneho vyrazu. Napr. v slovniku je "Meno Priezvisko" a bude hľadať aj "M. Priezvisko"

Podpora slovenskeho vyhladávania

- analyzer ktorý rieši diakritiku (napr vyhadzuje), spellcheck aj bez diakritiky - urobiť nad lucene

Extrakcia faktov

Extrakcia faktov zo slovenského webu alebo iných textových dokumentov. Nieco na spôsob knowItAll ale pre slovenčinu. Napríklad zo stránky FIIT vytiahnuť zoznam učiteľov, predmetov alebo študentov. Možno jednoduchý príklad. Proste vytiahnuť fakty ktore sa nedajú nájsť z jedného dokumentu. Iný príklad je napríklad vytiahnuť zoznam sklenárstiev v Bratislave.

- zoznam ľudí zo stránok ústavou SAV, automatické porovnanie so SAV.sk

Tvorba vyhľadávača kontaktov v mobile a iné

- aplikácia na spôsob písania SMS pomocou T9. Teda niečo kde termy v indexe budú vlastne čísla.
- Može to byť vyhľadávanie kontaktov alebo menších dokumentov pomocou zadania query cez cisla reprezentujúce písmená.

Focused Crawler

- sťahovanie dynamicky generovaných stránok tak aby sa nesťahoval rovnaký obsah (inak utriedený zoznam, verzie na wiki stránke, printová verzia stránky a podobne)
- sťahovanie toho istého obsahu v inej jazykovej verzii (tvorba corpusu pre štatistický prekladač)

Rozpoznávanie slovných druhov

Part of Speach Tagging (POS) taging pre slovenčinu. Založené na slovníkovom princípe z dostupných slovníkov (aspell, ispell, OpenOffice a pod.) alebo založený na štatistickom princípe. Možné poyrieť princíp OpenNLP. Netreba všetky slovné druhy ani nemusí 100 percentne fungovať.

Fazetový a fultextový prehliadač

Kto má prístup k databáze nejakých produktov alebo nejaký rozsiahlejší web. Treba urobiť fultext toho webu s kombináciou fazetového prehliadača. (Podobne ako na amazon.com a iných)

Je to možné urobiť pomocou systému Apache Sorl. Ide najmä o konfiguráciu a napojenie systému a jeho odladenie.

Spread Activation

spread activation algoritmus (pozri wikipediu) treba naprogramovať a použiť na nejakých dátach. Napr na extrakcii z emailov.

Tag Cloud

Generovanie Tag cloudu (pozri wikipediu) z webstránky a jej podstránok.

Teória okolo TF-IDF, stop slová, lematizácia....

Name Entity Recognition - Machine Learning

Rozpoznávanie mien (osoby, mesta, organizácie, ...) pomocou OpenNLP

2 projekty - jeden rozpoznavanie na Slovenskom a anglickom texte, druhý trenovanie na Slovenskom

Name Entity Recognition - Extrakcia Anchor Text

Pomocou extrakcie anchor textov liniek (text v ramci Tagu <a href>TEXT) robiť Named Entity recognition a Aliasy.

Analyzer ktorý vyhodí diakritiku

použitie vo vyhľadávači emailov alebo fajlov na disku spolu so spell checkom.

Advanced email search

Vyhladavac pomocou socialnej siete-grafu extrahovaneho z emailu.

Pomocou existujuceho softveru (Ontea alebo acoma = emailSocNet) sa extrahuje graph. Nad nim urobit vyhladavanie pomocou spread activation ?+lucene kde budu fazety podla objavenych typov objektov.

Porovnanie implementacii spread activation algoritmov

Z minuleho roka treba porovnat implementacie spread activation (3 projekty - Fridrich, Blazko, Marton) a zistit v com sa lisia a ako je najvyhodnejsie implementovat spread activation na Jung kniznici tak aby bol algoritmus vseobecny a pouzitelny

Vyhľadávanie s využitím anotácií (tagov)

V systémoch ako delicious.com, twiter alebo youtube uzivatelia generujú množstvo tagov ktoré v kombinácii s klasickými technikami vyhľadávanie (indexovanie) možu priniesť lepšie výsledky.

Úlohou je vytvoriť vyhľadávač ktorý tieto tag-y využije.

Extrakcia udalostí

extrahovanie udalostí z emailov alebo webových stránok.

- dátum, čas
- miesto
- názov udalosti