Spracovanie dokumentov, spracovanie metadát, extraxcia kľúčových slov z textu

Prezentácia aktuálneho stavu spracovania problému

Bc. Ján Chlpek

Tímový projekt II, d9vina

Marec 2011

Zhrnutie

- spracovanie TXT súborov – **OK**
- spracovanie PDF súborov – **OK**
- spracovanie DOC súborov – **OK**

- aj metadáta – autor, dátum, kľúčové slová

- extrakcia kľúčových slov z textu – **OK**

Spracovanie dokumenov

- Spracovanie **TXT, DOC, PDF**
  - V rámci spracovania sa v pôvodnej implementácii volala externá aplikácia ConvertDoc spušťaná s parametrami (vstup, výstup, možnosti). Keďze to bola trial verzia, bolo nutné hľadať alternatívy. Vymenené za **Java knižnice + Apache Tika – OK**
  - Problémy s kódovaním PDF – vyriešené – **OK**

Spracovanie textu

- Rozpoznanie jazyka dokumentu (**slovenský, anglický, nemecký**) – kvôli extrakcii keywords - **OK**
- Odstránenie stop slov a čísel **OK**
  - vlastný zoznam slovenských, anglických a nemeckých stop slov - **podpora 3 jazykov - OK**
  - odstraňovanie stop slov - dva krát. **OK**
    - 1. pred lematizáciou, aby sme zrýchlili proces lematizácie (aby neboli vyhľadávané základné tvary stop slov)
    - 2. po lematizácii (odstránenie stop slov prevedených do základného tvaru)
  - odstránenie čísel, pretože ponechanie čísel je nevhodné (číslované zoznamy...) - **OK**
- Lematizácia pre slovenský jazyk - **OK**
  - zmena slov do základného (slovníkového tvaru) - potreba slovníka - súbor form2lemma.cdb (pochádza z JÚĽŠ SAV - Garabík). práca s týmto databázovým súborom – využitie manuálu k CDB a konkrétne použitie knižnice strangegizmo.cdb.

Výber kľúčových slov, metadáta

- Početnosti slov – TF (term frequency) - **OK**
- Výber kľúčových slov – **MOŽNOSŤ - nastaviteľný “počet” KEYWORDS**
  - vybraté slová, ktorých početnosť v dokumente nie je menšia ako konštanta (0,5) * početnosť najviac vyskytujúceho sa slova. Nastaviteľná metóda.
- Získavanie informácií z metadát
  - Využitie knižnice ICEpdf, Apache Tika
    - potrebné informácie – autor, dátum, keywords **– OK**
  - Keywords z metadát, keywords z textu – zjednotenie, zoradené podľa relevantnosti, priorita – keywords z metadát, pretože boli zadané manuálne

Ďakujem za pozornosť.

- xchlpek [at] gmail [dot] com