

Porovnávanie dokumentov s využitím znalostí na pozadí

Martin Kováčik

mato.kovacik@gmail.com

<http://student.fiit.stuba.sk/~kovacik03/dipl.html>

Obsah

- ◆ Prezентácia projektu
 - ◆ Zvolený prístup
 - ◆ Metódy
 - ◆ Testovanie
- ◆ Cieľe projektu
 - ◆ Smerovanie projektu nasledujúci semester

Prezentácia obsahu projektu

***Porovnávanie dokumentov s
využitím znalostí na pozadí***

Motivácia (1)

- ◆ Dokument
 - ◆ Väčší rozsah
 - ◆ Slabo štruktúrované informácie
 - ◆ Prirodzený jazyk
- ◆ Prirodzený jazyk
 - ◆ Kontextový
 - ◆ Viacznačný
 - ◆ Jeden fakt je možné vyjadriť viacerými spôsobmi

Motivácia (2)

- ◆ Porovnávanie nie je invariantny pojem
- ◆ Cieľom je konceptuálna podobnosť
- ◆ Vyžaduje sa identifikácia konceptuálnych prvkov v dokumente
- ◆ Vytvorenie reprezentácie dokumentu, ktorá zmierňuje:
 - ◆ Chýbajúcu štruktúru
 - ◆ Kontextovosť a viacznačnosť prirodzeného jazyka

Aplikovaná metóda porovnávania

- ◆ Aplikované metódy chápu dokumentu ako súbor znakov
- ◆ Porovnávanie : Abstraktná postupnosť
 - ◆ Filtrovania znakov
 - ◆ Normalizácie znakov
 - ◆ Vytvorenie vektorovej reprezentácie
 - ◆ Znižovanie entropie identifikáciou latentných štruktúr
 - ◆ Vyhodnotenie podobnosti vektorovej reprezentácie

Znak dokumentu

- ◆ Znak reprezentuje vlastnosť, ktorá sa vyskytuje v dokumente
- ◆ Najjednoduchší znak dokumentu je slovo
- ◆ Zložitejšie znaky:
 - ◆ N-Gramy – sekvencie n slov
 - ◆ Itemsety – skupiny slov, spĺňajúce pravidlo
napr. prídavné meno – podstatné meno
 - ◆ Iné štruktúry ...

Znalosti vyžadované procesom porovnávania

- ◆ Znalosti vystupujúce v procese porovnávania
 - ◆ Filtrovania znakov ← ***Slovníky***
 - ◆ Normalizácie znakov ← ***Slovníky, pravidlá***
 - ◆ Vytvorenie vektorovej reprezentácie
 - ◆ Znižovanie entropie ← ***Transformácie***
identifikáciou latentných ***analýzou korpusu***
štruktúr
 - ◆ Vyhodnotenie podobnosti vektorovej
reprezentácie

Proces porovnávania (1)

- ◆ Filtrovanie znakov, ich vytváranie
- ◆ Normalizácia
 - ◆ Využitý je aplikačný rámec Apache Lucene
 - ◆ Filtrovanie je realizované na základe obsahovej informácie slova – spojky, predložky, jazykové prvky sú odstránené
- ◆ Vektorová reprezentácia
 - ◆ Bag-Of-Words
 - ◆ Znak dokumentu reprezentuje dimenziu, početnosť znaku hodnotu dimenzie

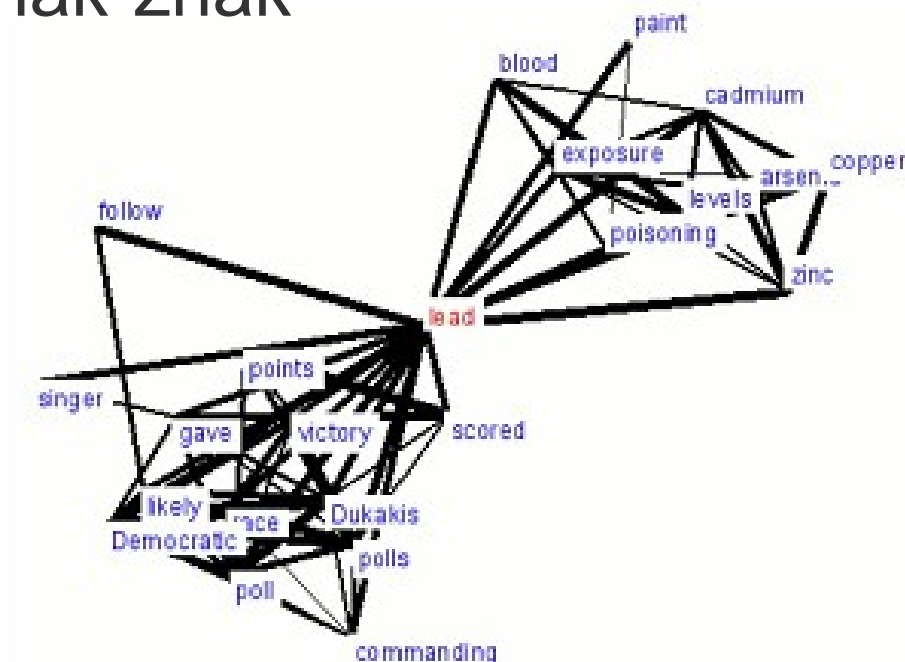
Proces porovnávania (2)

- ◆ Identifikácia latentných znakov
 - ◆ Latent Semantic Analysis – vzťah dokument-znak

Príklad:

(car, motor, wheel) $\rightarrow$ $(0.7 \cdot \text{car} + 0.6 \cdot \text{motor} + 0.4 \cdot \text{wheel})$

- ◆ Vzťah znak-znak



Proces porovnávania (3)

- ◆ Vyhodnotenie podobnosti

- ◆ Vzdialenostné metriky
Cosine, Euler, Manhattan, ...

- ◆ Information Theoretic

$$ItSim(A, B) = \frac{I(common(A, B))}{I(desc(A, B))}$$

Vyhodnocovanie úspešnosti (1)

- ◆ Problematické určenie podobnosti
- ◆ Zaradenie dokumentov do kategórie je deterministickejšie
- ◆ Predpokladom je využitie korpusov pre vyhodnocovanie klastrovacích algoritmov
 - ◆ Reuters-21578

Vyhodnocovanie úspešnosti (2)

- ◆ Porovnávanie je aplikované na všetky dvojice testovacích dokumentov
- ◆ Priemerná presnosť – presnosť je vyhodnocovaná pre korpus, nie pre dvojicu dokumentov
- ◆ Podobnosť na základe prieniku kategórií porovnávaných dokumentov

$$precision(p_k) = \frac{\#(\forall p_j: j < k \wedge ITP(p_j))}{k}$$

Ciele projektu

Ciele projektu

- ◆ Práce nasledujúci semester
 - ◆ Dokončenie implementácie procesu
 - ◆ Implementácia jednotlivých krokov
 - ◆ Vytváranie špecifických konfigurácií procesu porovnávania – kompozícia prezentovaných metód
 - ◆ Nástroj na vyhodnocovanie porovnávacieho procesu
 - ◆ Nástroj na vytváranie znalostí