Generalized Term-Vector Similarity Evaluation Framework

Bc. Martin Kovácik (kovacik03@student.fiit.stuba.sk), supervisor: Mgr. György Frivolt

Applications

Clustering Algorithms

- Taxonomy development
- Identification of major themes and sub-themes
- Navigation through document databases

Search Engine Enhancements

- Relevance ranking
- Discovering of relevant items, that are do not match query
- Recommendations based on similarity

Document Model

Basic Idea

Words describing similar topics are present in similar documents

Term-Vector Representations (Bag-Of-Words)

- highlights presence of words and occurrence probability
- discards word position

...But on Saturday Darwin walked in to a police station in central London. Darwin told police that he did not remember where he had been for last five years...

Language Processing Problems

Polysemy - ambiguity of an individual word or phrase that can be used (in different contexts) to express two or more different meanings

Synonym - two words that can be interchanged in context are said to be synonymous relative to that context

Inflection - a change in the form of word (usually by adding suffix) to indicate a change in its grammatical function

Noise - words with low influence on topic identification, but they occur very often - ***Stop-words***.

Saturday darwin walk police station ...

d = (1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1)

1. Tokenization

... {but}, {on}, {Saturday}, {darwin}, {walked} ...

2. Stemming

- Dictionary based stemming
- Porters' Steemer

... {walked→walk}, {told→tell} ...

3. Stop-Word Filtering

- Dictionary based
- Occurrence statistics

... ~~{but}~~, ~~{on}~~, {Saturday}, {darwin}, ~~{that}~~, ~~{in}~~ ...

4. Occurrence count

- Term Frequency (TF), Word Count
- Inverse Document Frequency (IDF)
- TF.IDF

5. Dimensionality Reduction

- Latent Semantic Analysis (uses SVD)

$X = U \Sigma V^T \rightarrow \quad X_k = U_k \Sigma_k V_k^T$

term-document matrix — k-rank term-document matrix

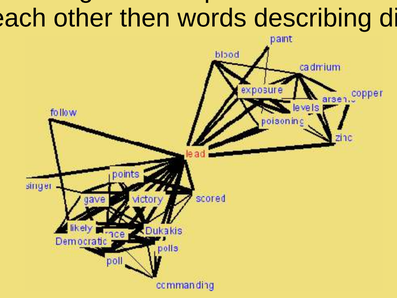
$d^T' = \Sigma_k^{-1} U_k d^T = (0.1, 0.3, \ldots)$

0.4 x d[police] + (-1) x d[station] + ...

- Transformation to Probabilistic Word Cluster Space

Distributional Hypothesis:

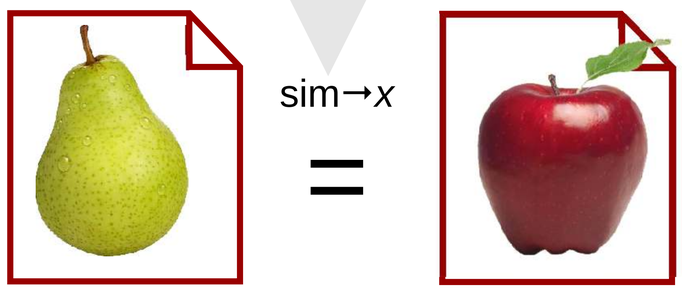
Words describing similar topic occur more significantly close to each other then words describing different topic.



6. Similarity Evaluation

- Cosine Similarity
- Information Theory Similarity

$$ItSim(A,B) = \frac{I(common(A,B))}{I(desc(A,B))} = \frac{2\sum_t min\{p_{A,t}, p_{B,t}\} log\pi(t)}{\sum_t p_{A,t} log\pi(t) + \sum_t p_{B,t} log\pi(t)}$$



Testing

Datasets

- Mapekus ACM Ontology
  - only metadata
  - wrapper for PDF downloading
  - approx. 38 000 downloadable papers
- Reuters-21578
  - reduced version - with categories

Performance Evaluation Method

- Two documents, that have same category assigned are similar

$$precision(p_k) = \frac{\#\ of\ \ intratopic\ \ pairs\ \ p_j\ \ for\ \ all\ \ j < k}{k}$$

- p_k - k-th most similar document pair
- Average precision for whole corpus is calculated
- Shows relative improvements over reference method

More info : http://student.fiit.stuba.sk/~kovacik03/dipl.html

This work was partially supported by scientific Grant Agency of Slovak Republic, grant No. VG1/3102/06