Generalized Term-Vector Similarity Evaluation Framework U • •

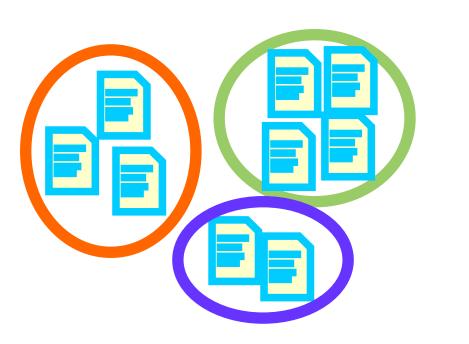
Bc. Martin Kovácik (kovacik03@student.fiit.stuba.sk), supervisor: Mgr. György Frivolt

Clustering Algorithms

T 0

00

- 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

Service and a service transfer service the threadest me has			Sand Fault - Had
The React Taylor Buylor upper Tealury Factory			Libe Long Louis
🐗 - 🕪 - 🎉 😳 🏠 🖓 tapajo ang sang sang sang sang sang sang sang		· > G · spik-	A - 10 - 20 -
Web mages Mare News Sharping and maxie		Sup =	IEEE@comp
Google Front and	Sever t Several Land		arra i vicni abriri
Web Blaze Ernes Colocal	Sestis 1, 10 Ja	e ou. i 3 540.000 (cr. <u>semantic web</u> . :0.16 selonde:	site sea.
Bernantic Web - Weipedia, the free encyclopedia Semantic Web - Aviapedia, the free encyclopedia Semantic Web - Aviapedia, the free encyclopedia Semantic Web - Aviapedia, the Semantic Web - Sector - Avia WSC Semantic Web Activity Semantic Web over La acon - or homework list, the WSC semantic Web reading Semantic Web readman Million - Aviapedia, the Common Common Semantic Web readman Million - Aviapedia, the Common Common Semantic Web readman Million - Aviapedia, the Common Common Semantic Web readman Million - Aviapedia, the Common - Avia Semantic Web - An Incode clien Mysteric Information - Avia Man Page - Semantic Web - An Incode clien Mon Page - Semantic Web - Avia Novel Semantic Web - Avia Novel Semantic Web - Avia Novel Semantic Web - Avia Man Page - Semantic Web - Avia Novel Semantic Web - The Voice of Semantic Web on Novel Semantic Web - The Voice of Semantic Web on Novel Semantic Web - The Voice of Semantic Web The Semantic Web The Semantic Voice - The Semantic Web The Semantic Web The Semantic Voice - The Semantic Web The Semantic Web The Semantic Voice - The Semantic Web The Semantic Web The Semantic Voice - The Semantic Voice - The Semantic Voice - The	e Wale nich en frei sem tahter er 1. onnes 2. Jake is be char od endrikt sole 2. Jake is be char od endrikt sole 2. Jahr 2. Jahr 3. De mei 3. De mei 3. De mei 3. De mei 3. De mei 3. De mei	Elemental Web Platform Transford - Disk (Rest) - annex Transford - Disk (Rest) - annex Tran	Escret fori Jathatti 2 Jat Solt By: Theorem Constant Solution Soluti
se nantickeb confectif 44-30 chorters manpagat		🖬 🕼 ig milan di 🖌 Tato transfera 🛞 🕸 🛒	* S -4 a

innet Sandr-Madis Tudas × Lanst Concert _ Ant Digly (Ante: Concert.	
n ganar ganati jawa wago ganati ganarki 1 - 10 - 20 ²⁰ - 10 ² - 10	e an Older e
	The second s
EE@computer society	
me plottilubring und some Cutrian und eine som skotternistet, de Loom	令 IEE
ite search Trans	
тейн хамай махалаас жилээ нэхэг хэлээ нэхэнхэг хөвөх 1 хорон хамахас холон хайос соорон хайос халас нэхэг	
Denne Toris Internet teate And the second and the s	estatus reportas faced vende e d avez an branchas para esta estas relianda or otra secto proceso no estas estas contra estas contra estas contra estas
Gigls, ->	Nul
Convertion of the second product Convertion of the second product of the second produ	
Einer beiter over zurdende. Filmen son filmen einer Filmen son filmen einer son	
Reutleally Driven "Revelling Contornations Space Institute a first the Contornations Space Institute a first the Contornation	
-, and Near A., And Andre. Low Folgithan 2005	

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...

1. Tokenization

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

Language Processing Problems

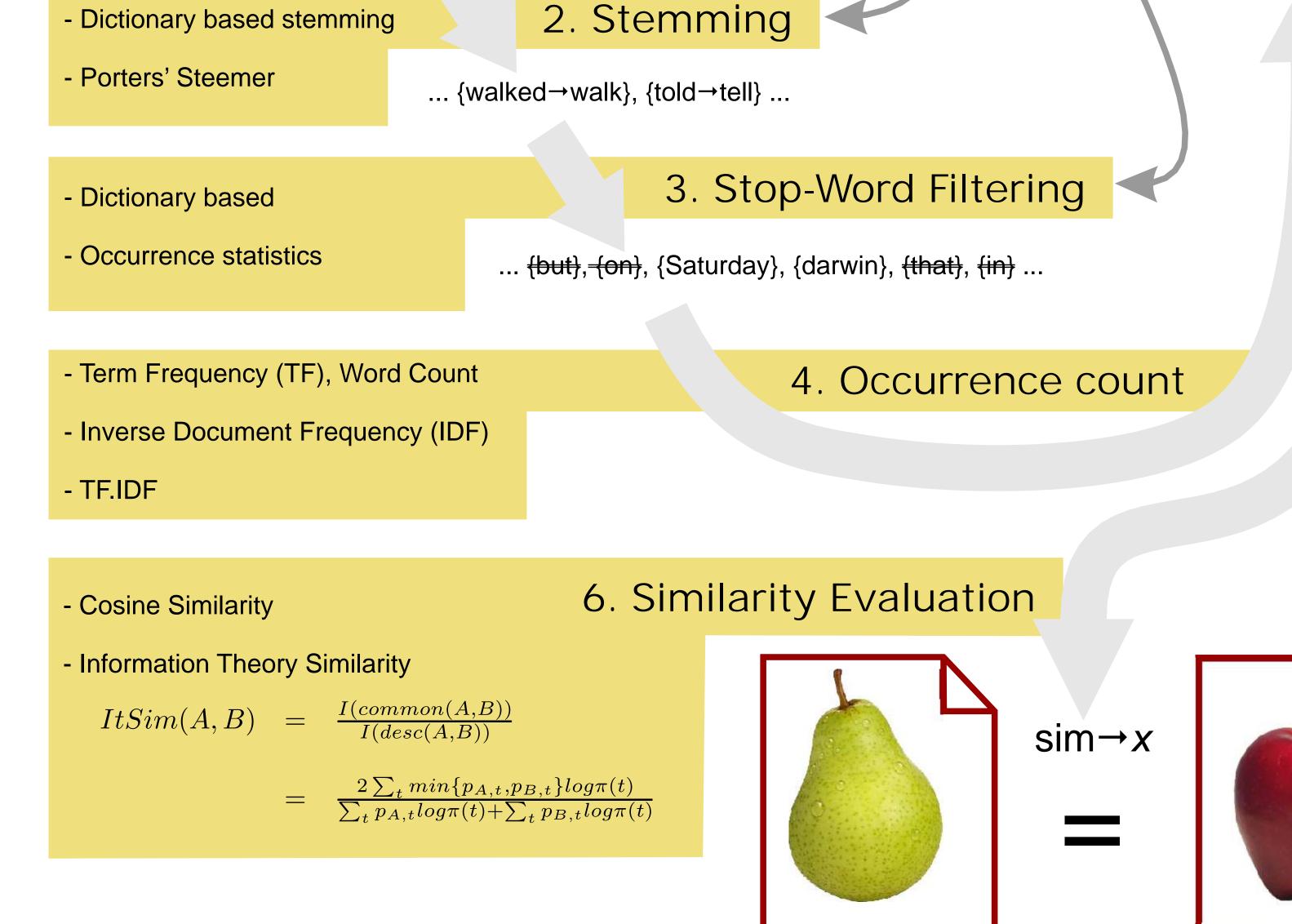
Polysemy - ambiguity of an individual word or phrase that can be used (in different contexts) to express two or more different meanings

- **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 . . . 1, 2, d = (

5. Dimensionality Reduction - Latent Semantic Analysis (uses SVD) $X = U \Sigma V^{\mathsf{T}} \rightarrow X_{\mathsf{k}} = U_{\mathsf{k}} \Sigma_{\mathsf{k}} V_{\mathsf{k}}^{\mathsf{T}}$ term-document matrix k-rank term-document matrix $d^{T} = \Sigma_{k}^{-1} U_{k} d^{T} = (0.1, 0.3, ...)$ 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.





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