Linking Historical Dictionaries

Karin Heß

Max-Planck-Institut für Informatik - Prof. Weikum

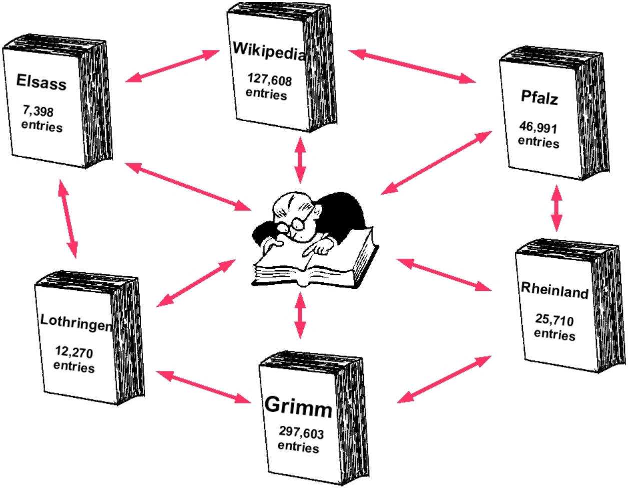
WS 06/07

Motivation

- Grimmsches Wörterbuch and Dialect Dictionaries of the University of Trier
- Dictionaries partially online accessible but difficult to use

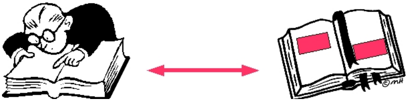
⇒ Goal: Make the Dictionaries more easily accesible by linking

Goal

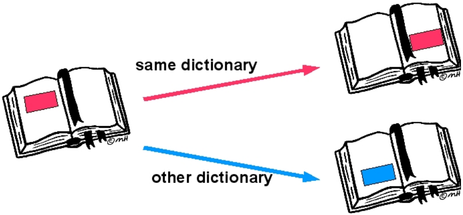

Elsass
7,398 entries
Wikipedia
127,608 entries
Pfalz
46,991 entries
Lothringen
12,270 entries
Rheinland
25,710 entries
Grimm
297,603 entries

Goal

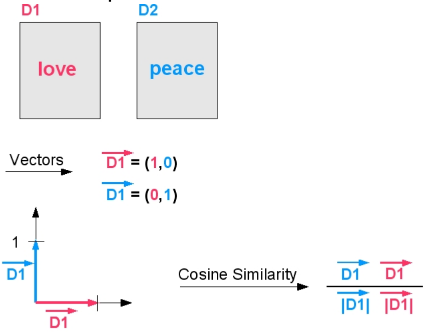
Keyword search:



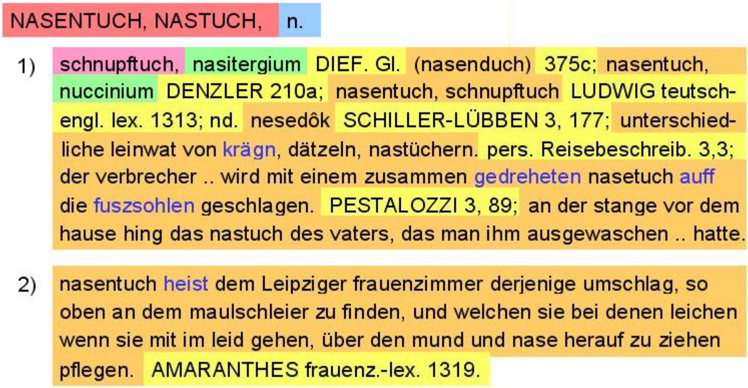
Document search:



1. Preprocessing of the data:
   - What kind of data do we have?
   - How can we preprocess the data to allow an efficient search?
2. Vector-Space Model



Example entry (Grimm)



NASENTUCH, NASTUCH, n.

1) schnupftuch, nasitergium DIEF. Gl. (nasenduch) 375c; nasentuch, nuccinium DENZLER 210a; nasentuch, schnupftuch LUDWIG teutsch-engl. lex. 1313; nd. nesedôk SCHILLER-LÜBBEN 3, 177; unterschiedliche leinwat von krägn, dätzeln, nastüchern. pers. Reisebeschreib. 3,3; der verbrecher .. wird mit einem zusammen gedreheten nasetuch auff die fuszsohlen geschlagen. PESTALOZZI 3, 89; an der stange vor dem hause hing das nastuch des vaters, das man ihm ausgewaschen .. hatte.

2) nasentuch heist dem Leipziger frauenzimmer derjenige umschlag, so oben an dem maulschleier zu finden, und welchen sie bei denen leichen wenn sie mit im leid gehen, über den mund und nase herauf zu ziehen pflegen. AMARANTHES frauenz.-lex. 1319.

Problems

- Latin terms raise problems with stemming *capere* → *cap* ← *cape canaveral*
- Composed words are difficult to link *igelbalg* → *igel*, *balg* → *blasebalg*
- Dialect expressions can cause mistakes *kalt (schweiz. gehalt*) → *stiftung schweiz*
- General problems: spelling variants and unique terms

⇒ We need several forms of preprocessing

Preprocessing

Initial example query: $q = \{der, capere, igels, vortheekocher\}$

Stopword elimination

$$stopwordElimination(q) = \{capere, igels, vortheekocher\}$$

Latin term elimination

$$latinElimination(q) = \{igels, vortheekocher\}$$

Preprocessing

Rule application - Software Andrea Ernst-Gerlach (Duisburg-Essen)

$RuleSet = \{(a, a')|a, a' \in W\}$
if any $(a, a') \in RuleSet$, if $match(t, a)$, a' is applied to a, thus yielding the word t
$ruleApplication(q) = \{igels, vortheekocher, vorteekocher\}$

Elimination of the prefixes

$prefixElimination(q) =$
$\{igels, vortheekocher, theekocher, vorteekocher, teekocher\}$

Preprocessing

Stemming - Treetagger

$stemming(q) =$
$\{igel, vortheekocher, theekocher, vorteekocher, teekocher\}$

Decomposition - Connexor

$decomposition(q) = \{igel, tee, kocher\}$

Vector-Space-Model

TfIdf

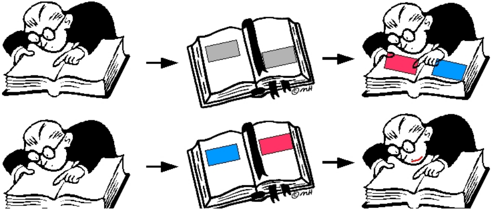
$$d_{ij} = \frac{(\log tf_{ij} + 1.0) \cdot idf_j}{\sum_{j=1}^{t}((\log tf_{ij} + 1.0) \cdot idf_j)^2}$$

with document i, term j

Cosine similarity

$$sim(d, q) = \frac{\sum_{j=1}^{t} w_{qj} d ij}{\sqrt{\sum_{j=1}^{t}(d_{ij})^2 \sum_{j=1}^{t}(w_{qj}^2}}$$

Relevance feedback



Roccio - Relevance feedback

New query vector:

$$q_i' = aq_i + b\frac{1}{|r|}\sum_{d_{ij}\in r} d_{ij} - c\frac{1}{|n|}\sum_{d_{ij}\in n} d_{ij}$$

parameters a, b, c, relevant documents r, non relevant documents n

BM25

Weighting function

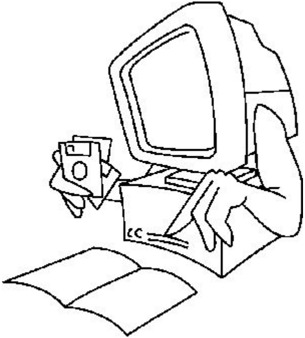
$$\frac{(k_3 + 1.0)qf_i}{k_3 + qf_i} \times \frac{(k_1 + 1.0)tf_i}{k_1 \cdot avdl \cdot tf_i}$$

term i
parameters k_1, k_3
query term frequency qf
term frequency tf

Outlook

- Efficiently and Scalability of the application (precomputations, indexing)
- Search refinement using external resources

Web Application for Dictionary search



Thank you for your attention!