Introduction to Information Retrieval and Semedico Evaluation

Erik Faessler

Lehrstuhl für Computerlinguistik
Friedrich-Schiller-Universität Jena

Oberseminar
15.12.2017

Jena University Language & Information Engineering Lab

Einführung

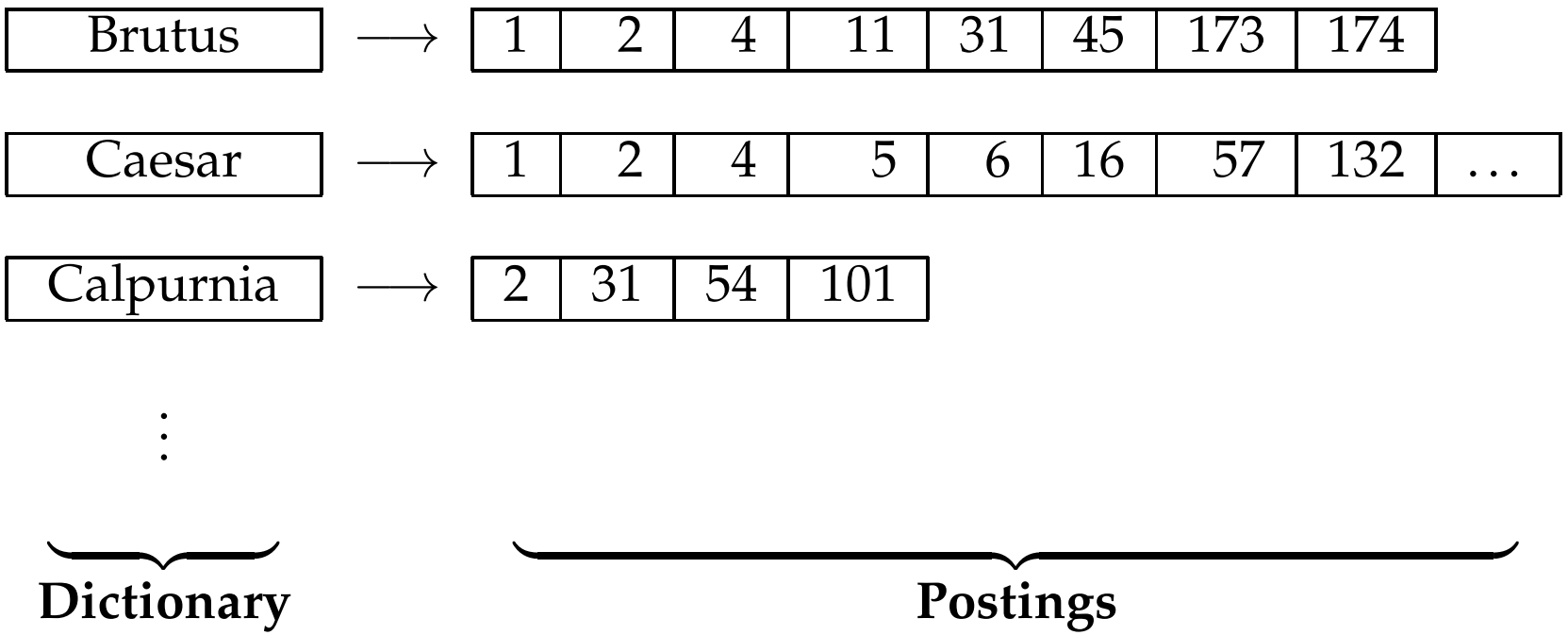
- Ausgangspunkt:
  - Informationbedürfnis (*information need*)
  - (große) Textkollektion
- Ziel:
  - Finde *Suchterme* im Corpus
  - Erstelle Liste von Trefferdokumenten

Beispiel / Intuition

- Corpus:
  - Reuters RCV1 Auszug
  - ~50k Dokumente von 810k
  - Textinhalte extrahiert
- Demo
  - Beispieldaten
  - Naive Suche
  - Einfach(st)er Index

Der invertierte Index

- **Die** *zentrale* Datenstruktur im IR
- *Un*invertiert: Normaler Zustand, Dokumente 'haben' Wörter
- Invertiert: Wörter 'haben' Dokumente
- Einfachstes Beispiel: Buchindex

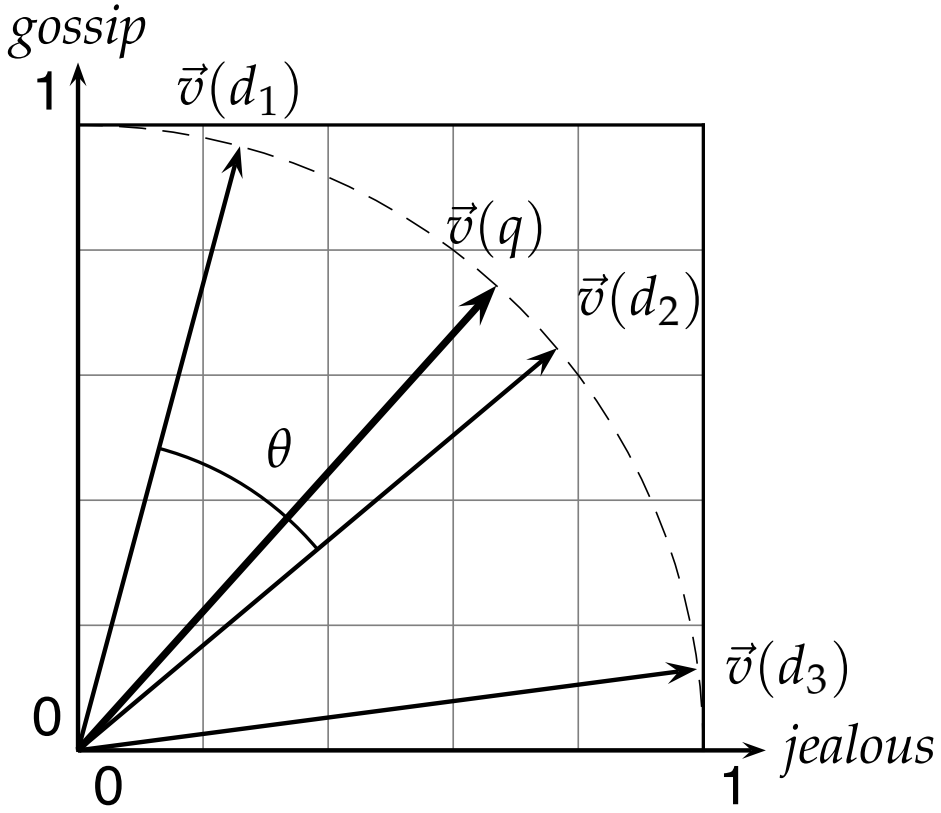


Erweiterungen des invertierten Index

- Termfrequenz pro Dokument
- Termposition im Dokument
- Textlängennorm
- Speichern von Dokumententextauszügen
- Massen von Optimierungen
- Mehrere Felder pro Dokument
- Lucenevortrag?

Ranking nach *Relevanz*

- Bislang boolesche Suche: Dokument getroffen / nicht getroffen
- Vektorraummodell (TF/IDF)



- Probabilistische Modelle (Okapi BM25)

Evaluation

- Evaluationscorpora
  - Dokumente
  - Anfragen (*queries*)
  - Relevanzbeurteilung von Dokumenten *pro Anfrage*
- **Die** Information Retrieval Challenge Serie: TREC
- http://trec.nist.gov/
- Evaluationmaße:
  - Precision, Recall, F-Score
  - *mean average precision*

Mean Average Precision

- Q ~ Menge von Queries
- $\{d_1, \ldots, d_{mj}\}$ ~ relevante Dokumente für Anfrage $q_j \in Q$
- R_{jk} ~ Menge der gerankten retrieval Ergebnisse vom ersten Ergebnis bis zum Dokument d_k
- Precision ~ $\frac{\#(\textit{relevante gefunden})}{\#(\textit{gefunden})}$

$$\mathrm{MAP}(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} \mathrm{Precision}(R_{jk})$$

Semedico

- Biomedizinische (noch) Suchmaschine am JULIE Lab
- http://semedico.org/
- Evaluation: TREC genomics 2005

Introduction to Information Retrieval and Semedico Evaluation

Erik Faessler

Lehrstuhl für Computerlinguistik
Friedrich-Schiller-Universität Jena

Oberseminar
15.12.2017