Einführung

Statistische Verfahren der automatischen Indexierung

Leistung linguistischer und wörterbuchbasierter Verfahren

- Können Indexeinträge zusammenführen
- verkleinern Index und erhöhen Recall
  - Stoppwortliste
  - Grundform/Stammform
  - verbessern Volltextretrieval, indem sie dabei helfen, nur sinntragende Worte zu indexieren (Thesaurusabgleich)
- können Deskriptoren vergeben, die nicht im Text vorkommen, indem sie Worte nach bestimmten Vorschriften auf einen Thesaurus/Wörterliste abbilden: nach dem Modell: "wenn maschin* neben indexi* oder Verschlagwortung steht, dann indexiere 'automatische indexierung'

Grenzen linguistischer Verfahren

- liefern keinen Anhaltspunkt, wie Deskriptoren bestimmt werden, die besonders **repräsentativ** für einen Text sind
- behandeln die Bedeutung aller Wörter gleich, haben keine Regel / Modell, um Wertigkeit von Dokumenten für das Retrieval zu gewichten (**Ranking**)
- jedes Dokument wird für sich allein betrachtet und **keine Beziehung zur Gesamtheit** aller Dokumente im Speicher hergestellt

Lösung: Statistik

- Statistische Eigenschaften von Texten:
  - Wenige Worte treten sehr häufig auf
    - die 2 häufigsten Wörter können 10% eines Textes ausmachen,
    - die 6 häufigsten Worte machen 20% des Textes aus,
    - die häufigsten 50 Worte machen 50% aus
    - Beispiel: in Textsammlung von über 300.000 Dokumenten macht "the" = 5.9% des Textes aus und "of" =3.1%
  - Viele Worte sind selten
  - Texte lassen sich nach der Verteilung der Worthäufigkeit unterscheiden

Zipfs Gesetz

- Es besteht eine konstante Beziehung (C) zwischen dem Rang (r) eines Wortes in einer Häufigkeitsliste und der absoluten Häufigkeit (f) des Wortes in einem Text
  - **r x f = C**

Wort	Häufigkeit	Rang	C-Wert
Wort 1	2.653	10	26.530
Wort 2	1.311	20	26.220
Wort 3	26	1000	26.000

26.000 ist der Einheitswert für James Joyce Roman „Ulysses“

Grundannahmen der statistischen Indexierung

- Nicht alle Worte eines Dokuments sind als Indexterme geeignet
  - Es muss eine **Auswahl** getroffen werden
- Nicht alle ausgewählten Indexterme sind gleich relevant
  - Es muss eine **Gewichtung** der Indexterme vorgenommen werden

Kann die Statistik helfen?

"It is here proposed, that the the frequency of word occurrence in an article furnishes a useful measurement of word significance." (H. P. Luhn)

Fragen:

- Was macht man mit ganz häufigen Worten?
- Was macht man mit Worten, die nur 1x vorkommen?
- In einer 10 zeiligen dpa Meldung kommt das Wort Doping 5x vor. In einem 3 seitigen Zeitungsartikel ebenfalls. Wie beurteilen Sie die Eignung von Doping als Deskriptor für die beiden Dokumente?

Berechnung der Termfrequenz

TF(td)= Häufigkeit eines Wortes im Dokument
Anzahl aller Wörter des Dokuments

Beispiel:

•in einem Text (a) aus 200 Wörtern kommt Gesundheit 5 x vor.

TF = 5/200 (0,025)

•In einem Text (b) aus 2000 Wörtern kommt Gesundheit 6 x vor.

TF = 6/2000 (0,003)

Problem?

Probleme

Beispiel:
in einer Datensammlung gibt es 1000 Dokumente. Das Wort Segelboot kommt in 5 Dokumenten vor.
In der Datensammlung gibt es 500 Dokumente, in denen Sport vorkommt.

Problem für die automatische Indexierung?

Berechnung der Inversen Dokumenthäufigkeit

IDF= $\frac{\text{Anzahl aller Dokumente in Datensammlung}}{\text{Anzahl Dokumente, in denen Suchbegriff vorkommt}}$

Beispiel:

•in einer Datensammlung gibt es 1000 Dokumente. Das Wort Segelboot kommt in 5 Dokumenten vor.

IDF= 1000/5 = 200

•In der Datensammlung gibt es 500 Dokumente, in denen Sport vorkommt

IDF = 1000/500 = 2