Information Retrieval

Sprachanalyse

Stefan Birkner

31. August 2010

Tokenisierung

Information Retrieval ist ein Fachgebiet, das sich mit computergestütztem Suchen nach komplexen Inhalten beschäftigt.

Information | Retrieval | ist | ein | Fachgebiet | das | sich | mit | computergestütztem | Suchen | nach | komplexen | Inhalten | beschäftigt

Stoppwörter

Entfernung von Termen, die nicht zur Semantik der Dokumente beitragen.

Stoppwörter

- Wirtschaft und Gesellschaft befinden sich derzeit im größten Umbruch seit der Industrialisierung. Die Ursache hierfür liegt in der globalen Verfügbarkeit leistungsfähiger und zugleich kostengünstiger Informations- und Kommunikationstechnologien. Das Informationszeitalter wird Realität.
- befinden, Das, der, derzeit, Die, Gesellschaft, globalen, größten, hierfür, im, in, Industrialisierung, Informations, Informationszeitalter, Kommunikationstechnologien, kostengünstiger, leistungsfähiger, liegt, Realität, seit, sich, Umbruch, und, Ursache, Verfügbarkeit, wird, Wirtschaft, zugleich

Stoppwörter

befinden, ~~Das~~, ~~der~~, derzeit, ~~Die~~, Gesellschaft, globalen, größten, hierfür, ~~im~~, ~~in~~, Industrialisierung, Informations, Informationszeitalter, Kommunikationstechnologien, kostengünstiger, leistungsfähiger, liegt, Realität, ~~seit~~, ~~sich~~, Umbruch, ~~und~~, Ursache, Verfügbarkeit, ~~wird~~, Wirtschaft, ~~zugleich~~

Stoppwörter

Warum werden Stoppwörter entfernt?

- reduzierter Speicherplatzbedarf
- höhere Performance der Matchingverfahren

Stoppwörter

Wie findet man Stoppwörter?

- Stoppwortlisten
- häufige und sehr seltene Wörter

Stemming

Problemstellung

- laufen, lief, gelaufen
- das Haus, des Hauses, die Häuser
- proben, Erprobung, Probe

Stemming

- Grundformreduktion
- Stammformreduktion

Stemming

Regelbasierte Verfahren

- Lovins-Algorithmus
- Porter-Stemmer-Algorithmu

Stemming

- Regelbasierte Verfahren für schwach flektierte Sprachen (z. B. Englisch)
- Wörterbuch-Verfahren für stark flektierte Sprachen (z. B. Deutsch)

Mehrwortgruppenidentifikation

- Bundeskanzlerwahl
- die Wahl des Bundeskanzlers
- die Wahl des Bundeskanzlers fiel auf Saumagen mit Sauerkraut

Mehrwortgruppenidentifikation

- keine Mehrwortgruppenidentifikation
- Zerlegung der Worte in ihre Bestandteile

N-Gramme

Eisenbahn

- 3-Gramme: eis, ise, sen, enb, nba, bah, ahn
- 2-Gramme: ei, is, se, en, nb, ba, ah, hn
- vor allem für längere Suchangfragen vorteilhaft

Metadaten

- Text wird mit zusätzlichen Attributen angereichert
- Beispiel: Autor, Sprache, Herausgeber, Datum, Schlagworte, Titel

Literatur

- Andreas Henrich: Information Retrieval 1
- `http://www.uni-bamberg.de/minf/ir1_buch/`
- Kapitel 3