Technologien für das Internet I

Übung vom 21.10.2011
Bearbeiten bis: 27.10.2011
Bei Fragen: demidova@l3s.de

1. Suchanfrage-Optimierung

1. Eine Dokumentsammlung mit insgesamt 100 000 Dokumenten beinhaltet Wetterbreichte. Gegeben sei die Anfrage:

Sommer AND (Sonne OR Gewitter) AND NOT (Wind)

Geben Sie die effizienteste Abfrage-Reihenfolge an, die sich aus folgender Tabelle bestimmen lässt:

Term	**DF (document frequency)**
Sommer	20 000
Sonne	5 000
Gewitter	1 500
Wind	25 000

Ist die gefundene Reihenfolge garantiert optimal?

2. Invertierte Dateien (Inverted Index)

Gegeben ist die folgende Dokumentsammlung:

Dokument 1:
In Fluch der Karibik 4 begegnet Johnny Depp einer Frau aus seiner Vergangenheit wieder.
Dokument 2:
Johnny Depp wird für seinen Auftritt in Fluch der Karibik 4 mit einer Rekordgage von über 50 Millionen Dollar entlohnt.

Erzeugen Sie aus den Dokumenten eine invertierte Datei (inverted index).
Tokenisierung-Regeln: Wort-weise (Satzzeichen ignorieren), alles klein geschrieben.
Stopwörter*: in, der, einer, aus, seiner, für, seinen, mit, von, über.*
Geben Sie dabei jeweils ***an passender Stelle*** für jeden Term die entsprechenden Werte für ***TF*** (Termhäufigkeit, term frequency) und ***DF*** (Dokumenthäufigkeit, document frequency) an.

Welche Ergebnisse liefern die folgenden Suchanfragen für diese Dokumentsammlung:

Q_1= *Johnny UND Depp UND NOT Frau*
Q_2= *Meerjungfrau*