Stefan Bunk 756006
Franz Liedke 757081
Cornelius Bock 754432

Week 2: Indexing

Assignment 1: Text Statistics

- Kepler’s third law
- Lotka’s law

Assignment 2: Text Processing

Raw text documents are usually pre-processed before an index is build.
a) The German language has many compound words. Give two examples of compound words that should be split into tokens to find all relevant documents. Give two counter examples where splitting the compound word would lead to retrieving irrelevant documents.
Split:

- Pflaumenkuchen
- Autohändler

Not to split:

- Stellplatz
- Stammspieler

b) What are the pros and cons of using a stopword filter?
Pros:

- reduce index space
- improve response time
- improve effectiveness

Cons:

- can’t find combinations (“if and only if”, “to be or not to be”)

c) What are the pros and cons of stemming and lemmatization?
Pros:

- words with similar meaning are found even if their different
- improves search effectiveness

Cons:

- takes time at indexing
- false positives worsen search performance

d) When does stemming take place, at indexing or query time? Explain your answer.
Both. Words from documents are stemmed during index time, the query keywords at query time.

e) Can stemming lower precision or recall in a simple keyword retrieval system? Explain your answer.

Recall will increase, because there are more documents found. Because of false positives, precision could be lowered.

Assignment 3: Index Compression

Encode the following posting lists. A posting list comprises pairs of (document ID, term frequency)

a) (2,1), (4,2), (5,2) → delta encoding

2 1 2 2 1 2

b) (2,1), (4,2), (5,2) → unary encoding

11010111101101111110110

c) (6,3), (7,1), (10,4) → Elias gamma encoding

11010 101 11011 111010 11000

Assignment 4: (Programming) Indexing

Print the list of Wikipedia titles in your development set whose articles match the queries:

a) "Uranisotope"

[There are no documents found, because "Uranisotope" was stemmed to "Uranisotop", but "des Uranisotops" from the Text stays as it is after stemming.]

b) "Artikel"

```
Alan Smithee
Ang Lee
```