Stefan Bunk 756006
Franz Liedke 757081
Cornelius Bock 754432

Week 4: Retrieval Models II

Assignment 1: Probabilistic Retrieval

a) What is "binary" in the binary independence model (BIM)?
The occurence of the terms within a document is assumed independent. This assumption is obviously wrong: When the word "president" occurs in a document, it is more likely that the word "lincoln" occurs in the text, than if the word "president" does not occur, i.e. P("lincoln" occurs|"president" occurs) is assumed to be the same as P("lincoln" occurs).
However, in practice, this assumption has shown to give good results.

b) What is "independent" in the binary independence model (BIM) and is this a reasonable assumption? Explain.
The occurence of the terms within a document is assumed independent. This assumption is obviously wrong: When the word "president" occurs in a document, it is more likely that the word "lincoln" occurs in the text, than if the word "president" does not occur, i.e. P("lincoln" occurs|"president" occurs) is assumed to be the same as P("lincoln" occurs).
However, in practice, this assumption has shown to give good results.

c) What are the differences between standard vector space tf-idf weighting and the BIM probabilistic retrieval model (in the case where no document relevance information is available)?
Both VSM and BIM contain an idf-component, which is almost identical: In VSM it is log (N/n_i), wheres in BIM it is log((N - n_i)/n_i). However, BIM does not contain an tf-term, i.e. a term that increases with the number of occurences of a term i in a document. This leads to 50 % less relevant documents in top 10 rankings.

d) What is parameter b good for in the BM25 model?
The parameter b regulates the normalization of the length, i.e. how much of the document length in relation to the average document length is taken into account. If b = 1, there is full-normalization and b = 0 means no normalization at all.

Assignment 2: Language Models

Build a query likelihood model using maximum likelihood estimates. Use Jelinek-Mercer smoothing with λ = 0.2. Compute the ranking of the four documents for the queries
a) "click"
2, 1, 4, 3

b) "test"
1, 4, 2, 3

Stefan Bunk 756006
Franz Liedke 757081
Cornelius Bock 754432

Assignment 3: (Programming) Ranking

Print the ranking of Wikipedia titles in your development set for the queries:
a) "artikel regisseur"

```
Alan Smithee
Ang Lee
```

b) "regisseur"

```
Alan Smithee
Ang Lee
```

c) "deutsch"

```
Ang Lee
Actinium
Alan Smithee
Anschluss (Soziologie)
```

d) "anschluss"

```
Ang Lee
Anschluss (Soziologie)
Anschlussfaehigkeit
```