Information Retrieval – Project 2

**Index**

We built up the index by passing once over the data. The index maps from words to tuples containing the doc id and the local term frequency. We built additional indices that mapped from doc ids to the length of a document and to the title of the document.

**Preprocessing**

We modified tokenization to additionally split words where small letters change to capital ones.

We preprocessed the documents by removing stop words, removing all words shorter than 4 letters, removing all words only appearing once inside a document, and lowercasing all words.

**Term-based-model**

We used standard tf-idf and added some reward for documents containing query words in their title.

**Language Model**

We used the model in the slides with Jelinek-Mercer smoothing. As hyperparameter we chose lambda=0.15. We did some optimizations to not have to scan the full set of documents again, when computing the scores.

An overview of different evaluation measures for these two models:

|  |  |  |
| --- | --- | --- |
| **Evaluation** | **Term model** | **Language model** |
| Mean precision | 0.21 | 0.291 |
| Mean Recall | 0.078 | 0.106 |
| Mean F1 | 0.099 | 0.136 |
| MAP | 0.285 | 0.37 |

**Performance**

It took approximately 4 minutes and 1GB of memory to build up the index for the full set of documents.