MUNI
FI

Cranfield collection IR system

Václav Sobotka
454828@mail.muni.cz

May 7, 2021

Document & query preprocessing techniques

- Tokenization – nltk word tokenizer
- Trimming of hyphens, apostrophes, slashes and similar characters
- Splitting tokens by hyphens and slashes into more tokens
- Filtering of ignored tokens – slightly expanded list of stopwords from nltk
- Stemming using SnowballStemmer from nltk
- Bigrams are derived and used as separate terms
  - Bigrams are created as a concatenation of two subsequent terms after both are completely preprocessed including the stemming

Inverted index structure & vector model

- Custom implementation
- Items in posting lists keep occurrence counts for title, authors, bibliography and body separately instead of total count for term in document
  - Further used for different importance of document sections in the TF-IDF model
- TF-IDF with (weighted) cosine similarity is used as the vector model

Pseudo-relevance feedback

- Proved to have a significant impact on the results
- After obtaining the initial result, all documents with scores of at least 90 % of the best document are taken as relevant
- Top 100 important terms (based on TF-IDF) per document are kept
- The second exapanded query is enriched by the most important terms from all the documents that were assumed to be relevant
- Only one additional query is constructed and executed

Topic modeling

- In addition to the main index used for TF-IDF searching, 3 more indices with topic modeling were used
- The topic modeling code was taken from the example notebook for LSI
- The three indices use 300, 150 and 50 topics respectively
- The final ordering of answers is based on a weighted scheme
  - The main TF-IDF index is the most important
  - The importance of the topic indices decreases with the decreasing number of topics
  - The final score is just a weighted sum of scores achieved across all of the four indices

MUNI
FACULTY
OF INFORMATICS