Indexer

The indexer first reads in the entire document collection and processes each document sequentially. For each document, its contents are split into tokens and then processed according to the given user parameters.

If activated, stemming reduces every token to its root form. Common words are discarded from indexing if the option to remove stopwords was used when starting the indexer. Each resulting term is then stored in an internal dictionary data structure along with the number of occurrences encountered so far in the collection. After calculating the frequency of each term across the collection, optional minimum and maximum thresholds are applied to the dictionary, removing terms whose number of occurrences do not fall within the specified range.

Once the finalized pairs of terms and their frequency in the document collection has been created, the term frequency-inverse document frequency for every term is calculated for every document. These values are converted into Weka Instances objects, with each Instance representing a single document with attributes for the TF-IDF weights of every term in the dictionary. The resulting index is then written incrementally to disk for later user by the search.

Search

For the similarity method uses a Vector Space Model in combination with TF-IDF weighting and cosine distance function (formula as in lecture slides).
TF-IDF according to following formula:
w = 1+ log(tf) x log (N / df)
N… document list size
df… document frequency
tf… term frequency

Before the search starts the index is read from an ARFF file into memory.
Afterwards the search engine starts with parsing the lines of the input topics file and calculating the distance (via cosine distance function) to all the other topic vectors in the vector space whose TF-IDF are greater than 0 for each input topic vector.
Then the result list is sorted by their cosine score and the output is written into the folder “/output/”

Usage

- Indexer: -i [-stem] [-stop] [-min <value>] [-max <value>]
    - -i indicates the program should run the indexer
    - -stem activates stemming during index creation (optional)
    - -stop enables the removal of stopwords during the creation of the index (optional)

    - -min <value> terms that do not occur at least the given number of times will not be indexed (optional)
    - -max <value> terms that occur more than the given number of times will not be indexed (optional)
- Search: -s [-t <value>] [-n <value>]
    - -s indicates the program should run the search
    - -t defines the path of the input topics text file
    - -n is the number of results per search (default 10)