Information Retrieval Systems

Practice 1.2: Normalization (*stopper)*

Description

In this practice, we will continue with the preprocessing and normalization of the documents that we started in the previous practice. In this session we are going to reduce the number of terms in the documents by removing common words which are of little value to select documents.

Therefore, in this practice, it is requested to make a module that allows removing **stop words**. Given as input a text and a list of stop words, this module must return the input text, but removing the stop words present in it.

You can download the list of stop words at the following address: http://members.unine.ch/jacques.savoy/clef/. In the column *Stopword list* you can find the list of stop words for each language.

Once this module is done, it will be included in the main program of the previous practice that allowed us to read all the files of the collection. You will have to process the documents with this module and store them in a new folder called stopper with the same file names.

Existing libraries can be used to help with processing.

Documentation

In order to include additional information in the final report of practice 1, you must obtain some statistics on the number of words before and after performing the filtering of stop words (stopper). The information that should be included in the report is as follows:

- Total number of words in the collection (before and after).
- Average, maximum and minimum number of words in the documents of the collection (before and after).
- The 5 most frequent words (before and after).