Information Retrieval Systems

Practice 1.3: Normalization (*stemmer)*

Description

In this practice, we will continue with the preprocessing and normalization of the documents. In this session we are going to reduce the number of terms in the documents by reducing various forms of a term to a common root.

Therefore, in this practice, it is requested to make a module that allows the **extraction of roots**. Given as input a text, this module must return the input text, but with the words lemmatized with the corresponding stemmer. There are several algorithms to perform *stemming*, but the most popular is Porter Stemmer (for English). You can use external libraries to develop this module, such as Snowball, which you can find at the following address: http://snowball.tartarus.org/download.html. This library has algorithms for several languages, including Spanish.

Once this module is done, modify the main program of the previous practice so that it processes the files of the *stopper* folder with this module and stores them in a new folder called *stemmer* with the same 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 extraction of roots (stemming). 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 and their frequency (before and after).

You can compare this information with the one generated in the previous practice.

In addition, it must be included in the final memory which version of stemmer has been used (Porter, Snowball, ...), what problems have appeared when using it and how they have been solved.