Information Retrieval Systems

Practice 1.4: Creation of word-frequency pairs

Description

Let's start with the development of the index. Until this session, the system has several modules already completed and tested:

- Information extraction from the web pages of the collection
- Processing of each word of each extracted document:
    - Normalization
    - Stopper
    - Stemmer

Next, we're going to create a dictionary as a storage structure for each different term. This dictionary will assign a unique numeric value to each different term. Likewise, each file in the collection will also have a unique identifier value and will be stored in another structure so that at the end we can obtain the URL or file name, for example, from the ID of each relevant document.

It’s time to build the data structure that you will use for your index, which will contain the number of times each term appears in each file. Once this data structure is fixed, implement the methods to save the word-frequency pairs going through all the documents in the collection. Think, before implementing anything, about the different solutions and the performance differences of each one.

Documentation

To understand the volume of information handled, the program must generate a log file with the following information:

- Time in seconds to calculate and generate the selected structure
- Size in bytes of the structure stored in memory
- Characteristics of the machine where the calculation is made.
- OPTIONAL: If you make the calculation in several machines (home and classroom) you can put the information of both to compare the differences.