Information Retrieval Systems

Practice 1.1: Process HTML collection

Description

One of the necessary things for an RI system is the collection of documents. Each document in the collection must be processed prior to the indexing process. It will consist of cleaning the text and creating the documents in plain text.

In this first practice, the aim is to carry out this process of cleaning up the texts of a collection. To do this, you must download the collection provided through Docencia Virtual platform:

> colecciónESuja2019.zip: 838 HTML files in Spanish with news from the Diario Digital of the University of Jaén.

In this practice, two modules will be carried out:

1. **Filtered.** Given the name of a text file, the module will return the text without HTML tags (content only). This text must contain at least the title of the article and the body of the article.
   OPTIONAL: Other sections of the document can be included such as the date, categories, tags...
2. **Normalization and tokenization.** This method will take as input the filtered text and return the same text but removing all the "rare" characters,We will change the text to lowercase and replace the accented letters with their corresponding ones without accentuation, finally, we will put one word per line (tokenization) in the file. In our case we will remove all the characters that are not included in the following range:

   [‘a’...’z’, ‘A’...’Z’, ‘ ‘, ‘_’, ‘-‘, ‘\n’,‘0’..’9’]

Once these two modules have been completed, a main program must be created that allows us to read all the files in the collection, we must process each file with the two previous modules and store them in a new folder with the same file name. This main program should reflect on screen the time it takes to perform the process.

Existing libraries can be used for HTML processing and tokenization.

Documentation

You should save the following information and include it in your final practice documentation: number of files processed, time to complete the process, number of total collection tokens and average number of collection tokens.