Information Retrieval Systems

Practice 1.5: Term frequency, inverse document frequency, weights and normalized weights

Description

We continue with the processing of the words from the collection and, before going on to calculate the weights without normalizing, we will make a simple improvement of the system, the normalization of frequencies.

It consists of taking the highest frequency of each document in the collection, and dividing all the frequencies of the words in that document by this maximum value. In this way, this maximum value, when divided by itself, will remain with value 1 and the rest of frequency values between 0 and 1. With this we manage to normalize the frequency values between documents and it will no longer influence the results if we have in the collection very extensive documents and other very brief.

$$tf_{ij} = \frac{f_{ij}}{max\{f_{xj} \forall x \in j\}}$$

Where, f_{ij} is the frequency of word *i* in document *j*

Next, we apply the steps seen in theory about the vector space model, in order to calculate IDFs, unnormalized weights, and normalized weights:

$$idf_i = \log(\frac{N}{df_i})$$
$$w_{ij} = tf_{ij} \times idf_i$$
$$wn_{ij} = \frac{w_{ij}}{\sqrt{\sum_{\forall x \in j} w_{xj}^2}}$$

Where:

- *N* is the total number of documents in the collection
- df_i is the documentary frequency of the word *i*
- w_{ij} is the weight of the Word *i* in the document *j*

It will be necessary **to save this last table of normalized weights**, to work later with the queries, **as well as the IDFs**. All this will be part of the index of the collection, so it should be saved in a separate folder.

Documentation

In order to include additional information in the final report, you must obtain some statistics:

- Time in seconds to calculate the normalized weights.
- 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.