Information Retrieval

Practical session n°3: Weighting

Create a directory named *practice3*. In this directory, create a new file named *practice3_report.txt*.
During the practical session, for each exercise, copy-paste some outputs of your program into this file to demonstrate that you have completed the exercise and it works correctly. Add some explanations.
At the end of the practical session, copy-paste the source code of your program(s) in the directory *practice3*.
Compress the directory in a file named *practice3_YourTeamName.zip* (e.g.: *practice3_ VictorAlbertJules.zip*).
Upload this compressed file (one file / team) on the website of the course (**deadline: November 2nd** (first try: exo 1-2-3); **November 9th** (final version)).

Exercise 1: Increasing (again) the size of the collection

Download the collection of documents *Practice_03_data.zip* on the website of the course. It contains 9,804 documents stored in a single file (76.4MB).

Index this collection using your indexing program (tokenization, dictionary, *df*, *tf*):

- Simple tokenization: terms without digits or special characters.
- No stop-words list is used.
- No stemming is applied.

Compute the following collection statistics:

- Total indexing time (#sec).
- Total number of tokens occurrences in the entire collection (#tokens), before any further processing (normalization, case folding, stop-words, stemming, etc.).
- Total number of distinct tokens in the entire collection (#distinct tokens), before any further processing.
- Average length of distinct tokens (#characters).
- Total number of terms occurrences in the entire collection (#terms).
- Total number of distinct terms in the entire collection, i.e. vocabulary size (#distinct terms).
- Average document length (#terms).
- Average length of vocabulary terms (#characters).

Exercise 2: Collection Statistics using stop-words and stemmer

Refresh the index by removing stop-words (stop-words-english4.txt) and applying Porter's stemmer.
Recompute the statistics of the exercise n°1.

Exercise 3: Ranked Retrieval (SMART *ltn* weighting)

Compute a weighted index using the SMART *ltn* weighting function.
Calculate the score for each document in response to the query « web ranking scoring algorithm », using the index based on the SMART *ltn* weighting function.

Provide the following statistics:

- Total weighting time (#sec).
- Weight of the term "ranking" in document #23724.
- RSV (Retrieval Status Value) of document #23724.
- Top-10: a list of the ten most relevant documents along with their RSV.

Exercise 4: Ranked Retrieval (SMART *ltc* weighting)

Answer the same questions as in Exercise 3, but using the SMART *ltc* weighting function.

Exercise 5: *BM25* weighting

Answer the same questions as in Exercise 3, but using the *BM25* weighting function with these usual values: k_1 = 1.2 and b = 0.75.