Information Retrieval

Practical session n°2: Pre-Processing and Dictionary

Create a directory named *practice2*. In this directory, create a new file named *practice2_report.txt*.
During the practical session, for each exercise, copy-paste in this file some outputs of your program, showing that you complete 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 *practice2*.
Compress the directory in a file named *practice2_YourTeamName.zip* (or *.tar*, *.gz*, *.rar*, etc.) (e.g.: *practice2_ VictorAlbertJulesIsaac.zip*).
Upload this compressed file (one file / team) on the website of the course **(deadline October 2nd)**.

Exercise 1: Increasing the size of the collection

Several collection files of increasing size are available on the website of the course:

```
55k 01-Text_Only-Ascii-Coll-1-10-NoSem.gz
52k 02-Text_Only-Ascii-Coll-11-20-NoSem.gz
103k 03-Text_Only-Ascii-Coll-21-50-NoSem.gz
96k 04-Text_Only-Ascii-Coll-51-100-NoSem.gz
357k 05-Text_Only-Ascii-Coll-101-200-NoSem.gz
559k 06-Text_Only-Ascii-Coll-201-500-NoSem.gz
747k 07-Text_Only-Ascii-Coll-501-1000-NoSem.gz
1.2M 08-Text_Only-Ascii-Coll-1001-2000-NoSem.gz
4.1M 09-Text_Only-Ascii-Coll-2001-5000-NoSem.gz
```

Index each of these files using your indexing program (cf. Practical session n°1: dictionary, postings lists, *df*, *tf*). Build a time efficiency graph of your program (x = size of the collection, y = seconds).

Variants: read several files instead of only one, uncompress a file if it is compressed, insert an option to print or not the index, print the indexing time. For large collections, you must not print the index!

Exercise 2: Collection Statistics

Modify your indexing program so that it computes different statistics on the indexed collection, for instance:

1. document length,
2. term length,
3. vocabulary size,
4. collection frequency of terms.

Plot the evolution of these statistics as the collection size grows.

Exercise 3: Stop-words

Download a stop-words list (cf. list, lecture n°2).
Refresh the index of the 9th file of the exercise n°1, removing stop words.
Compute again the statistics of the exercise n°2.

Exercise 4: Porter's Stemmer

Download a Porter's Stemmer (cf. list, lecture n°2)
Refresh the index of the 9th file of the exercise n°1, applying Porter's stemmer.
Compute again the statistics of the exercise n°2.