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 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 *practice2*.
Compress the directory in a file named *practice2_YourTeamName.zip* (e.g.: *practice2_ VictorAlbertJules.zip*).
Upload this compressed file (one file / team) on the website of the course **(deadline: v1 October 12th ; v2 November 2nd)).**

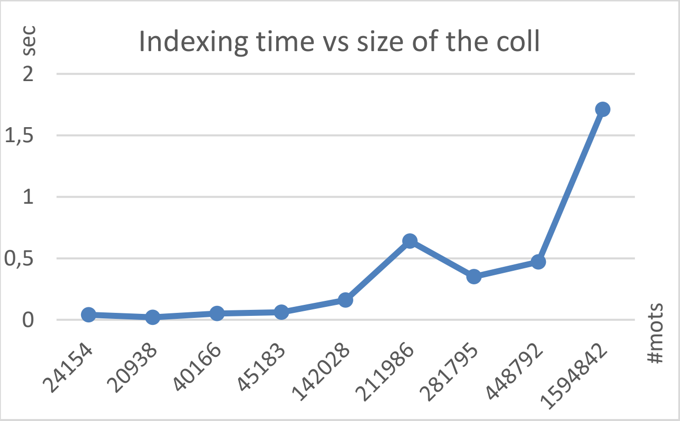
Exercise 1: Increasing the size of the collection

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*). Implement simple tokenization: terms without digits or special characters. Then, build a time-efficiency graph (x-axis = size of the collection, y-axis = 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

2.1) Modify your indexing program so that it computes different statistics on the indexed collection:

1. average documents length (#terms),
2. average terms length (#char),
3. vocabulary size (#distinct terms in the whole collection).

Include these statistics in your report.

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

Exercise 3: Stop-words

3.1) Download this stop-words list on the website of the course: stop-words-english4.txt
Update the index of the 9th file from exercise n°1 by removing stop words. Recompute the statistics from exercise n°2 and include them in your report.

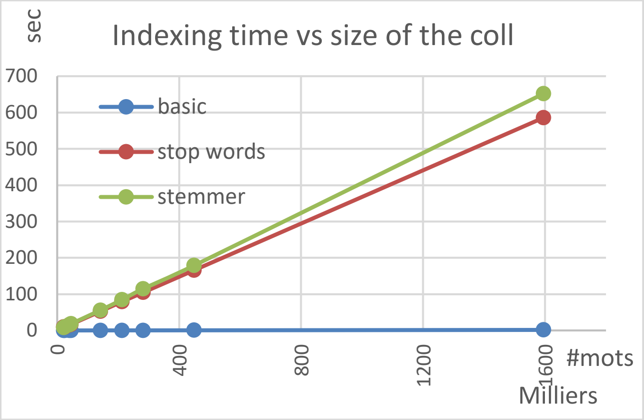
3.2) Plot the evolution of these statistics as the collection size grows.

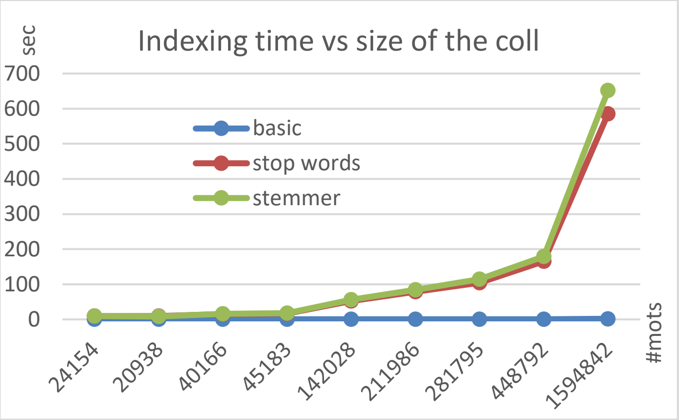
Exercise 4: Stop-words + Porter's Stemmer

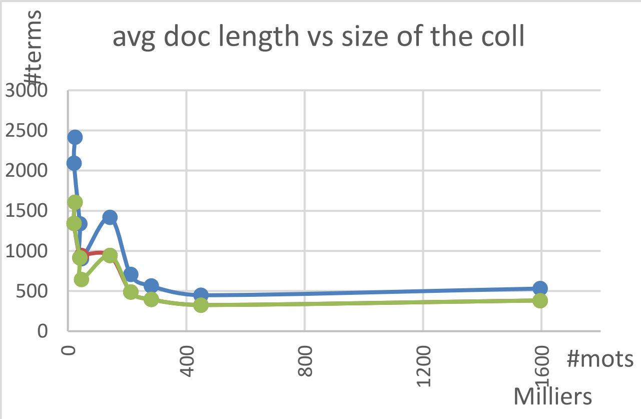
4.1) Download a Porter's Stemmer (cf. list, lecture n°2).
Update the index of the 9th file from exercise n°1 by applying Porter's stemmer. Recompute the statistics from exercise n°2 and include them in your report.

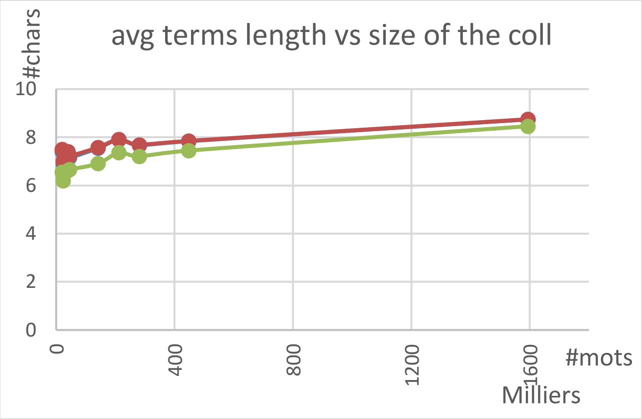
4.2) Plot the evolution of these statistics as the collection size grows.

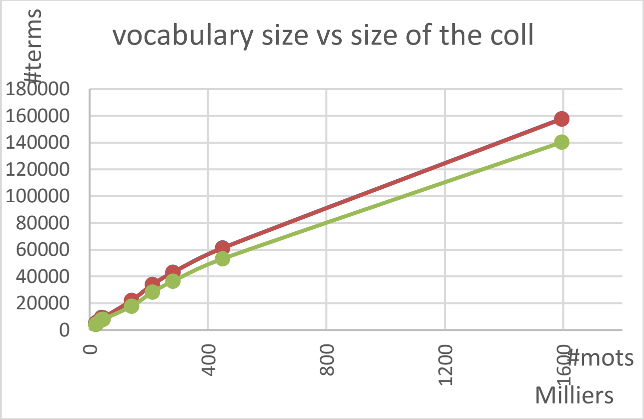
Annexes: graphs


Indexing time vs size of the coll
sec
basic
stop words
stemmer
#mots
Milliers


Indexing time vs size of the coll
sec
basic
stop words
stemmer
#mots


avg doc length vs size of the coll
#terms
#mots
Milliers


avg terms length vs size of the coll
#chars
#mots
Milliers


vocabulary size vs size of the coll
#terms
#mots
Milliers