Information Retrieval

Exercises session n°3: weighting

Create a directory named *exercises3*. In this directory, create a new file named *exercises3_report.txt*.
For each exercise, answer and / or explain in this file.
Add your file *Exercises_03_IR_weighting.xlsx* / *.ods* (cf. below) in the directory *exercises3*.
Compress the directory in a file named *exercises3_YourTeamName.zip* (or *.tar*, *.gz*, *.rar*, etc.) (e.g.: *exercises3_VictorAlbertJulesIsaac.zip*).
Upload this compressed file (one file / team) on the website of the course. **Deadline October 16th** (first try); **October 23th** (final version).

Exercise 1: Jaccard & Log frequency weighting

Compute the Jaccard matching score for the following query-document pairs:

```
q1: information on cars
d1: all you've ever wanted to know about cars

q2: information on cars
d2: information on trucks, information on planes, information on trains

q3: red cars and red trucks
d3: cops stop red cars more often
```

Exercise 2: Count Matrix

The two documents d_1 and d_2 belong to a collection of 1000 XML documents that validate the DTD file *article.dtd*.

```
<!DOCTYPE article [
  <!ELEMENT article (title, abs?)>
  <!ELEMENT title (#PCDATA)>
  <!ELEMENT abs (#PCDATA)>
]>
```
article.dtd

```
<article>
  <title>c d d d e e e e</title>
  <abs>a d e</abs>
</article>
```
d_1

```
<article>
  <title>a b b</title>
  <abs>b b c</abs>
</article>
```
d_2

Collection statistics say that:

- 10 documents on 1000 contain the word "a", 25 contains the word "b", 10 contains the word "c", 24 contains the word "d" and 250 contains the word "e".
- Only 800 XML documents on 1000 contain an *abstract* element.
- The whole collection contains 20 000 words, including 3 000 words appearing in a *title* element.

Give a representation of the index of d_1 and d_2 as a count matrix in the file:
Exercises_03_IR_weighting.xlsx / *.ods*

Exercise 3: SMART *ltn* weighting

In the same file, compute the weight of each term for d_1 and d_2 using the *SMART ltn* weighting function.
Compute the score of d_1 and d_2 considering the query q = "a e".
Which document is the more relevant?

Exercise 4: SMART *ltc* weighting

Same question using the *SMART ltc* weighting function.

Exercise 5: *BM25* weighting

Same question using the *BM25* weighting function, with b=0.5, k_1=1.