

Année Universitaire : 2023/2024 Master 2 : SII Module : Recherche d'Information	Université des Sciences et de la Technologie Houari Boumediene Faculté d'Informatique Département d'Intelligence Artificielle et Sciences des Données	TP N°5 Recherche de l'Information : Appariement Partie 2
---	---	---

Support :

1. Extraction de termes (Tokens) à l'aide de l'expression régulière suivante :

```
nltk.RegexpTokenizer('(?:[A-Za-z]\.)+|[A-Za-z]+[\-@]\d+(?:\.\d+)?|\d+[A-Za-z]+|\d+(?:[\.\,\,]\d+)?%?|\w+(?:[\-\/]\w+)*')
```

2. Appariement :

2.1. Modèle Probabiliste

2.1.1. Modèle BM25

Entrée (requête) :

Un ensemble de termes normalisés

Sortie :

Une liste de documents ordonnés selon leurs degrés de pertinences. Le degré de pertinence **RSV** d'un document **d** par rapport à une requête **Q** est calculé à l'aide de la méthode probabiliste **BM25** comme suit :

$$RSV(Q, d) = \sum_{i \in Q} \left(\frac{freq(t_i, d)}{K \left((1 - B) + B * \frac{dl}{avdl} \right) + freq(t_i, d)} \right) * \left(\log \left(\frac{N - n_i + 0.5}{n_i + 0.5} \right) \right)$$

Avec :

freq(t_i, d) : la fréquence du terme **i** dans le document **d**.

dl : la taille du document **d** (nombre de termes). La taille est calculée sur la base des fréquences des termes.

avdl : la taille moyenne des documents (nombre de termes).

K & B, sont des constantes.

N : le nombre de documents dans la collection.

n_i : le nombre de documents contenant le terme **i**.

log : c'est le Log de **10**.

Exercice :

I. Implémenter la méthode probabiliste BM25.

II. Visualiser les résultats retournés de la méthode probabiliste BM25.

Introduire la requête

Query document ranking **Search** ☐ Queries Dataset 1

Processing ☒ Tokenization ☒ Porter Stemmer

Index ☐ ☒ DOCS per TERM ☐ TERMS per DOC

Matching ☐ Vector Space Model ☒ Probabilistic Model (BM25) ☐ Boolean Model ☐ Data Mining Model

Scalar Product

K 1.50
B 0.75

Results

N°doc	Relevance
2	-0.4531
3	-0.4597
6	-0.4771
5	-0.5028
4	-0.5372

Résultats retournés par le
modèle probabiliste **BM25**