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 $\boldsymbol{RSV}$ d'un document $\boldsymbol{d}$ par rapport à une requête $\boldsymbol{Q}$ est calculé à l'aide de la méthode probabiliste **BM25** comme suit :

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

Avec :

$\boldsymbol{freq(t_i,d)}$: la fréquence du terme $\boldsymbol{i}$ dans le document $\boldsymbol{d}$.
$\boldsymbol{dl}$ **:** la taille du document $\boldsymbol{d}$ (nombre de termes).
$\boldsymbol{avdl}$: la taille moyenne des documents (nombre de termes).
$\boldsymbol{K}$ & $\boldsymbol{B}$, sont des constantes.
$\boldsymbol{N}$: le nombre de documents dans la collection.
$\boldsymbol{n_i}$: le nombre de documents contenant le terme $\boldsymbol{i}$.
$\boldsymbol{\log}$: 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

Résultats retournés par le modèle probabiliste **BM25**