Université des Sciences et de la Technologie Houari Boumediene	
Faculté d'Informatique	

Département d'Intelligence Artificielle et Sciences des Données

TP N°2Représentation de l'Information : Indexation

Partie 2

Support:

Master 2: SII

Année Universitaire: 2024/2025

Module: Recherche d'Information

1. Création d'un fichier descripteur basé sur les fréquences

```
Création d'un dictionnaire :
>>> TermesFrequence = {}
>>> for terme in TermesNormalisation:
        if (terme in TermesFrequence.keys()):
              TermesFrequence[terme] += 1
        else:
              TermesFrequence[terme] = 1
>>> TermesFrequence
>>> {'d.z.': 1, 'post': 1, 'print': 1, 'cost': 1, '120.50da': 1, '...': 1}
>>> TermesFrequence.keys()
>>> dict keys(['d.z.', 'post', 'print', 'cost', '120.50da', '...'])
>>> TermesFrequence.items()
>>> dict items([('d.z.', 1), ('post', 1), ('print', 1), ('cost', 1), ('120.50da', 1), ('...', 1)])
>>> TermesFrequence = nltk.FreqDist(TermesNormalisation)
>>> TermesFrequence
>>> FreqDist({'d.z.': 1, 'post': 1, 'print': 1, 'cost': 1, '120.50da': 1, '...': 1})
Tri d'un dictionnaire:
>>> collections.OrderedDict(sorted(TermesFrequence.items()))
```

2. Pondération des termes normalisés

$$poids(t_i, d_j) = \left(\frac{freq(t_i, d_j)}{Max(freq(t, d_j))}\right) * \log\left(\frac{N}{n_i} + 1\right)$$

Avec:

```
poids(t_i, d_j): le poids du terme i dans le document j. freq(t_i, d_j): la fréquence du terme i dans le document j. Max(freq(t, d_j)): la fréquence max dans le document j. N: le nombre de documents dans la collection. n_i: le nombre de documents contenant le terme i. log: log: 10.
```

Exercice:

I. Création des index :

. Mettre à jour les fichiers descripteurs, comme suit :

<N° document> <Terme> <Fréquence> <Poids>

. Mettre à jour les fichiers inverses, définis comme suit :

<Terme> <N° document> <Fréquence> <Poids>

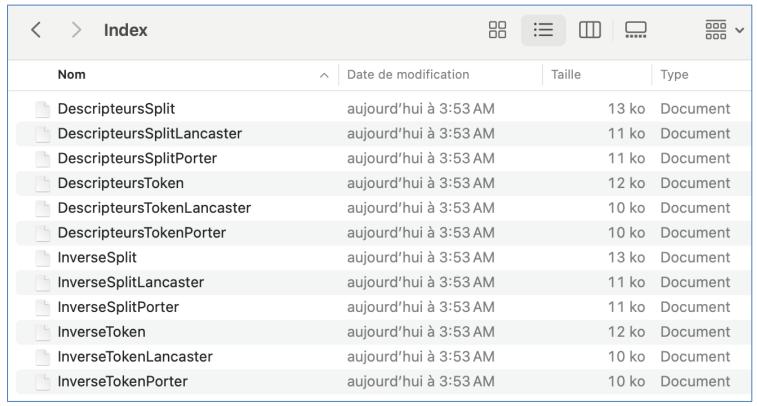


Fig.1 – Index à mettre à jour

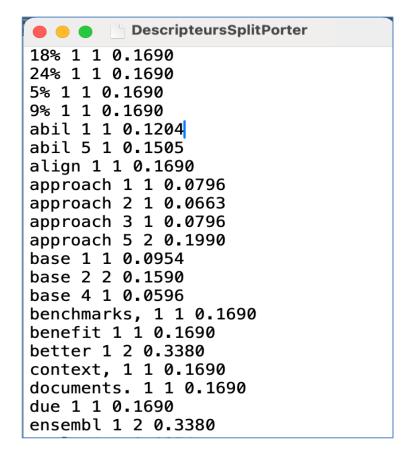


Fig.2 (a) – DescripteursSplitPorter

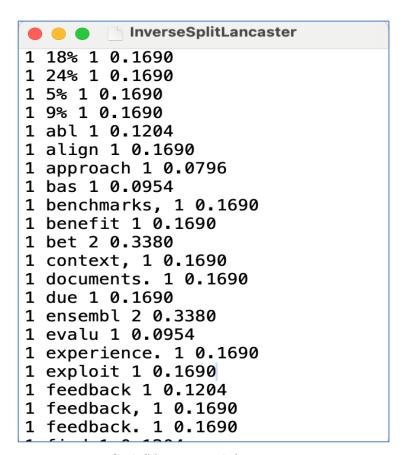
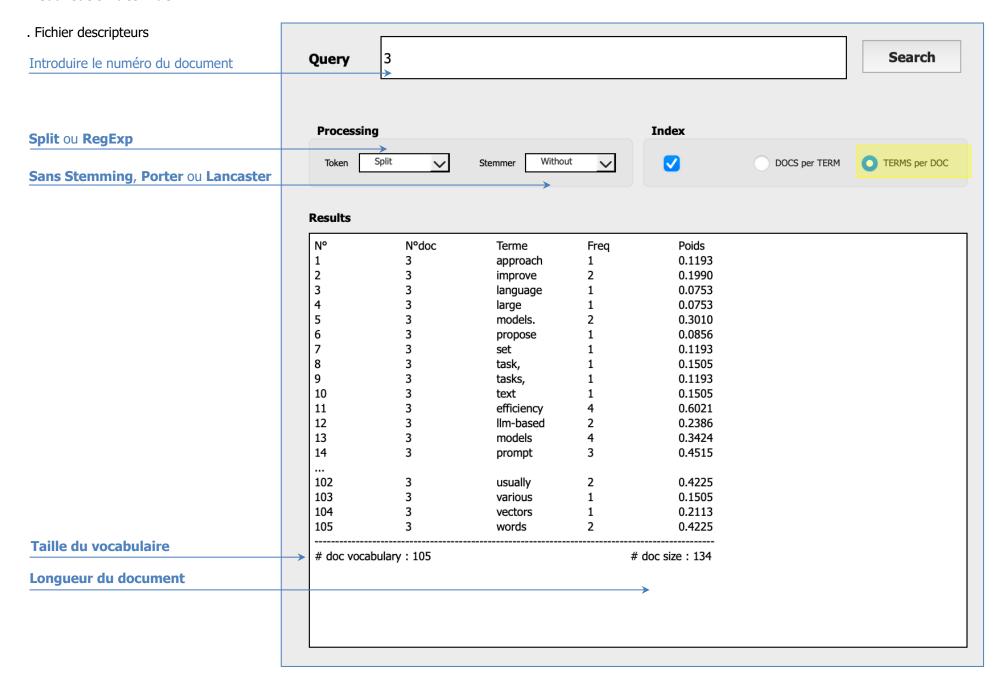


Fig.2 (b) – InverseSplitLancaster

II. Visualisation des index:



. Fichier inverse

