

Support :

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>













Index				
Nom	^	Date de modification	Taille	Type
 DescripteursSplit		aujourd'hui à 3:53 AM	13 ko	Document
 DescripteursSplitLancaster		aujourd'hui à 3:53 AM	11 ko	Document
 DescripteursSplitPorter		aujourd'hui à 3:53 AM	11 ko	Document
 DescripteursToken		aujourd'hui à 3:53 AM	12 ko	Document
 DescripteursTokenLancaster		aujourd'hui à 3:53 AM	10 ko	Document
 DescripteursTokenPorter		aujourd'hui à 3:53 AM	10 ko	Document
 InverseSplit		aujourd'hui à 3:53 AM	13 ko	Document
 InverseSplitLancaster		aujourd'hui à 3:53 AM	11 ko	Document
 InverseSplitPorter		aujourd'hui à 3:53 AM	11 ko	Document
 InverseToken		aujourd'hui à 3:53 AM	12 ko	Document
 InverseTokenLancaster		aujourd'hui à 3:53 AM	10 ko	Document
 InverseTokenPorter		aujourd'hui à 3:53 AM	10 ko	Document

Fig.1 – Index à mettre à jour

```
DescripteursSplitPorter
18% 1 1 0.1690
24% 1 1 0.1690
5% 1 1 0.1690
9% 1 1 0.1690
abil 1 1 0.1204
abil 5 1 0.1505
align 1 1 0.1690
approach 1 1 0.0796
approach 2 1 0.0663
approach 3 1 0.0796
approach 5 2 0.1990
base 1 1 0.0954
base 2 2 0.1590
base 4 1 0.0596
benchmarks, 1 1 0.1690
benefit 1 1 0.1690
better 1 2 0.3380
context, 1 1 0.1690
documents. 1 1 0.1690
due 1 1 0.1690
ensembl 1 2 0.3380
```

Fig.2 (a) – DescripteursSplitPorter

```
InverseSplitLancaster
1 18% 1 0.1690
1 24% 1 0.1690
1 5% 1 0.1690
1 9% 1 0.1690
1 abl 1 0.1204
1 align 1 0.1690
1 approach 1 0.0796
1 bas 1 0.0954
1 benchmarks, 1 0.1690
1 benefit 1 0.1690
1 bet 2 0.3380
1 context, 1 0.1690
1 documents. 1 0.1690
1 due 1 0.1690
1 ensembl 2 0.3380
1 evalu 1 0.0954
1 experience. 1 0.1690
1 exploit 1 0.1690
1 feedback 1 0.1204
1 feedback, 1 0.1690
1 feedback. 1 0.1690
```

Fig.2 (b) – InverseSplitLancaster

II. Visualisation des index :

. Fichier descripteurs

Introduire le numéro du document

Split ou RegExp

Sans Stemming, Porter ou Lancaster

Taille du vocabulaire

Longueur du document

Query

Processing

Token Stemmer

Index
☒ ☐ DOCS per TERM ☒ TERMS per DOC

Results

N°	N°doc	Terme	Freq	Poids
1	3	approach	1	0.1193
2	3	improve	2	0.1990
3	3	language	1	0.0753
4	3	large	1	0.0753
5	3	models.	2	0.3010
6	3	propose	1	0.0856
7	3	set	1	0.1193
8	3	task,	1	0.1505
9	3	tasks,	1	0.1193
10	3	text	1	0.1505
11	3	efficiency	4	0.6021
12	3	llm-based	2	0.2386
13	3	models	4	0.3424
14	3	prompt	3	0.4515
...				
102	3	usually	2	0.4225
103	3	various	1	0.1505
104	3	vectors	1	0.2113
105	3	words	2	0.4225

doc vocabulary : 105 # doc size : 134

. Fichier inverse

Introduire un terme

Query

Processing
Token Stemmer

Index
☒ ☒ DOCS per TERM ☐ TERMS per DOC

Results

N°	Terme	N°doc	Freq	Poids
1	gpt-3.5	6	1	0.1056