

Année Universitaire : 2024/2025 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	<b>TP N°2</b> Représentation de l'Information : Indexation Partie 2
---	---	---

## 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>

Nom	Modifié le	Type	Taille
DescripteurSplit	20/10/2024 11:34	Fichier	9 Ko
DescripteurSplitLancaster	20/10/2024 11:34	Fichier	8 Ko
DescripteurSplitPorter	20/10/2024 11:34	Fichier	8 Ko
DescripteurToken	20/10/2024 11:34	Fichier	9 Ko
DescripteurTokenLancaster	20/10/2024 11:34	Fichier	7 Ko
DescripteurTokenPorter	20/10/2024 12:57	Fichier	8 Ko
InverseSplit	20/10/2024 11:34	Fichier	9 Ko
InverseSplitLancaster	20/10/2024 11:34	Fichier	8 Ko
InverseSplitPorter	20/10/2024 11:34	Fichier	8 Ko
InverseToken	20/10/2024 11:34	Fichier	9 Ko
InverseTokenLancaster	20/10/2024 11:34	Fichier	7 Ko
InverseTokenPorter	20/10/2024 11:34	Fichier	8 Ko

**Fig.1** – Index à mettre à jour

```
DescripteurSplitLancaster - Blo...
Fichier Edition Format Affichage Aide
1 . 6 0.3010
1 /destalling/ 1 0.1408
1 aerodynam 1 0.1003
1 agr 1 0.1408
1 angl 1 0.1408
1 attack 1 0.1408
1 bas 1 0.0795
1 boundary-layer-control 1 0.1408
1 comp 1 0.1408
1 config 1 0.1408
1 curves, 1 0.1408
1 destal 2 0.2817
1 determin 1 0.1408
1 diff 3 0.2386
1 distribut 1 0.0663
1 due 2 0.2817
1 effect 2 0.1326
1 empir 1 0.1408
1 evalu 2 0.2817
1 evidence, 1 0.1408
1 experty 3 0.4225
1 flow 1 0.0571
1 found 1 0.1408
1 fre 1 0.1003
1 incr 1 0.1408
1 increas 1 0.1003
1 increment, 1 0.1408
1 integr 1 0.0795
1 intend 1 0.1408
1 investig 1 0.1003
1 lift 3 0.4225
1 lift, 1 0.1003
1 load 1 0.1408
1 mad 2 0.2007
1 ord 1 0.1408
1 part 2 0.2817
1 pot 1 0.1408
Ln 37, C 100% Unix (LF) UTF-8
```

**Fig.2 (a) – DescripteursSplitLancaster**

```
InverseSplitPorter - Bloc-notes
Fichier Edition Format Affichage Aide
. 1 6 0.3010
. 2 10 0.3010
. 3 4 0.3010
. 4 12 0.3010
. 5 12 0.3010
. 6 4 0.3010
/destalling/ 1 1 0.1408
aerodynam 1 1 0.1003
aerodynam 6 1 0.1505
agre 1 1 0.1408
angl 1 1 0.1408
attack 1 1 0.1408
basi 1 1 0.1003
basi 4 1 0.0502
boundary-layer-control 1 1 0.1408
compar 1 1 0.1408
configur 1 1 0.1408
curves, 1 1 0.1408
destal 1 2 0.2817
determin 1 1 0.1408
differ 1 3 0.2386
differ 2 1 0.0477
differ 5 2 0.0795
distribut 1 1 0.0663
distribut 3 1 0.0995
distribut 4 1 0.0332
distribut 6 1 0.0995
due 1 2 0.2817
effect 1 2 0.1326
effect 2 1 0.0398
effect 3 1 0.0995
effect 6 3 0.2985
empir 1 1 0.1408
evalu 1 2 0.2817
evidence, 1 1 0.1408
experi 1 1 0.1408
experiment 1 2 0.2817
Ln 1, Col 100% Unix (LF) UTF-8
```

**Fig.2 (b) – InverseSplitPorter**

```
InverseSplit - Bloc-notes
Fichier Edition Format Affichage Aide
. 1 6 0.3010
. 2 10 0.3010
. 3 4 0.3010
. 4 12 0.3010
. 5 12 0.3010
. 6 4 0.3010
/destalling/ 1 1 0.1408
aerodynamics 1 1 0.1408
agree 1 1 0.1408
angles 1 1 0.1408
attack 1 1 0.1408
basis 1 1 0.1003
basis 4 1 0.0502
boundary-layer-control 1 1 0.1408
comparative 1 1 0.1408
configuration 1 1 0.1408
curves, 1 1 0.1408
destalling 1 2 0.2817
determine 1 1 0.1408
different 1 3 0.3010
different 2 1 0.0602
distribution 1 1 0.0663
distribution 3 1 0.0995
distribution 4 1 0.0332
distribution 6 1 0.0995
due 1 2 0.2817
effect 1 1 0.0795
effect 3 1 0.1193
effect 6 1 0.1193
effects 1 1 0.0795
effects 2 1 0.0477
effects 6 2 0.2386
empirical 1 1 0.1408
evaluation 1 2 0.2817
evidence, 1 1 0.1408
experiment 1 1 0.1408
experimental 1 2 0.2817
Ln 1, C 100% Unix (LF) UTF-8
```

**Fig.2 (c) – InverseSplit**

## II. Visualisation des index :

. Fichier descripteurs

Introduire le numéro du document

**Split** ou **RegExp**

**No stem** ou **Porter Stemmer** ou **Lancaster**

**Query**  **Search**

**Processing**

**Tokenization**

☒ Split  
☐ RegExp

**Normalization**

☒ No stem  
☐ Porter Stemmer  
☐ Lancaster Stemmer

**Index**

☒ ☐ DOCS per TERM ☒ **TERMS per DOC**

**Results**

N°	N°doc	Terme	Freq	Poids
1	5	.	12	0.3010
2	5	flow	5	0.1427
3	5	theory	4	0.2007
4	5	considered	1	0.0398
5	5	hypersonic	3	0.1505
6	5	layer	2	0.0663
7	5	possible	1	0.0502
8	5	region	1	0.0502
9	5	shock	5	0.2509
10	5	simple	2	0.1003
11	5	wave	2	0.1003
12	5	equations	1	0.0502
13	5	uniform	1	0.0502
14	5	axial	1	0.0502
15	5	bodies	1	0.0502
16	5	calculated	1	0.0398
17	5	developed	1	0.0502
18	5	equation	1	0.0502
19	5	number	1	0.0502
20	5	surface	3	0.1505
21	5	..	1	0.0704
22	5	0.1	1	0.0704

## I. Visualisation des index :

. Fichier inverse

Introduire un terme

**Query**

**Processing**

**Tokenization**  
☒ Split  
☐ RegExp

**Normalization**  
☒ No stem  
☐ Porter Stemmer  
☐ Lancancer Stemmer

**Index**  
☒ ☒ DOCS per TERM ☐ TERMS per DOC

**Results**

N°	Terme	N°doc	Freq	Poids
1	theory	1	1	0.1003
2	theory	5	4	0.2007