

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°2 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 :

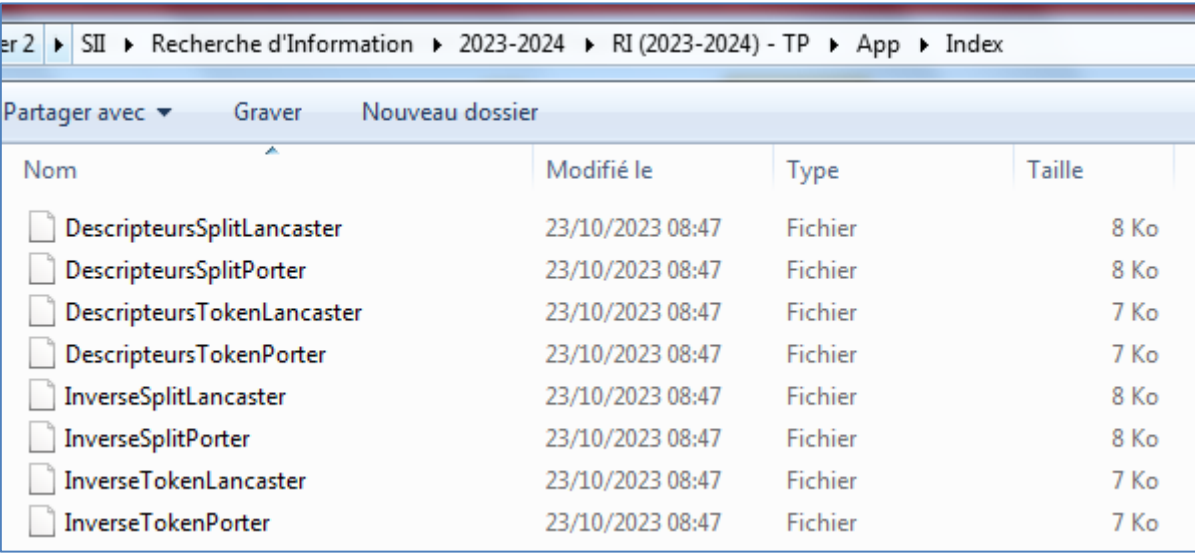
I. Création des index :

- . 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>



The screenshot shows a Windows File Explorer window with the address bar displaying the path: "er 2 > SII > Recherche d'Information > 2023-2024 > RI (2023-2024) - TP > App > Index". The window contains a table of files with the following columns: "Nom", "Modifié le", "Type", and "Taille". There are eight files listed, all of which are "Fichier" type and were last modified on "23/10/2023 08:47". The files are: "DescripteursSplitLancaster" (8 Ko), "DescripteursSplitPorter" (8 Ko), "DescripteursTokenLancaster" (7 Ko), "DescripteursTokenPorter" (7 Ko), "InverseSplitLancaster" (8 Ko), "InverseSplitPorter" (8 Ko), "InverseTokenLancaster" (7 Ko), and "InverseTokenPorter" (7 Ko).

Nom	Modifié le	Type	Taille
DescripteursSplitLancaster	23/10/2023 08:47	Fichier	8 Ko
DescripteursSplitPorter	23/10/2023 08:47	Fichier	8 Ko
DescripteursTokenLancaster	23/10/2023 08:47	Fichier	7 Ko
DescripteursTokenPorter	23/10/2023 08:47	Fichier	7 Ko
InverseSplitLancaster	23/10/2023 08:47	Fichier	8 Ko
InverseSplitPorter	23/10/2023 08:47	Fichier	8 Ko
InverseTokenLancaster	23/10/2023 08:47	Fichier	7 Ko
InverseTokenPorter	23/10/2023 08:47	Fichier	7 Ko

Fig.1 – Index à mettre à jour

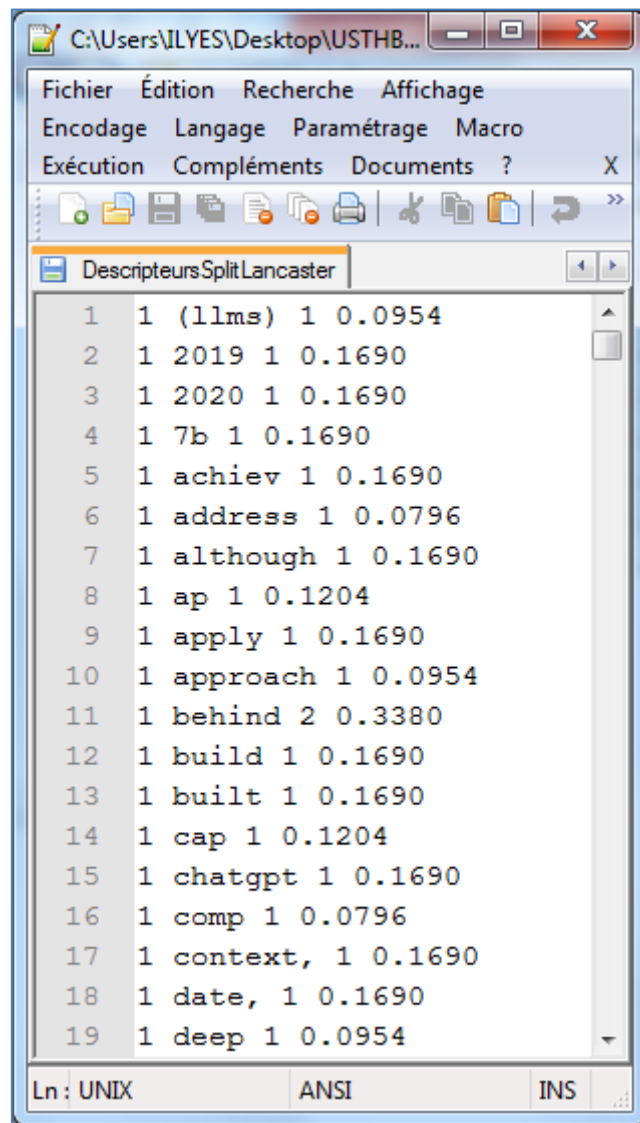


Fig.2 (a) – DescripteursSplitLancaster

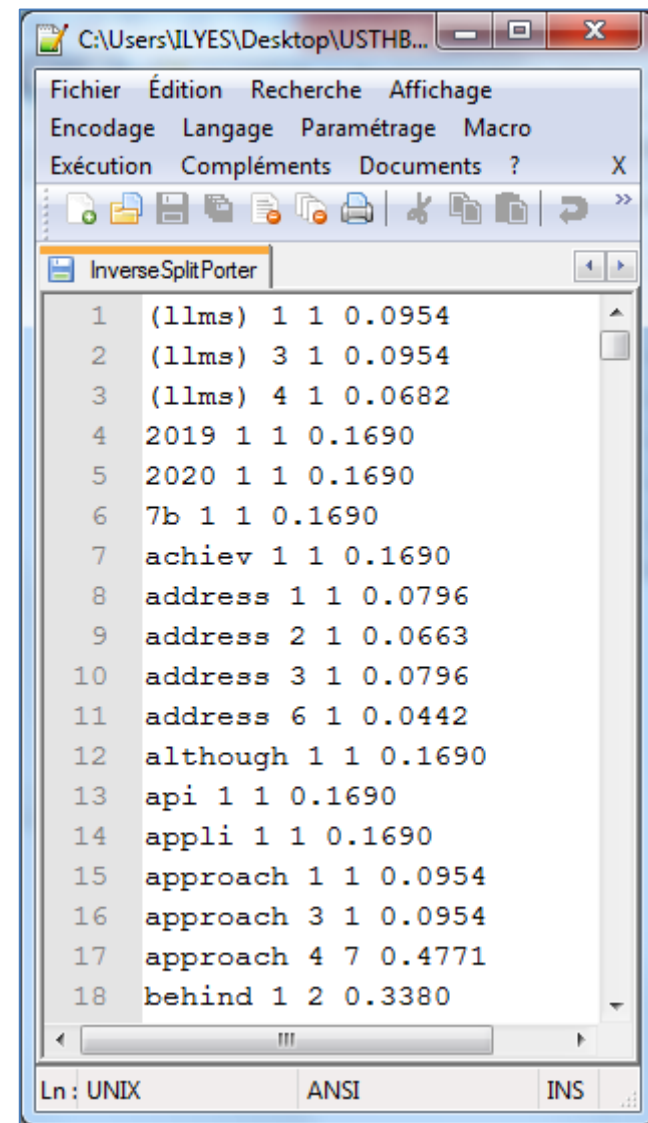


Fig.2 (b) – InverseSplitPorter

II. Visualisation des index :

. Fichier descripteurs

Introduire le numéro du document

Split ou RegExp

Porter Stemmer ou Lancaster Stemmer

Query

Processing
☐ Tokenization ☒ Porter Stemmer

Index
☒ ☐ DOCS per TERM ☒ TERMS per DOC

Results

N°	N°doc	Terme	Freq	Poids
1	1	(llms)	1	0.0954
2	1	2019	1	0.1690
3	1	2020	1	0.1690
4	1	7b	1	0.1690
5	1	achiev	1	0.1690
6	1	address	1	0.0796
7	1	although	1	0.1690
8	1	api	1	0.1690
9	1	appli	1	0.1690
10	1	approach	1	0.0954
11	1	behind	2	0.3380
12	1	build	1	0.1690
13	1	built	1	0.1690
14	1	capabl	1	0.1204
15	1	chatgpt	1	0.1690
16	1	compar	1	0.0796
17	1	context,	1	0.1690
18	1	date,	1	0.1690
19	1	deep	1	0.0954
20	1	effect	2	0.1370
21	1	endpoints.	1	0.1690
22	1	experiment	2	0.3380

I. Visualisation des index :

. Fichier inverse

Introduire un terme

Query

Processing
☒ Tokenization ☒ Porter Stemmer

Index
☒ ☒ DOCS per TERM ☐ TERMS per DOC

Results

N°	Terme	N°doc	Freq	Poids
1	gpt-3	1	1	0.1690