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>`

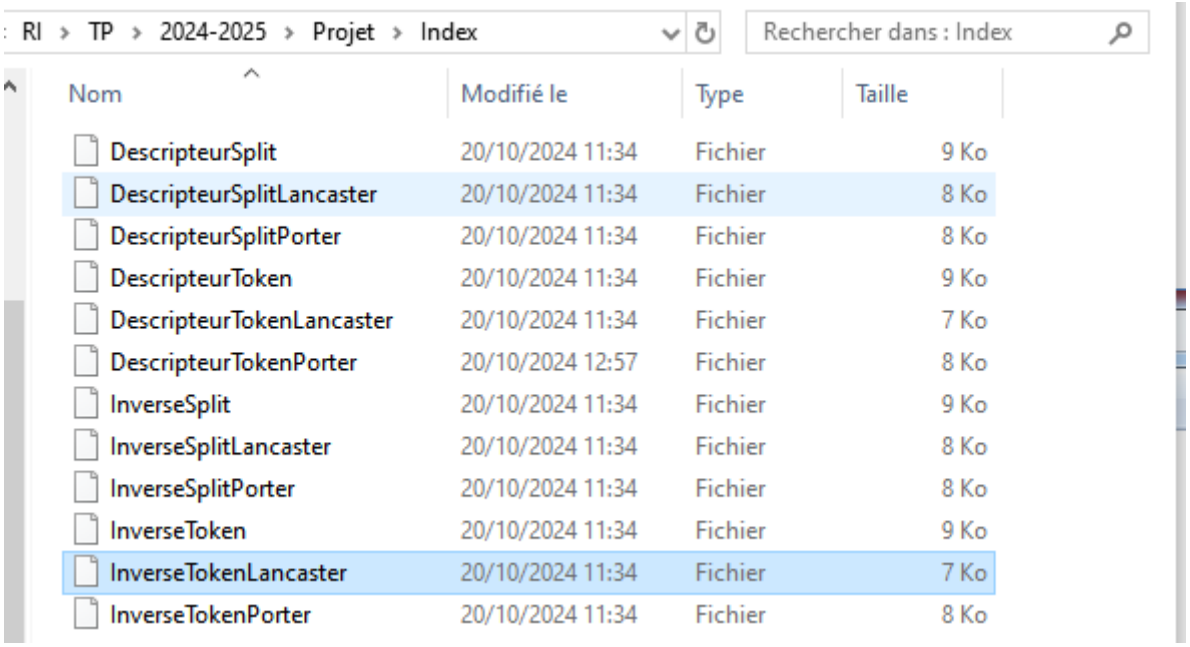


Fig.1 – Index à mettre à jour

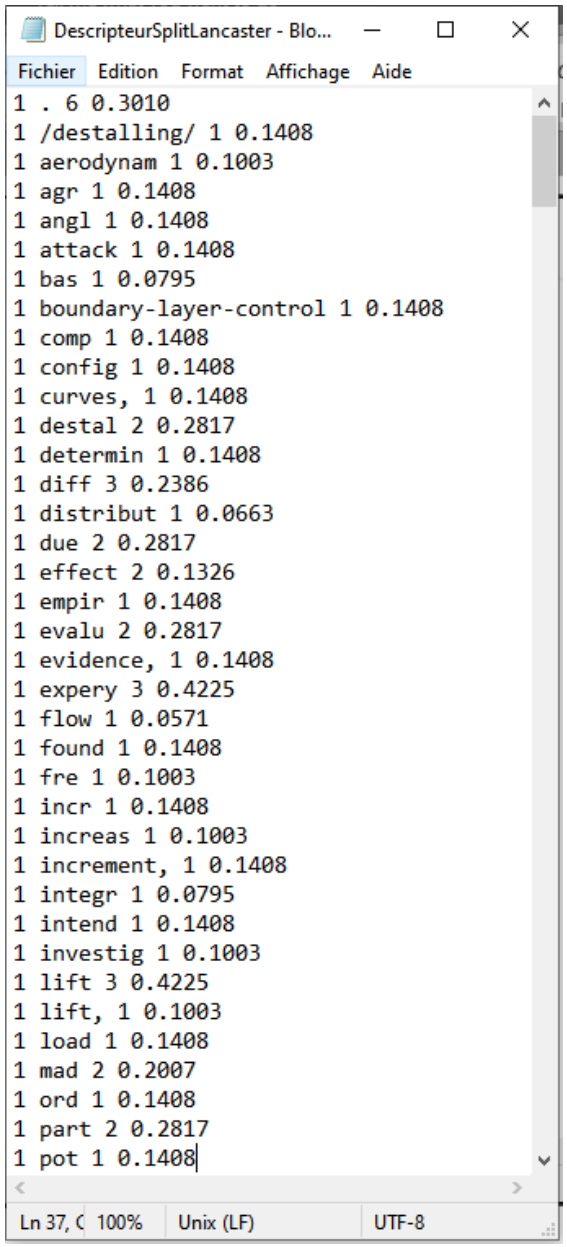


Fig.2 (a) – DescripteursSplitLancaster

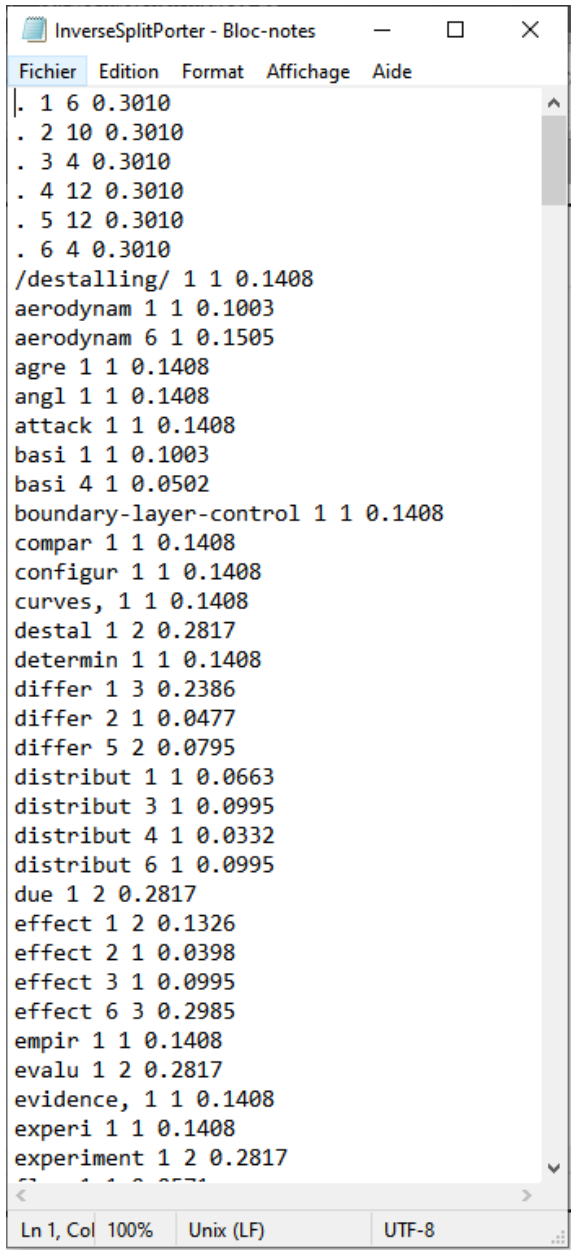


Fig.2 (b) – InverseSplitPorter

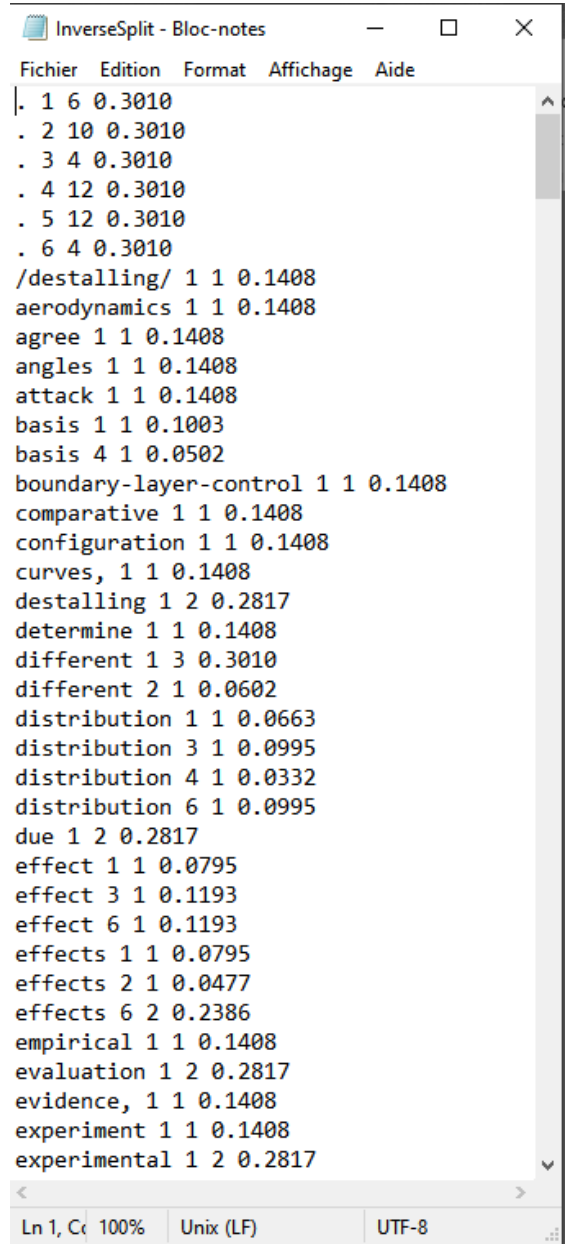


Fig.2 (c) – InverseSplit

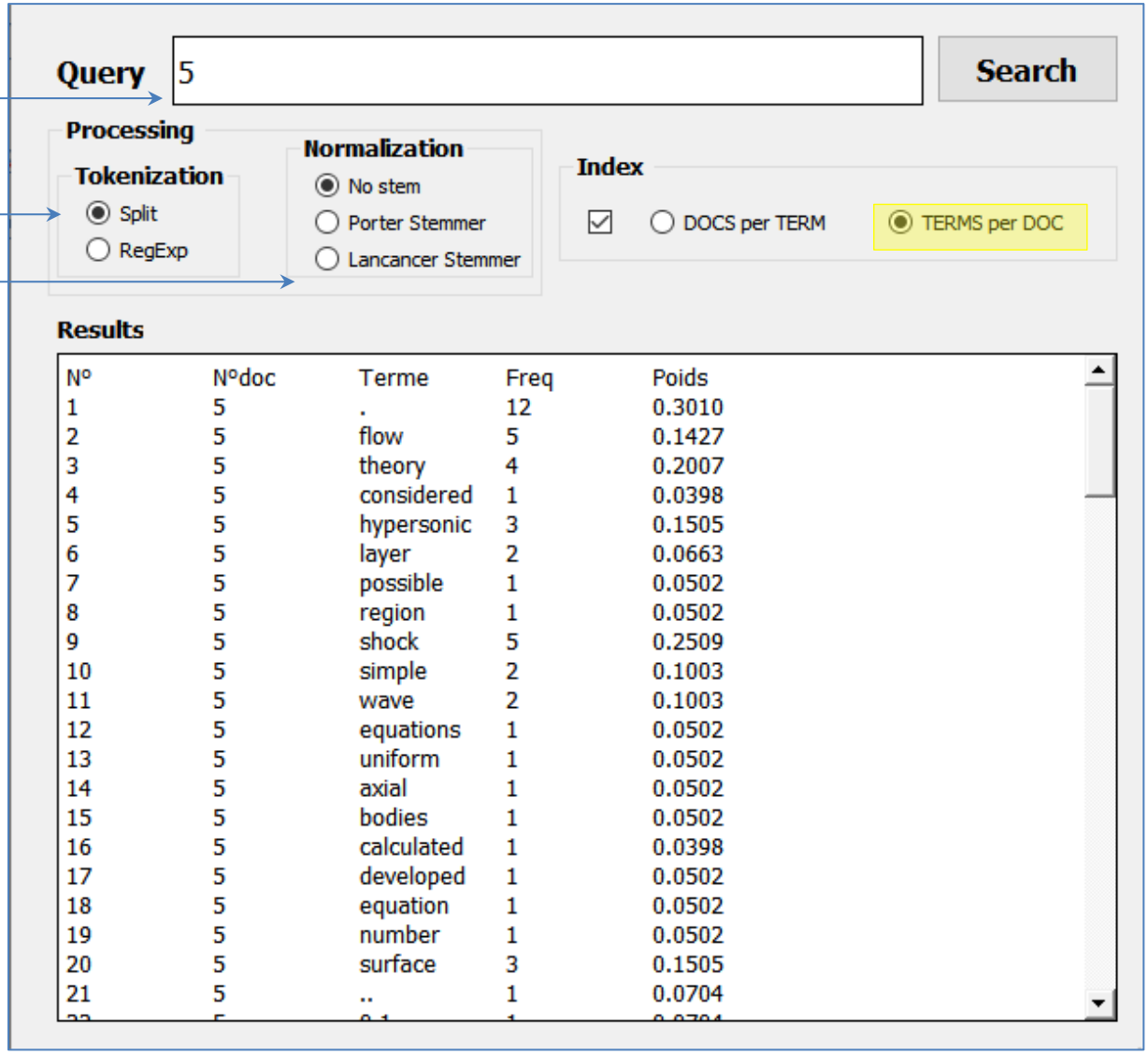
II. Visualisation des index :

. Fichier descripteurs

Introduire le numéro du document

Split ou **RegExp**

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



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

I. Visualisation des index :

. Fichier inverse

Introduire un terme