Collection de documents textuels

Extraction de la liste des termes présents au moins 1 fois

Matrice de données représentant le corpus, avec choix de pondération

Corpus

Dictionnaire

Matrice documents-termes

Exemple :
<doc A>*the movie begins in the past where a young boy named sam attempts to save celebi from a hunter .* ***</doc>***
<doc B>*she , among others excentricities , talks to a small rock , gertrude , like if she was alive .* ***</doc>***
<doc C> *etc…* ***</doc>***
Etc….

{the, movie, begins, in, the, past, where, young, boy, named, sam, attempts, …}

doc	the	movie	begins	…	she	among	…
A	1	1	1	…	0	0	…
B	0	0	0	…	2	1	…
C	…						

Algorithme des K-Means