Nous utilisons plusieurs fonctions pour récupérer le contenu des corpus et traiter le texte. Nous utilisons la bibliothèque NLTK.

Nous « tokenisons » le texte et nous supprimons les mots qui ne sont pas « alphabétiques » afin de supprimer les titres de tableaux, les années, mots incorrects, etc…

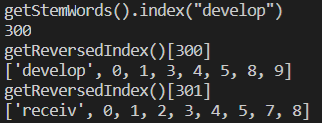
Nous appliquons ensuite un lemmatizer sur les mots, puis nous supprimons les « stop words ». Ensuite nous utiliser la classe PorterStemmer() pour stem notre liste de mots. Nous retournons ensuite une liste sans duplicatas dans un dictionnaire (grâce à la clé de la classe dictionnaire).

Nous avons deux fonctions nous permettant de récupérer le contenu des corpus (utilisant des fonctions différentes également) :

* getStemWords() permet de récupérer le dictionnaire de tous les corpus mélangés, afin d’obtenir la liste réel de tous les mots présents dans nos corpus.
* getStemWordsByCorpus() nous permet de récupérer 10 listes, une pour chaque corpus, afin de fabriquer nos index plus tard.

Avec ces deux listes ainsi obtenus, nous pouvons avec la fonction getIncidenceMatrix() générer la matrice d’incidence.

La fonction getReversedIndex() nous permet de récupérer l’index inversé pour chaque corpus dans le dossier « corpus ».



La fonction booleanRequest() nous permet ensuite de faire des requêtes utilisant l’index / la matrice d’incidence.