

Activité 8 : Analyse du champ lexical de façon automatique

Définition du champ lexical : *Un champ lexical désigne un ensemble de noms, d'adjectifs et de verbes liés par leur sémantique, c'est-à-dire traitant d'un domaine commun.*

Objectif : Quels sont les mots qui sont le plus souvent associés au champ lexical lié au mot dessin animé ?

Utiliser un fichier magazine « avec espace entre les mots » disponible sur picassciences (*peu importe la taille ici, les temps de calcul sont raisonnables*)

Commencer un script vierge, en repartant du résultat de l'activité 1 qui se présente sous forme suivante, en ajoutant les lignes en gras :

```
fichier_magazine = open('G:\\Mes Cours\\NSI\\1ere\\Mes
cours\\projetRetroMagazineSearch\\Joystick.txt', 'r',
encoding="utf-8")

lignes_magazine = fichier_magazine.readlines()

Drapeau = 0
liste_pages_resultats = []

liste_tous_les_mots_d_une_page = []
dictionnaire_tous_les_mots = {}

for i in range(0,len(lignes_magazine)):

    if "dessin anim" in lignes_magazine[i]:
        Drapeau = 1

        liste_tous_les_mots_d_une_page = #compléter
        print(liste_tous_les_mots_d_une_page)
        break #arret pour analyser vos résultats

    if Drapeau == 1:
        if '----NOUVELLEPAGE----' in lignes_magazine[i]:
            liste_pages_resultats.append(lignes_magazine[i])
            Drapeau = 0

print(liste_pages_resultats)
print(len(liste_pages_resultats))
```

Etapes de résolution :

1. Quand le terme de recherche est trouvé, utiliser la méthode `.split(" ")` pour stocker tous les mots de la page dans la liste `liste_tous_les_mots_d_une_page`.
2. Remplir le dictionnaire `dictionnaire_tous_les_mots` avec
 - Comme clé : les mots
 - Comme valeur : le nombre de mots rencontrés

Vous obtenez un dictionnaire de tous les mots et le nombre d'occurrences, mais celui-ci n'est pas trié.

3. Classer votre dictionnaire par ordre de nombre d'occurrences croissant. (on utilise ici une fonction intégrée de tri dans Python exemple :

```
dictionnaire_tous_les_mots =  
sorted(dictionnaire_tous_les_mots.items(), key=lambda x:x[1])
```

Le résultat obtenu ressemble à ceci :

```
-----  
, 650), ('ou', 659), ('sont', 684), ('tout', 724), ('avec', 772), ('mais', 801),  
, ('par', 860), ('on', 866), ('se', 924), ('ne', 941), ('au', 946), ('il', 990),  
, ('plus', 1005), ('jeu', 1044), (':', 1069), ('ce', 1348), ('pas', 1396), ('sur',  
, 1400), ('une', 1475), ('qui', 1515), ('dans', 1571), ('du', 1580), ('est', 1691  
, ('pour', 1710), ('vous', 1897), ('que', 1951), ('en', 2410), ('des', 2469), (  
, 'un', 2730), ('les', 3362), ('a', 3514), ('la', 3663), ('et', 3762), ('le', 3939  
, ('', 7974), ('de', 8553)]
```

Le résultat est peu exploitable puisqu'il est pollué par des articles, des conjonctions, etc... des mots ne portant que peu de sens.

Améliorons le résultat ...

4. Piste d'amélioration 1 : n'ajouter que les mots de 4 lettres ou plus au dictionnaire.
5. Piste d'amélioration 2 : ne pas ajouter les « stopwords ». Vous trouverez le script 8 sur [picassciences](#) comportant une liste de « stopwords »

Définition d'un « stopwords » :

En recherche d'information, un mot vide (ou stop word, en anglais) est un mot qui est tellement commun qu'il est inutile de l'indexer ou de l'utiliser dans une recherche. En français, des mots vides évidents pourraient être « le », « la », « de », « du », « ce ». Un mot vide est un mot non significatif figurant dans un texte. On l'oppose à mot plein. La signification d'un mot s'évalue à partir de sa distribution (au sens statistique) dans une collection de textes. Un mot est dit « vide » si sa distribution est uniforme sur les textes de la collection. En d'autres termes, un mot qui apparaît avec une fréquence semblable dans chacun des textes de la collection n'est pas discriminant car il ne permet pas de distinguer les textes les uns par rapport aux autres.

6. Exécuter l'algorithme sur chaque magazine (ne garder que le top 50), puis sur l'intégralité des magazines.