TP1 - Analyse de corpus de texte

Première partie : Analyse et comparaison de corpus

Le but de cette première séance est d'apprendre à manipuler des corpus de texte en différents formats, différentes langues, afin de réaliser des études lexicométriques simples permettant de qualifier chaque corpus à travers son lexique et la distribution de ses éléments.

Mise en pratique

- Partie 1 : Manipulation de corpus
    - Vous développerez la partie pratique de ce cours sous Unix, dans un environnement de type *shell* tel que le *Bash shell* disponible généralement par défaut dans les distributions Ubuntu.
    - Pour les programmes de manipulation de données que vous développerez, vous avez toute liberté pour le langage de programmation, néanmoins vous êtes encouragé à utiliser autant que possible les outils commandes Unix du shell permettant par exemple de trier (*sort*) et compter (*uniq -c*) des éléments, ainsi que l'enchaînement de commandes dans des scripts écris en *bash*. Une des contraintes principales auxquels vous serez confronter est la vitesse de traitement, certains corpus pouvant être de taille très importante.
    - Vous trouverez sur les liens suivants l'ensemble des corpus :
        - corpus_multilingue.tgz (*39 Mo*) : répertoire contenant des corpus de dépêches de presse au format XML dans 4 langues (Anglais, Arabe, Espagnol, Français) - https://pageperso.lis-lab.fr/frederic.bechet/M1_TAL_TP_data/corpus_multilingue.tgz
        - corpus_topic.tgz (*1.1 Mo*) : répertoire contenant des corpus des dépêches de presse en mode *plain text* en français, classées selon 4 thématiques (Culture, Economie, International, Sport) - https://pageperso.lis-lab.fr/frederic.bechet/M1_TAL_TP_data/corpus_topic.tgz
        - frwiki-from-polyglot.txt.gz (*903.5 Mo*) : dump (ancien) de *Wikipedia* pour le français en mode *plain text* - *https://pageperso.lis-lab.fr/frederic.bechet/M1_TAL_TP_data/frwiki-from-polyglot.txt.gz*
        - frwiki-sample.txt.gz (*19.2 Mo*) : extrait du dump (ancien) de *Wikipedia* pour le français en mode *plain text* - *https://pageperso.lis-lab.fr/frederic.bechet/M1_TAL_TP_data/frwiki-sample.txt.gz*
        - orfeo_dump.txt.gz (*6 Mo*) : corpus de transcription de conversations orales au format *plain text* - *https://pageperso.lis-lab.fr/frederic.bechet/M1_TAL_TP_data/orfeo_dump.txt.gz*

- Vous pouvez télécharger les corpus séparemment, selon vos besoins (en particulier il n'est pas nécessaire de télécharger le dump complet de *Wikipedia* avant d'avoir fini le TP). Bien sûr ces corpus devront être décompressés avant utilisation.
- Ces corpus représentent différents types de langage :
  - par rapport à la langue elle-même dans les corpus du répertoire corpus_multilingue
  - par rapport au niveau de langue, langue orale transcrite dans le corpus orfeo et langue écrite dans les autres corpus
  - par rapport au domaine, corpus journalistiques pour les textes de corpus_topic et texte encyclopédique pour les corpus provenant de *Wikipedia*
  - par rapport au thème dans les différentes dépêches de presse du corpus corpus_topic.
- Question 1 : obtenir le lexique d'un corpus donné. Pour un corpus, le lexique correspond à l'ensemble des mots *différents* qu'il contient. Nous utiliserons pour tous les exercices de cette séance une définition très naïve de la notion de mots : toute séquence de caractères séparée par un espace, une tabulation ou un retour à la ligne. Déterminez le lexique des corpus frwiki-sample.txt et orfeo_dump.txt.

Question 2 : vous allez maintenant déterminer la relation qui lie la taille de ces corpus et la taille de leur lexique. Pour cela vous allez écrire un programme qui va afficher la taille du lexique lorsque l'on garde uniquement les *n* premiers mots du corpus, pour les valeurs de *n* suivantes : *100 500 1000 1500 2000 4000 5000 10000 20000 30000 40000 50000 80000 100000 200000 500000 1000000 2000000 5000000*. Par exemple vous allez obtenir:

```
100 63
500 201
1000 323
1500 406
2000 488
4000 737
5000 863
10000 1426
20000 2279
30000 2965
40000 3641
50000 4261
80000 5647
100000 6316
200000 9248
500000 16157
1000000 26516
2000000 39908
5000000 60182
100000000 60182
```

-

Question 3 : vous allez maintenant comparez les distributions *taille corpus/taille lexique* obtenues sur les 2 corpus frwiki-sample.txt et orfeo_dump.txt en les affichant avec la commande Unix *gnuplot*. Cette commande permet de produire une courbe à partir d'un ou plusieurs fichiers de valeurs. Par exemple, si votre distribution sur le fichier frwiki-sample.txt s'appelle wiki.data et celle sur le fichier orfeo_dump.txt, orfeo.data, vous pourrez produire une courbe avec la commande suivante (après avoir lancé l'interpréteur *gnuplot* :

```
set grid ; set logscale x ; set logscale y ; set xlabel 'taille corpus' ; set ylabel 'taille lexique' ; plot 'wiki.data' w lp t 'ecrit' , 'orfeo.data' w lp t 'oral'
```

Si vous voulez écrire la courbe dans un fichier, il suffit de rajouter :

```
set terminal pdf ; set output 'taille_lexique.pdf' ; replot
```

-
    - Question 4 : que déduisez vous des courbes précédentes, tout d'abord concernant le rapport qui lit taille de corpus et taille de lexique, puis sur les caractéristiques de la langue orale par rapport à la langue écrite ?
- Partie 2 : Loi de Zipf et comparaison de corpus
    - Dans cette partie vous allez produire des courbes afin de vérifier si vos données respectent ou pas la loi de Zipf. Pour cela vous allez produire, pour un lexique donné, un fichier contenant sur chaque ligne les mots classés par fréquence décroissante, la fréquence et le rang du mot (selon l'ordre de ces fréquences). Affichez également le produit *rang x fréquence* pour vérifier s'il s'approche d'une constante comme le prévoi la loi de Zipf. Enfin vous produirez des courbes rang/fréquence en utilisant *gnuplot* en affichant les courbes correspondant à plusieurs corpus pour pouvoir les comparer.
    - Question 1 : comparez les courbes correspondant aux différents thèmes des corpus du répertoire corpus_topic. Qu'en déduisez vous ?
    - Question 2 : comparez les courbes correspondant aux différentes langues des corpus du répertoire corpus_multilingue. Qu'en déduisez vous ?
    - Question 3 : produisez la courbe pour le fichier frwiki-from-polyglot.txt. Le loi de Zipf est-elle respectée ?