Master 1 — Langages IL

Ce projet comporte plusieurs parties ; pour obtenir le maximum des points, il vous faut les traiter toutes. Vous pouvez travailler en groupes de deux étudiants. Il fera l'objet de soutenances (individuelles) en salle informatisée. La soutenance compte pour moitié dans la note finale.

Vous rendrez votre projet sous la forme d'un fichier PDF ou ODT (corps de texte en *Times New Roman* 11 points, niveaux de titre en *Verdana* ou *Gill Sans*, code et commandes en *Courier New* 11 points, interligne standard). Il contiendra vos commandes et scripts, vos commentaires et explications.

La date de remise de vos devoirs dépendra des mesures prises par l'UFR EILA pour l'organisation des sessions d'examens au terme de la grève actuelle. Néanmoins, n'attendez pas le dernier moment pour commencer !

Première partie

Le but de cette partie est d'automatiser la constitution de votre corpus suivant les règles énoncées au 1er semestre.

Vous vous aiderez de deux fichiers `fr.url` et `en.url`, contenant *a minima* chacun une URL par ligne (et toute autre information que vous jugerez utile). Évidemment, le fichier `fr.url` contient les adresses de vos documents en français et `en.url` contient celles en anglais. Par exemple, vos deux fichiers pourraient avoir la forme suivante :

> http://blabla DOE J., FOO D. (2005), foo, bar, baz

On considère dans cette partie que vos sources sont au format HTML ou PDF.

Pour télécharger un fichier PDF ou une page HTML, vous pouvez utiliser les commandes `wget(1)`, `lynx(1)` ou `w3m(1)`.

Nota Bene : wget(1) permet aussi d'aspirer une arborescence d'un site :

```
$ wget –mirror –no-parent http://www.example.com/foo/bar/
```

Vous pouvez vous aider des commandes `gs(1)` et `pdftotext(1)` pour transformer les fichiers PDF en texte et `lynx(1)` ou `w3m(1)` pour les fichiers HTML. N'oubliez pas de vérifier l'encodage de vos fichiers ni de les transformer en UTF-8 le cas échéant.

Rappel : vous devez obtenir dans un répertoire `CORPUS_NOM_PRENOM`, deux sous-répertoires `ANGLAIS` et `FRANCAIS` contenant vos fichiers TXT. Ces fichiers portent le nom de l'auteur et l'année de publication (séparés par le caractère `_`) : `JDOE_2005.txt`, `JDOE_2005_a.txt`, `DFOO_2007.txt`.

Commentez commandes et résultats ; qu'en déduisez-vous ?

Deuxième partie

Rappel : Un **bi-gramme** est une suite de deux mots consécutifs.

Extraire les bi-grammes de vos deux corpus monolingues et calculer leur fréquence d'apparition (fréquences exprimées en pourcentage). Le résultat attendu est un fichier contenant un bi-gramme par ligne précédé de sa fréquence (4 chiffres après la virgule) et séparé d'elle par le caractère `:` (« deux points »).

Rappel : la commande `printf(1)` utilise des « formats » pour afficher les nombres flottants (« à virgule ») avec la précision désirée. Si vous préférez utiliser `awk(1)`, vous disposez aussi d'une fonction `printf()`.

Troisième partie

À l'aide de quelques fichiers de votre corpus, écrire un script qui extrait les N mots les plus fréquents d'un corpus. Le résultat attendu est un fichier contenant un mot par ligne.

Quatrième partie

On se focalise désormais sur le corpus français. Dans cette partie, on utilise un anti-dictionnaire (aussi appelé dictionnaire d'exclusion) `fr_freq.txt` téléchargeable à l'adresse :

`http://wall.eila.univ-paris-diderot.fr/~pc/m1-08-09-s2/fr_freq.txt`

Il contient les 9000 mots les plus fréquents trouvés dans les articles du journal *Le Monde* (version en ligne, articles depuis fin mars 2008).

1.Tokeniser votre corpus.

2.Supprimer les mots présents dans `fr_freq.txt`.

3.Pour chacun des mots restants, extraire les bi-grammes correspondants et les fréquencer. Commentez votre commande.

4.Que peut-on dire ? Qu'observe-t-on ?

Cinquième partie

Dans cette dernière partie, on utilise le logiciel *TreeTagger*, disponible ici :

`http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

Commencez par rédiger un manuel succinct (une page dactylographiée maximum) décrivant l'utilisation du logiciel (il n'est pas question ici d'étudier son fonctionnement mais uniquement son utilisation).

Avec *TreeTagger*[1], marquer votre corpus. Extraire les bi-grammes Nom-Nom. Quel est leur fréquence ? Que peut-on en dire ?

Sixième partie

Soit l'archive XMLs.zip (http://wall.eila.univ-paris-diderot.fr/~pc/m1-08-09-s2/XMLs.zip) contenant des fichiers .xml provenant du même site (www.plos.org).
Explorer la structure de ces fichiers en les ouvrant par exemple avec firefox. Il s'agit de transformer les fichiers .xml en fichiers .txt. Nous allons dans un premier temps nous concentrer sur un seul fichier, et, une fois trouvée la bonne manipulation, nous allons l'appliquer à tous les autres.

- La solution la plus simple consisterait à effacer toutes les balises
- Améliorer cette solution en effaçant d'abord les tableaux (entre balises de type <table-wrap>), les figures (entre balises de type <fig>)
- Trouver un moyen de garder les légendes des figures précédées par la chaîne de caractères : **FIGURE :**
- Proposer d'autres améliorations possibles de cette solution
- Maintenant, en une seule commande, appliquer la même commande à tous les fichers .xml

A partir des fichiers .txt obtenus à l'exercice 2., trouver un moyen (consulter les sed oneliners) pour obtenir un fichier contenant uniquement les sections “Introduction” des textes, puis une autre pour les sections “Results”.
Que pouvez vous dire sur l'avantage de travailler avec des fichiers .xml ?

1La commande est installée en salle 515B, dans l'amphithéâtre 3B (« libre service », partie droite) et au CRL.