SL0720X - Perl et TreeTagger

Franck Sajous (CLLE-ERSS) - sajous@univ-tlse2.fr
http://fsajous.free.fr/

1 Comptage des catégories et des mots inconnus

1. Écrivez en Perl un programme qui lit un fichier étiqueté par TreeTagger et qui :
   - comptabilise le nombre de noms, d'adjectifs et de verbes ;
   - comptabilise le nombre de noms inconnus, d'adjectifs inconnus et de verbes inconnus ;
   - affiche le nombre de mots de chaque catégorie.
   - affiche le nombre et le pourcentage de mots inconnus de chaque catégorie.
2. Testez ce programme sur différents corpus étiquetés (en commençant par un « exemple-jouet » pour tester s'il marche).

Note : on demande ici de comptabiliser les nombres d'occurrences.

2 Extraction de la liste des mots inconnus

1. Écrivez en Perl un programme qui lit un fichier étiqueté par TreeTagger et qui reproduit en sortie la liste des noms et des adjectifs mots inconnus, précédés de leur catégorie syntaxique.
2. Testez ce programme sur différents corpus étiquetés (toujours en commençant par un petit texte).

3 Lexique auxiliaire

1. Étiquetez un petit texte (par exemple, `neologismes.xml`, ou un autre) et appliquez le programme mis au point en section 1 pour compter le nombre de mots inconus.
2. Lancez ensuite le programme créé en section 2 afin de lister les mots inconnus.
3. Créez un lexique auxiliaire à partir des mots inconnus extraits.
4. Étiquetez à nouveau le texte avec ce lexique auxiliaire.
5. Relancez à nouveau le programme créé en section 2 sur ce dernier étiqutage et comparez les nombres de mots inconnus par catégorie avec ceux obtenus précédemment

4 Corpus LexiMédia2007

Ce corpus est constitué d'articles des quotidiens *Le Monde, Libération, Le Figaro* et *L'Humanité* parus durant la campagne présidentielle de 2007 [1].

1. Téléchargez ce corpus depuis IRIS. Le fichier, d'un peu moins d'un million de mots, est un extrait du corpus initial.
2. Étiquetez le corpus.
3. Appliquez le programme d'extraction de mots inconnus (section 2)
4. Sous Notepad, il est possible de trier ces mots en ne conservant qu'une seule instance de toutes les occurrences d'un même mot :
   - Ouvrez le fichier
   - Dans le menu *TextFx/TextFx Tools*, cliquez sur *+Sort outputs only UNIQUE*
   - Puis Dans le menu *TextFx/TextFx Tools*, cliquez sur *Sort lines case sensitive*
5. Observez les différents mots inconnus et proposez une classification (vous pouvez au besoin observer les contextes des mots inconnus de TreeTagger en important le corpus initial sous AntConc)

1. `http://redac.univ-tlse2.fr/LexiMedia2007/`