

Master Informatique (M2) – TP Fouille de Textes

2016-2017

→ Etape 1 : Etiquetage grammatical

Télécharger le fichier : http://textmining.biz/Staff/Roche/ECD_M2/Etiqueteur/brill.zip

En décompressant ce fichier, un répertoire BRILL est créé. Vous aurez ainsi tous les fichiers nécessaires pour l'exécution de l'étiqueteur de Brill (sur des textes en anglais) tout au long de ce TP.

Exercice 1

Appliquer l'étiqueteur de Brill sur la phrase ci-dessous écrite dans le fichier `corpus.txt`.

The policy excludes France , Germany , Russia and Canada
from bidding on construction important projects (\$ 500000) .

Pour appliquer l'étiqueteur de Brill, exécuter la commande : `./tagger LEXICON_vide corpus.txt
BIGRAMS LEXICALRULEFILE_vide CONTEXTUALRULEFILE_vide`

Question 1 : Que remarquez-vous ?

Question 2 : Calculer le Rappel et la Précision des noms (étiquettes NN, NNS, NNP, NNPS).
Discuter le résultat.

Exercice 2

Appliquer à chaque mot du texte son étiquette grammaticale correcte. Par exemple, le mot France étant un nom propre, vous devez lui associer l'étiquette NNP dans le fichier `LEXICON_vide` :
France NNP

Question : Appliquer l'étiqueteur de Brill avec ce nouveau lexique formé et vérifier alors que le texte donné en sortie est correctement étiqueté.

Exercice 3 : Utilisation du lexique seul

Question 1 : Éliminer le mot construction du Lexique. Conclure.

Question 2 : Éliminer les noms des pays du Lexique. Conclure.

Question 3 : Éliminer les mots important et 500000 du Lexique. Conclure.

Exercice 4 : Impact des règles lexicales

Dans le fichier LEXICALRULEFILE_vide propre aux règles lexicales, ajouter la règle : NN ant fhassuf 3 JJ x (*Attention !! Ne pas oublier d'ajouter un Retour Chariot à la fin de la ligne*).

Question 1 : Appliquer l'étiqueteur de Brill en utilisant le lexique modifié à l'exercice précédent. Que remarquez-vous ?

Question 2 : Proposer une interprétation de cette règle.

Exercice 5 : Impact des règles contextuelles

Question 1 : Éliminer le mot The du lexique. Conclure et expliquer le résultat obtenu.

Question 2 : Dans le fichier CONTEXTUALRULEFILE_vide propre aux règles contextuelles, ajouter la règle : NN CD PREVTAG \$. Appliquer l'étiqueteur de Brill. Que remarquez-vous ? Proposez une interprétation de cette règle.

Master Informatique (M2) – TP Fouille de Textes

2016-2017

→ Etape 2 : *Extraction de la Terminologie*

- 1) Constituer un corpus liés à la COP 21 (en français ou en anglais)
- 2) Extraction de la Terminologie avec BioTex (www.tubo.lirmm.fr/biotex)
- 3) Calculer l'AUC à partir des 5 premiers termes donnés par différentes mesures