

TP 2 - Fouille de Données - Master Pro

Extraction de la Terminologie

2006/2007

Exercice 1 - Étiquetage

Appliquer l'étiqueteur de Brill sur un corpus normalisé constitué de quelques dépêches relatives aux élections en Ukraine. Ce corpus écrit en anglais (fichier `corpusNormalise.txt`) ainsi que l'étiqueteur de Brill (fichier `brill.zip`) sont disponibles à l'adresse :

http://www.lirmm.fr/~mroche/Enseignements/FdD_M2P/

Question : Appliquer l'étiqueteur de Brill. Pour cela exécuter la commande :

```
./tagger LEXICON corpusNormalise.txt BIGRAMS LEXICALRULEFILE CONTEXTUALRULEFILE  
> corpusNormaliseEtiquete.txt
```

Le résultat de l'étiquetage sera écrit dans le fichier `corpusNormaliseEtiquete.txt`.

Exercice 2 - Extraction des candidats termes

Le but de cet exercice est d'extraire les candidats termes de type "adjectif nom" à partir du corpus (c'est-à-dire les couples de mots dont le premier mot est un adjectif et le second un nom). Nous rappelons que l'étiquette de Brill JJ correspond aux adjectifs et les étiquettes NN, NNS, NNP NNPS correspondent aux noms.

Question : Écrire un programme (en Perl, en Java, en C ou tout autre langage de votre choix) qui extrait les candidats termes. Chaque candidat terme et son nombre d'occurrences dans le corpus seront écrits dans un fichier. Ainsi, chaque ligne du fichier comportera le candidat terme et son nombre d'occurrences (séparés par un caractère "espace"). Ces candidats termes seront classés de manière décroissante selon leur nombre d'occurrences.

Exemple de fichier de sortie :

```
adj1 nom1 10  
adj2 nom2 8  
.  
.  
.  
adjn nomn 1
```

Donner les trois premiers candidats termes (les candidats termes les plus fréquents).

Exercice 3 - Classement des candidats termes

Cet exercice consiste à classer les candidats termes de type “adjectif nom” en utilisant deux mesures statistiques (l’Information Mutuelle et l’Information Mutuelle au Cube) :

$$IM(x, y) = \frac{nb(x, y)}{nb(x, *)nb(*, y)}$$

$$IM^3(x, y) = \frac{(nb(x, y))^3}{nb(x, *)nb(*, y)}$$

Dans ces mesures, notons $nb(x, y)$ le nombre de candidats termes “ $x y$ ”, $nb(x, *)$ le nombre de candidats termes commençant par “ x ”, $nb(*, y)$ le nombre de candidats termes se terminant par “ y ”.

Question 1 :

Classer les candidats termes en utilisant les deux mesures :

- l’Information Mutuelle (IM)
- l’Information Mutuelle au Cube (IM^3)

Donner les cinq premiers candidats termes avec ces deux mesures.

Question 2 :

Discuter le résultat obtenu en donnant la mesure qui vous semble la mieux adaptée pour extraire des termes utiles pour la construction d’une classification conceptuelle? Illustrez et justifiez vos propos avec des exemples issus du corpus étudié.

Question 3 :

Pour évaluer la qualité des résultats issus de ces étapes d’extraction et de classement de la terminologie, plusieurs critères d’évaluation vus dans les différents cours de “fouille de données” peuvent être utilisés. Discuter ces différentes mesures d’évaluation en vous appuyant sur le contexte du TP.

Exercice 4 - Extraction de la terminologie à partir d’un texte libre

Constituez un corpus (ensemble de textes homogènes) en anglais sur un thème de votre choix.

Après avoir appliqué l’étiqueteur de Brill sur votre corpus, effectuez une tâche d’extraction des termes de type “adjectif nom” (en vous appuyant sur les différentes étapes décrites dans les exercices précédents).

Attention!! Un prétraitement du corpus consistant à placer un espace entre les caractères spéciaux tels que les ponctuations peut être nécessaire pour avoir un étiquetage de bonne qualité. Par exemple, le fragment “xx, yy” devient “xx , yy”

Question : Quels types de termes sont privilégiés avec l’Information Mutuelle? Illustrez et justifiez vos propos avec des exemples issus du corpus que vous avez acquis.

MODALITÉS : Dans une archive (zip ou autre) placez le rapport propre à ce TP, le ou les programmes, le corpus relatif au dernier exercice. Cette archive (pour chaque binôme) devra avoir le format suivant : Nom1_Nom2_UMINP406.zip. L’archive doit être envoyée à l’adresse suivante : mroche@lirmm.fr