TP2 - Etiquetage morpho-syntaxique

Cours d'initiation au TAL - L2 MIASHS

10 février 2014

Pour chacun des exercices, sauvegardez le script Python dans un fichier qui a comme nom *exo < numéro de l'exercice >.py* puis mettre tous les fichiers dans une archive .zip qui a comme nom *TP2_<nom de l'étudiant>_<prénom de l'étudiant >.zip*. A la fin de la séance, envoyez l'archive en fichier attaché à l'adresse *perrier@loria.fr* en mettant comme objet du message *Initiation au TAL : TP2*.

A l'aide d'un navigateur, affichez le tutoriel NLTK en allant à l'URL : *nltk.org/book*

Ouvrez le tutoriel au chapitre 5 *Categorizing and Tagging Words*. Lisez attentivement les passages qui vous seront recommandés en exécutant les exemples dans une fenêtre *Idle* ouverte en parallèle.

1 Corpus étiquetés

Dans la section *5.2 Tagged Corpora*, allez directement au paragraphe *A Simplified Part-of-Speech Tagset*. Dans ce paragraphe et ceux qui suivent, sont présentées certaines fonctions qui permettent d'explorer un corpus étiqueté. Allez jusqu'à la fin de la section.

Exercice 1.1 *Stockez dans une variable le corpus de mots de Brown de la catégorie* belles_lettres *étiqueté avec le jeu d'étiquettes simplifié.*

A l'aide de la fonction FreqDist *de NLTK, déterminez les 500 mots les plus fréquents du corpus et commentez le résultat.*

A l'aide de la fonction ConditionalFreqDist *de NLTK, déterminez les étiquettes dans le corpus de chacun des mots suivants rangées de la plus fréquente à la moins fréquente :* best, look, most, that.

Déterminez ensuite les mots du corpus qui ont à la fois des occurrences étiquetées comme adjectifs et d'autres étiquetées comme adverbes.

Déterminez enfin les mots les plus ambigus du corpus, c'est-à-dire ceux qui ont au moins 4 étiquettes.

2 Etiqueteur n-gramme

Allez à la section *5.5 N-Gram Tagging* et étudiez la complètement.

Exercice 2.1 *Stockez dans une variable le corpus de phrases de Brown de la catégorie* belles_lettres *étiqueté avec le jeu d'étiquettes simplifié. Déterminez la précision d'un étiqueteur unigramme entraîné sur 90% du corpus et testé sur les 10% restant. Faites de nouveau l'expérience en partageant moitié moitié le corpus entre corpus d'entraînement et corpus de test.*

En partageant toujours le même corpus en 90% de corpus d'entraînement et 10% de corpus de test, considérez maintenant un étiqueteur bigramme.

Déterminez l'étiquetage qu'il assigne à la phrase numérotée 6489. Comparez avec l'étiquetage initial de la phrase et commentez.

Déterminez ensuite la précision de cet étiqueteur sur l'ensemble du corpus d'entraînement, puis sur l'ensemble du corpus de test. Commentez.

Choisissez maintenant un étiqueteur bigramme entraîné sur le même corpus mais qui fait appel à un étiqueteur unigramme en cas de bigramme inconnu. Cet étiqueteur unigramme fait lui-même appel à un étiqueteur par défaut qui assigne l'étiquette NN *en cas de mot inconnu. Déterminez la précision du nouvel étiqueteur bigramme sur le corpus de test.*

3 Etiqueteur à base de règles de correction

Allez à la section *5.6 Transformation-Based Tagging* qui présente l'étiquetage à base de règles de correction et étudiez la démonstration qui y est présentée.

Exercice 3.1 *Avec le même corpus d'entraînement que pour l'étiqueteur bigramme, créez un étiqueteur de Brill en suivant les indications de la documentation que vous pouvez trouver sur le Web à l'adresse :*

`http://nltk.org/api/nltk.tag.html#module-nltk.tag.brill`

Expliquez les 5 premières règles engendrées par apprentissage sur le corpus.

Déterminez la précision de l'étiqueteur sur le même corpus de test que l'étiqueteur bigramme.