TP3 - Classification textuelle

Cours d'initiation au TAL - L2 MIASHS

28 mars 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 *TP3_<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 : TP3*.

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

Ouvrez le tutoriel au chapitre 6 *Learning to Classify Text*. Lisez attentivement les passages qui vous seront recommandés en exécutant les exemples dans une fenêtre *Idle* ouverte en parallèle.

1 Classification bayésienne naïve

La section 6.1 *Supervised Classification* présente la classification supervisée à travers diverses méthodes. Etudiez les deux premiers paragraphes de cette section *Gender Identification* et *Choosing the Right Features* où est utilisée la classification bayésienne naïve.

Passez ensuite directement au paragraphe *Part-of-Speech Tagging*. Celui-ci présente l'étiquetage morpho-syntaxique comme un problème de classification. Le classifieur qui est utilisé ici est fondé sur les arbres de décision. L'entrainement du classifieur peut être très long donc vous pouvez omettre l'utilisation de ce classifieur.

Exercice 1.1 *Vous allez étiquetez le sous-corpus littéraire du corpus* Brown*, nommé* belles_lettres*, avec le jeu d'étiquettes morpho-syntaxiques du Penn Treebank en utilisant un classifieur bayésien naïf. Ecrivez un programme Python qui exécute les tâches suivantes :*

a) Stocker les 10000 premiers mots du corpus étiqueté en question dans une variable tagged_words*.*

b) En faisant comme dans le tutoriel, choisir comme traits discriminants les 50 suffixes les plus fréquent du corpus avec 3 lettres maximum.

c) En utilisant la fonction pos_features *du tutoriel, transformer le corpus étiqueté en une liste* featuresets *de couples (dictionnaire de traits, étiquette morpho-syntaxique).*

d) Diviser la liste featuresets *en corpus d'entrainement et corpus de test. Le corpus de test sera formé des 10% premiers mots et corpus d'entrainement des 90% derniers mots.*

e) Entrainer un classifieur bayésien sur le corpus d'entraînement et l'évaluer sur le corpus de test.

f) Déterminer à l'aide du classifieur l'étiquette du mot set.

Modifiez les paramètres du classifieur pour améliorer sa fidélité : taille globale du corpus, taille maximum des suffixes, nombre de suffixes, choix des suffixes, utilisation du contexte, répartition entre corpus d'entraînement et corpus de test. Notez en commentaires les conclusions de votre recherche.

2 Arbres de décision

La section 6.4 *Decision Trees* présente la classification supervisée à l'aide des arbres de décision. Etudiez complètement cette section.

Exercice 2.1 *Vous allez étiquetez le sous-corpus littéraire du corpus* Brown, *nommé* belles_lettres, *avec le jeu d'étiquettes morpho-syntaxiques du Penn Treebank en utilisant un classifieur fondé sur un arbre de décision. Ecrivez un programme Python qui exécute les tâches suivantes :*

a) Stocker les 10000 premiers mots du corpus étiqueté en question dans une variable tagged_words.

b) En faisant comme dans le tutoriel, choisir comme traits discriminants les 50 suffixes les plus fréquent du corpus avec 3 lettres maximum.

c) En utilisant la fonction pos_features *du tutoriel, transformer le corpus étiqueté en une liste* featuresets *de couples (dictionnaire de traits, étiquette morpho-syntaxique).*

d) Diviser la liste featuresets *en corpus d'entrainement et corpus de test. Le corpus de test sera formé des 10% premiers mots et corpus d'entrainement des 90% derniers mots.*

e) Extraire la liste total_labels *des étiquettes des mots du corpus d'entrainement en gardant les répétitions.*

f) Extraire la liste e_labels *des étiquettes des mots terminant par "e" du corpus d'entrainement en gardant les répétitions.*

g) Extraire la liste none_labels *des étiquettes des mots ne terminant pas par "e" du corpus d'entrainement en gardant les répétitions.*

h) Déterminer l'entropie de chacune des trois listes qui viennent d'être calculées.

i) Calculer le gain d'information obtenu en choisissant comme trait discriminant la propriété pour un mot de se terminer ou non par "e".

j) Entrainer un classifieur fondé sur un arbre de décision sur le corpus d'entraînement et l'évaluer sur le corpus de test.

k) Déterminer à l'aide du classifieur l'étiquette du mot set.

Modifiez les paramètres du classifieur pour améliorer sa fidélité : taille globale du corpus, taille maximum des suffixes, nombre de suffixes, choix des suffixes, utilisation du contexte, répartition entre corpus d'entraînement et corpus de test. Notez en commentaires les conclusions de votre recherche.