TP d'introduction au Traitement Automatique des Langues

UPMC - 2015/2016

L'objectif de ce TP est de vous faire tester différents types d'analyses en traitement automatique des langues. Vous utiliserez d'abord l'outil nltk.
Note : vous pouvez trouver une documentation de nltk à l'adresse suivante :
`http://www.nltk.org/`.

Exercice 1 - Principe du TP

1 - Configuration
Pour faire ce TP, vous aurez besoin d'installer un certain nombre de modèles dans nltk.

```
python
import nltk
nltk.download()
```

Puis installez :

- punkt
- treebank
- maxent_treebank_pos_tagger
- maxent_ne_chunker
- words (corpus)

Enfin, sauvegardez dans le répertoire de votre choix les fichiers `test*.txt` qui se trouvent à l'adresse `http://www.limsi.fr/Individu/annlor/enseignement/corpus.tar.gz`. Les deux fichiers principaux sont `test_article.txt` qui contient un article du Time [1] et `test_srt.txt` qui contient les sous-titres d'un épisode de série [2].

Pour chaque exercice :

- Comparez les outils sur les deux types de texte (sous-titres et article de journal).
- Analysez les résultats, et essayez de généraliser les erreurs pour trouver les cas qui posent problème.

Exercice 2 - Segmentation en phrases

L'outil de segmentation en phrases de nltk permet de détecter la fin d'une phrase.

2 - Pour tester la segmentation en phrases, vous pouvez utiliser les commandes suivantes :

```
f=open('corpus/test_article.txt')
raw = f.read()
sents = nltk.sent_tokenize(raw)
print(sents)
```

Exercice 3 - Segmentation en mots (*tokens*)

L'outil de segmentation en mots permet de découper une chaîne de caractères en *tokens* c'est-à-dire en mots, ponctuation, nombres...

1. du 22 novembre 2015 sur le congé de paternité de Mark Zuckerberg
2. The Good Wife, saison 7 épisode 6

3 - Pour tester la segmentation en mots, vous pouvez utiliser la commande suivante :

```
tokens = nltk.word_tokenize(raw)
```

Exercice 4 - Reconnaissance d'entités nommées

Les entités nommées correspondent aux noms de personnes, lieux, organisations...

4 - Pour tester la reconnaissance des entités nommées, vous pouvez utiliser la commande suivante :

```
tagged = nltk.pos_tag(tokens)
print(nltk.ne_chunk(tagged))
```

Exercice 5 - Analyse morpho-syntaxique

L'analyseur morpho-syntaxique indique pour chaque token la catégorie de mot associée (nom commun, nom propre, préposition...).

5 - Pour tester l'étiquetage morpho-syntaxique, utilisez la commande suivante :

```
nltk.pos_tag(tokens)
```

La liste des étiquettes avec leur signification est disponible en annexe 1.

Exercice 6 - Analyse syntaxique

6 - Analyse en constituants

L'analyse en constituants construit l'arbre syntaxique de la phrase regroupant les tokens en groupes de mots (groupe nominal, verbal...) hiérarchiques.
Pour afficher une analyse en constituants, utilisez la commande suivante :

```
from nltk.corpus import treebank
t = treebank.parsed_sents('wsj_0001.mrg')[0]
t.draw()
```

Exercice 7 - Comparaison avec un autre outil

La suite d'outils de Stanford CoreNLP offre les mêmes outils. Afin de comparer les performances des outils deux à deux, analyser vos textes avec la suite coreNLP, et comparer les résultats.
Vous pouvez consulter la page `http://nlp.stanford.edu/software/corenlp.shtml` pour plus d'informations et la page `http://nlp.stanford.edu:8080/corenlp/` pour une démonstration en ligne.

coreNLP peut être téléchargé à l'adresse suivante : `http://www.limsi.fr/Individu/annlor/enseignement/corenlp.tar.gz`
Pour l'utiliser :

```
java -cp "stanford-corenlp-full-2014-08-27/*" -Xmx2g
edu.stanford.nlp.pipeline.StanfordCoreNLP -annotators
tokenize,ssplit,pos,lemma,ner,parse,dcoref -outputDirectory votre-repertoire -file
fichier-a-analyser.txt
```

Rem : vous pouvez aussi positionner la variable CLASSPATH
Voir aussi la page `http://webia.lip6.fr/~guigue/wikihomepage/pmwiki.php?n=Course.CourseTALTME6`
Le résultat est dans un fichier .xml, qui peut être visualisé avec la feuille de style que vous aurez recopié chez vous : CoreNLP-to-HTML.xsl

Liste des étiquettes

- adjectifs
  - JJ : adjectif
  - JJR : adjectif comparatif
  - JJS : adjectif superlatif
- adverbes
  - RB : adverbe, négation
  - RBR : adverbe comparatif
  - RBS : adverbe superlatif
  - WRB : adverbe interrogatif utilisé dans un sens temporel (When/WRB he finally arrived, ...)
- noms
  - NN : nom commun singulier ou indénombrable
  - NNS : nom commun pluriel
  - NNP : nom propre singulier
  - NNPS : nom propre pluriel
- conjonctions, prépositions...
  - CC : conjonction de coordination
  - IN : préposition ou conjonction de subordination
- verbes
  - MD : verbe modal
  - VB : verbe à l'impératif, infinitif ou subjonctif
  - VBN : verbe au participe passé
  - VBG : verbe au participe présent
  - VBP : verbe au présent
  - VBZ : verbe ou présent, 3e personne du singulier
  - VBD : verbe au passé
- pronoms
  - PRP : pronom personnel
  - PRP$: pronom possessif
  - WP : pronom interrogatif
  - WP$: pronom interrogatif possessif (whose/WP$)
- DT : déterminant
- WDT : déterminant interrogatif
- CD : nombre cardinal
- UH : exclamation, interjection
- EX : existentiel (There/EX is a party)
- FW : mot étranger
- RP : particule
- TO : *to*
- LS : marqueur de liste
- POS : marque du possessif
- PDT : prédéterminant (both/PDT the girls)
- SYM : symbole