Université Mohammed Premier
Faculté des sciences
Département d'Informatique
Oujda

Année universitaire 2015/2016
Master M2I
Moteurs de Recherche
Semestre 2

TP numéro 2
Traitement Automatique des Langues Naturelles (TALN)

1 Transformation de textes

Écrire un programme java qui permet de lire un fichier texte (codé en utf-8), par exemple « **test.txt** » et permet de produire un autre fichier texte (codé en utf-8) avec ajout de **_min** au nom du fichier (par exemple « **test_min.txt** »). Le nouveau fichier aura les caractéristiques suivantes :

- tous le texte sera en minuscule ;
- suppression des caractères non-latin ;
- garder les nombres ;
- supprimer les chaînes qui commencent par un chiffre.

2 Stanford POS

Part-Of-Speech Tagger (Étiqueteur grammatical de Stanford).

- Téléchargez la version complète : **stanford-postagger-full-2015-12-09.zip** du site : http://nlp.stanford.edu/software/tagger.shtml
- Vous avez besoin de jdk-8.
- Lisez le fichier « README.txt » qui se trouve dans le dossier « stanford-postagger-full-2015-12-09 ».
- Pour la signification des différentes étiquettes, veuillez consulter les liens : https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html et http://www.comp.leeds.ac.uk/amalgam/tagsets/upenn.html

3 Alkhalil Morpho Sys

AlKhalil Morpho Sys est un logiciel open source développé avec Java. Il consiste à faire une analyse morphologique permettant pour chaque mot du texte arabe, pris hors contexte, d'identifier ses différentes étiquettes morphosyntaxiques possibles.

Lien de téléchargement : http://oujda-nlp-team.net/?p=93