Université Mohammed Premier
Faculté des sciences
Département d'Informatique
Oujda

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

TP numéro 5

Exercice 1

Soit la collection http://lakhouaja.oujda-nlp-team.net/wp-content/uploads/2016/03/Donnees.zip, contenant un ensemble de textes arabes sous format utf-8. Les fichiers ont été importés de façon automatique de « المكتبة الشاملة » (http://www.shamela.ws).

Écrivez une classe (en Java) qui permet de fournir les informations concernant chaque livre :

- Nom : الكتاب
- Auteur : المؤلف
- Date d'édition : تاريخ النشر
- Nombre de pages : عدد الصفحات
- Origine du livre : مصدر الكتاب
- Site web d'origine du livre.

Les informations précédentes ne sont pas toutes fournies.

Exercice 2 (normalisation)

Veuillez lire le document (http://lakhouaja.oujda-nlp-team.net/wp-content/uploads/2016/03/TextOperationsZivianiBis.pdf) et écrivez une classe qui permet d'appliquer la partie normalisation aux textes données.

Exercice 3 (tokenisation)

Écrivez une classe qui découpe un texte (normalisé) en tokens.

Exercice 5 (indexation)

1. Écrivez une classe qui construit l'index de la collection normalisé et le sauvegarde dans un fichier. Pour chaque token, l'index comportera les documents dans lequel il appartient ainsi que la ou les position(s) dans le document.
2. Construisez un autre index qui ne tiendra pas compte des mots vides fournis dans le fichier « stopwords_ar.txt ».
3. En utilisant Alkhalil, construisez un index des stems et un autre des lemmes.

Exercice 4 (recherche)

En utilisant les index créés précédemment, écrivez un programme qui contient trois méthodes. Une méthode :

1. qui retourne les documents qui contiennent tous les mots.
2. qui retourne les documents qui contiennent au moins un mot de la requête.
3. qui ne tient pas compte des mots vides dans la recherche.
4. qui fait la recherche par stem.

Exercice 5 (indexation avec pondération)

Écrire une classe qui ajoute à l'index précédent une pondération pour chaque mot de la façon suivante :

— si le mot se trouve dans le titre ou l'auteur lui affecter 7 points
— si le mot se trouve dans le corps du document lui affecter 1 point.

Si le mot apparaît plusieurs fois, on additionne les occurrences.

Par exemple, si dans un document, le mot «الشعر» apparaît :

— une fois dans le titre : 7 points
— sept fois dans le texte : 7 points
— **Total :** 7 + 7 = 14 points.

Le mot «الشعر » sera enregistré dans l'index comme suit :

— « الشعر » dans le document X 14 points
— « الشعر » dans le document Y 17 points

...

Lors de la recherche, si vous saisissez le mot « الشعر », alors le document Y sera classé avant X.

Si vous cherchez plus de 2 mots, par exemple :

« الشعر العربي الجاهلي »

Supposons que le résultat séparé de la recherche donne le résultat suivant :

— « الشعر » : document X 14 points et document Y 17 points
— « العربي » : document X 10 points, document Y 7 points et document Z 12 points
— « الجاهلي » : document Y 10 points et document Z 19 points

L'addition des points pour chaque document donne le résultat suivant :

— document X : 14 + 10 = 24
— document Y : 17 + 7 + 10 = 34
— document Z : 12 + 19 = 31

Le classement sera comme suit : document Y, document Z puis document X.