Module GMD
Telecom Nancy 2A, SIE
Adrien Coulet et Mario Lezoche

TP encadré n°4 :

"Full text indexing with the Lucene API"

"Indexation de documents textuels avec l'API Lucene "

Pour réaliser ce TP vous aurez besoin de la librairie Java suivante :

*lucene-core-4.0.0.jar

*lucene-analyzers-common-4.0.0.jar

(http://lucene.apache.org/core/

A vous de la télécharger à partir du site d'apache et d'inclure leurs locations dans votre CLASSPATH. La Javadoc des classes proposées par la librairie Lucene est en ligne à l'adresse suivante :
http://lucene.apache.org/core/4_0_0/core/index.html

N.B. : l'exemple donné en cours utilisait la version 3.0.3 de Lucene. Certaines classes et méthodes ont depuis été dépréciées. Ceci est expliqué dans la Javadoc.
Des exemples de code avec la nouvelle version de l'API sont disponibles :

Ici pour créer un index :
http://lucene.apache.org/core/4_0_0/demo/src-html/org/apache/lucene/demo/IndexFiles.html

Ici pour faire une recherche dans un index :
http://lucene.apache.org/core/4_0_0/demo/src-html/org/apache/lucene/demo/SearchFiles.html

La base de données DrugBank est une source de données électronique très riche sur les médicaments. Elle contient des données sur 6707 médicaments (le 16/01/2012).
Le fichier /home/depot/2A/gmd/tp4/drug_bank/drugbank.txt.zip est une archive qui contient ces données en format texte.

L'objectif de ce TP est de

(A) créer un index Lucene sur cette ressource et

(B) de permettre de retrouver des noms de médicaments et leurs identifiants (de la forme `DB00001`) dans DrugBank à partir de noms de médicaments ou de noms de symptômes ou maladies pour lequel le médicament est indiqué.

Pour cela dans votre index :

* vous stockerez l'identifiant des médicaments;

* vous stockerez et indexerez le champ `Generic_Name` (le nom générique du médicament);

*vous indexerez les champs `Synonyms, Brand_Names, Description, Indication, Pharmacology, Drug_Interactions`

Questions :

A. Création d'un index

Créez l'index décrit ci-dessus.

1- Quelle taille mémoire occupe le fichier .txt qui contient DrugBank ?

2- Quelle taille occupe l'index que vous avez généré ?

3- Mesurez le temps nécessaire à la constitution de l'index (vous pouvez utiliser la class Date).

TIP 1 : Un enregistrement de médicament dans le fichier txt de DrugBank commence par la chaîne "#BEGIN_DRUGCARD" et se termine par la chaîne "#END_DRUGCARD".
TIP 2 : Créez un fichier avec les 2 premiers médicaments de DrugBank (DB00001 et DB00002) pour tester votre code.

B. Interrogation de l'index

Developpez un programme qui vous permette d'interroger votre index et répondez aux questions suivantes. Pour chaque question, mesurez et notez le temps nécessaire pour obtenir une réponse à votre requête.

4- Quels sont les médicaments dont soit le nom générique, soit un synonyme, soit un nom de commercialisation (Generic_Name, Synonyms et Brand_Names) contient le nom *aspirin* (aspirine en anglais)?
Pouvez vous interpréter ce résultat en utilisant la GUI de DrugBank (http://www.drugbank.ca/) ?

5- Quels sont les médicaments qui peuvent être utilisés pour traiter le diabète ? Pour cela vous chercherez les médicaments pou lesquels le mot *diabetes* (diabète en anglais) apparaît à la fois dans l'indication et la description.

6– Quels sont les médicaments qui interagissent avec le médicament appelé *mercaptopurine* ? Est ce que vous pouvez interprétez ces résultats et les comparer aux résultats visibles sur le GUI de DrugBank ?