Université Mohammed Premier
Faculté des sciences
Département d'Informatique
Oujda

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

**Galago search**


# 1 Téléchargement

- Téléchargez **Galago Search** de : http://lakhouaja.oujda-nlp-team.net/teaching/master-level/search-engines/ et décompressez le.
- Téléchargez le corpus de test de (**CACM**) : http://lakhouaja.oujda-nlp-team.net/teaching/master-level/search-engines/ et mettez le dans dossier **galagosearch-1.04**.

Lorsque vous décompressez le fichier **galagosearch-1.04-bin.zip**, dans le dossier **galagosearch-1.04**, se trouve le dossier **bin** qui contient les scripts **galago** (Linux) et **galago.bat** (Windows).

Sous Windows, pour voir les différentes options de **galago**, tapez la commande :

```
bin\galago.bat help
```

# 2 Construction d'un index

Pour indexer le corpus **cacm.corpus**, il faut taper la commande :

```
bin\galago.bat build cacm.index cacm.corpus
```

L'index créé utilise deux listes :

1. une liste utilise les stems (utilise le stemmer de Porter http://tartarus.org/martin/PorterStemmer) ;
2. l'autre liste n'utilise pas les stems.

Vous pouvez choisir la première lise ou la deuxième dans la phase de recherche, lors de l'exécution des requêtes. Les mots vides (stopwords) ne sont pas éliminés.

# 3 Recherche

Pour faire la recherche, tapez la commande :

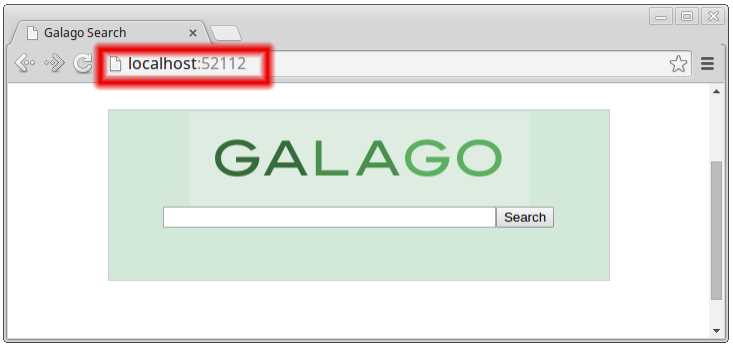
```
bin\galago.bat search cacm.index cacm.corpus
```

Comme dans le cas de l'indexation, vous allez obtenir une sortie comme suit :

```
2015-05-09 06:24:26.721::INFO:  Logging to STDERR via
org.mortbay.log.StdErrLog
2015-05-09 06:24:26.725::INFO:  jetty-6.1.5
2015-05-09 06:24:26.744::INFO:  Started
SocketConnector@0.0.0.0:35717
Server:  http://localhost:52112
```

Le numéro de port change à chaque exécution.

Utilisez l'adresse `http://localhost:52112` dans votre navigateur. Vous allez obtenir une interface comme celle d'un moteur de recherche. Elle se présente comme suit :




Tapez par exemple "**programming**" dans la zone de recherche et cliquez sur le bouton "**search**" pour faire la recherche. En cliquant sur "**debug**", vous allez obtenir comment Galago a traité votre requête.

# 4 Examen des fichiers

La plupart des fichiers **galago** sont écrits par IndexWriter, y compris les fichiers corpus. IndexWriter construit une liste triée de paires clé-valeur. Dans le cas de fichiers corpus, les clés sont les identifiants (noms) des documents. Dans les fichiers inverses, les clés sont les termes.

Dans les deux cas, l'option **dump-keys** affiche les clés du fichier index.

## 4.1 Corpus

Pour voir la liste des documents du corpus, tapez la commande :

```
bin\galago.bat dump-keys cacm.corpus
```

Pour voir le contenu d'un document, utilisez l'option **doc**. Par exemple :

```
bin\galago.bat doc cacm.corpus CACM-2481
```

### 4.2 Fichiers inverses

Pour voir tous les termes du corpus **cacm.corpus** :

```
bin\galago.bat dump-keys cacm.index/parts/postings
```

Pour voir tous les stems :

```
bin\galago.bat dump-keys cacm.index/parts/stemmedPostings
```

Pour voir tous les donnés du fichier inverse, utilisez l'option **dump-index** :

- pour les termes
  ```
  bin\galago.bat dump-index cacm.index/parts/Postings
  ```
- pour les stems
  ```
  bin\galago.bat dump-index cacm.index/parts/stemmedPostings
  ```

## 5 Recherche par lot

Pour faire une recherche par lot, écrivez un fichier (par exemple requete) et mettez dedans les lignes suivantes :

```
<parameters>
  <query>
    <number>requette 0</number>
    <text>programming</text>
  </query>

  <query>
    <number>requette 1</number>
    <text>algorithmic</text>
  </query>
</parameters>
```

Tapez ensuite la commande :

```
bin\galago.bat batch-search --index=cacm.index --count=20 requete
```

L'option **count** pour indiquer le nombre maximal de résultats par requête.