

EZLucene

(dernière mise a jour: 2008.07.04)

I. Qu'est-ce ?

II. Références

III. Contact

IV. Modules Perl requis

V. Légende

VI. Installation

VII. Tutoriel 1: comment déployer EZLucene pour des fichiers XML

VIII. Tutoriel 2: comment déployer EZLucene pour des fichiers TXT

I. Qu'est-ce ?

EZLucene est un environnement permettant de rapidement indexer, interroger des documents XML/TXT en utilisant l'implémentation C++ de Lucene.

EZLucene permet de déployer facilement une formulaire web permettant la creation de requetes et l'interrogation des index sans connaître la syntaxe Lucene.

II. Références

HeliaGene (<http://www.heliagene.org>)

Legoo (<http://www.legoo.org>)

Bacteria@LIPM<http://iant.toulouse.inra.fr/bacteria>Bacteria@LIPM ()

III. Contact

Sébastien Carrere (LIPM)

Sebastien.Carrere@toulouse.inra.fr

IV. Modules Perl requis

[Lucene](#)

[HTML::Entities](#)

[CGI](#)

[File::Temp](#)

[MIME::Base64](#)

[XML::Twig](#)

V. Légende

Contenu des fichiers

Commandes unix

VI. Installation

```
tar xzf EZLucene.tgz
cd EZLucene
./install_EZLucene.pl
```

VII. Tutoriel 1: comment déployer EZLucene pour des fichiers XML

Preparatifs (i): Creer un dossier contenant les fichiers XML a indexer:

```
mkdir data/raw/test
cat > data/raw/test/fichier1.xml
<elt1 att1="toto" att2="tata">
  <elt2 att1="pi">une phrase</elt2>
</elt1>

cat > data/raw/test/fichier2.xml
<elt1 att1="toto2" att2="tata2">
  <elt2 att1="pi">une phrase pour toto2</elt2>
</elt1>
```

Preparatifs (ii): Creer un fichier de conf avec les requetes xpath des champs a indexer:

```
cat > data/cfg/test.xml.cfg
analyzer=Lucene::Analysis::Standard::StandardAnalyzer
#Format de description d'un champ
#field:nom_du_champ=Indexer./requete/xpath
field:id=Text./elt1/@att1
#docid est le champs considere comme une cle primaire
docid=id

field:attribut2=Text./elt1/@att2
field:element2_content=Text./elt1/elt2
field:element2_attribut1=Text./elt1/elt2/@att1

#ce champs est particulier: il permet de stocker le fichier complet
#dans l'index (il est compresse puis encode en base64 - Lucene n'autorisant que
#l'ajout de donnees de type text
#le mot clef « file » indique au script indexer de charger le fichier complet
field:xmlfile=UnIndexed.file

#champs interroge par default
default_search=element2_content
#sortie par default (ici le fichier XML complet)
default_out=xmlfile
#si vous souhaitez rajouter un element XML racine
#dans le cas ou la requete retourne plusieurs fichiers XML
xmlfile_xmltag=elt0 #definition
des extension des fichiers a ajouter dans l'index
#si l'on donne a l'indexeur un repertoire comme input
suffix=xml
```

Types d'Indexer:

Text: analyse, indexe et stocke

Keyword: indexe et stocke

UnIndexed: stocke uniquement (**la donnee est compressee et encodee en base64; utile pour stocker le fichier brut par exemple**)

UnStored: analyse et indexe

Creation index:

```
./bin/int/iANT.idx.xml.pl --parser data/cfg/test.xml.cfg \
--input data/raw/test \
--store data/store/test
```

Informations sur l'index:

Nombre d'enregistrements

```
./bin/int/iANT.idx.search2.pl --parser data/cfg/test.xml.cfg \  
--store data/store/test\  
--info
```

Lister les valeurs d'un champ (id)

```
./bin/int/iANT.idx.search2.pl --parser data/cfg/test.xml.cfg \  
--store data/store/test\  
--query id --list
```

Interrogation de l'index:

```
./bin/int/iANT.idx.search2.pl --parser data/cfg/test.xml.cfg \  
--store data/store/test \  
--query "element2_content:phrase"
```

```
./bin/int/iANT.idx.search2.pl --parser data/cfg/test.xml.cfg \  
--store data/store/test \  
--query "id:toto" -output_fields='id,element2_content'
```

Suppression de donees de l'index:

```
./bin/int/iANT.idx.remove.pl --parser data/cfg/test.txt.cfg \  
--store data/store/test \  
--query "id:toto"
```

Création Formulaire web:

Prés-requis: le répertoire cgi/ de EZLucene doit etre connu par le serveur apache pour executer les scripts CGI (ExecCGI ou ScriptAlias)

Editer le fichier cfg/EZLucene.cfg

```
cat cfg/EZLucene.cfg  
#exemples de parametres specifiques a chaque banque  
#Nom de la premiere banque  
#(ce nom doit etre le meme que le suffixe des parametres  
# 'lucene_index_SUFFIX' et 'lucene_cfg_SUFFIX')  
lucene_db_1=test  
#repertoire de l'index test (doit etre dans "lucene_index_storage")  
lucene_index_test=/test  
#fichier de configuration (xpath) pour les donnees sotckees dans "lucene_index_storage"  
lucene_cfg_test=%I/data/cfg/test.xml.cfg
```

Accédez au formulaire

<http://votre.serveur.web/chemin/vers/EZLucene/cgi/EZLucene.cgi>

VIII. Tutoriel 2: comment déployer EZLucene pour des fichiers TXT

Preparatifs (i): Creer un dossier contenant les fichiers TXT a indexer:

```
mkdir data/raw/test
cat > data/raw/testfichier1.txt
#premier_fichier
scarrere      sebastien   carrere@bezeril.gers
pat   patrick      durel@kamoulox.fr
rob   robert       john@nawak.de
cat > data/raw/test/testfichier2.txt
#deuxieme_fichier
dan   daniel       cavo@lagrotte.ar
jfk   john         ileou@fbi.com
```

Preparatifs (ii): Creer un fichier de conf avec les expressions regulieres des champs a indexer:

```
cat > data/cfg/test.txt.cfg

analyzer=Lucene::Analysis::Standard::StandardAnalyzer

#docid correspond a une «clef primaire»
docid=login

##Format de description d'un champ
#field:nom_du_champ=Indexer.regular_expression
#les variables $file,$abs_path, $line testees avec les expressions regulieres
#sont reservees (techniquement un eval est effectue dans le script iANT.idx.txt.pl
#$abs_path correspond a l'adresse absolue du fichier analyse
#$file correspond au contenu du fichier analyse
#$line correspond au contenu d'une ligne du fichier analyse

field:title=Text.$file=~/(\\S+)/

#si on veut indexer des parties du nom du fichier
field:directory=Text.$abs_path=~/.+\\/(\\S+)\\/(\\w+)/
field:filename=Text.$abs_path=~/.+\\/(\\S+)\\/(\\w+)/

#si on veut garder la ligne complete (rappel: UnIndexed==>Gzip + EncodeBase64)
field:line=UnIndexed.$line
#si on veut garder le chemin absolu du fichier analyse
field:abs_path=UnIndexed.$abs_path
#si on veut garder le fichier complet (mot clef file) (compresse)
field:rawfile=UnIndexed.file

#si on veut indexer tous les mots du fichier
field:tout_les_mots=Text.$file=~/(\\w+)/g

#si on veut créer une entree dans l'index par ligne
field:login=Text.$line=~/^((\\S+)\\s+\\S+\\s+\\S+)/
field:firstname=Text.$line=~/^\\S+\\s+(\\S+)\\s+\\S+/
field:email=Text.$line=~/^\\S+\\s+\\S+\\s+(\\S+)/

default_search=login
default_out=rawfile
=====
#definition du suffixe des fichiers a ajouter dans l'index
suffix=txt
```

Types d'Indexer:

Text: analyse, indexe et stocke

Keyword: indexe et stocke

UnIndexed: stocke uniquement (la donnee est compressee et encodee en base64; utile pour stocker le fichier brut par exemple)

UnStored: analyse et indexe

Creation index:

```
./bin/int/iANT.idx.xml.pl --parser data/cfg/test.txt.cfg \  
--input data/raw/test \  
--store data/store/test
```

Informations sur l'index:

Nombre d'enregistrements

```
./bin/int/iANT.idx.search2.pl --parser data/cfg/test.txt.cfg \  
--store data/store/test\  
--info
```

Lister les valeurs d'un champ (login)

```
./bin/int/iANT.idx.search2.pl --parser data/cfg/test.txt.cfg \  
--store data/store/test\  
--query login --list
```

Interrogation de l'index:

```
./bin/int/iANT.idx.search2.pl --parser data/cfg/test.txt.cfg \  
--store data/store/test \  
--query "tout_les_mots:carrere*"
```

```
./bin/int/iANT.idx.search2.pl --parser data/cfg/test.txt.cfg \  
--store data/store/test \  
--query "login:jfk" -output_fields='line'
```

Suppression de donees de l'index:

```
./bin/int/iANT.idx.remove.pl --parser data/cfg/test.txt.cfg \  
--store data/store/test \  
--query "login:scarrere"
```

Création Formulaire web:

Prés-requis: le répertoire cgi/ de EZLucene doit etre connu par le serveur apache pour executer les scripts CGI (ExecCGI ou ScriptAlias)

Editez le fichier cfg/EZLucene.cfg

```
cat cfg/EZLucene.cfg  
#exemples de parametres specifiques a chaque banque  
#Nom de la premiere banque  
#(ce nom doit etre le meme que le suffixe des parametres  
# 'lucene_index_SUFFIX' et 'lucene_cfg_SUFFIX')  
lucene_db_1=test  
#repertoire de l'index test (doit etre dans "lucene_index_storage")  
lucene_index_test=/test  
#fichier de configuration (regex) pour les donnees sotckees dans "lucene_index_storage"  
lucene_cfg_test=%I/data/cfg/test.txt.cfg
```

Accédez au formulaire

<http://votre.serveur.web/chemin/vers/EZLucene/cgi/EZLucene.cgi>