Topic Classification Engine

Installation guide based on Olivier Grisel's documentation.

http://www.iks-project.eu/sites/default/files/Topic-Classification.pdf

Created by:

Carlo Baraldi - carlo.baraldi@studio.unibo.it

Umberto Marini - umberto.marini2@studio.unibo.it

Indice

Topic Classification Engine Build and Deployment

It is possible to download the sources of the 'topic classification Engine' from the Apache Stanbol[1] svn repository. To build and deploy the Topic Engine (0.10.0 version) we have followed the documentation made available by Olivier Grisel[2], presented at IKS Salzburg Workshop (June 2012).

Below there is a description of the procedure we have followed in order to build and deploy the Topic Engine. Note: you need to install version 0.10.0 of Stanbol.

Step 1 - Build

Start the Stanbol java executable:

```
$ cd ~/stanbol/launchers/stable/target/
$ java -jar org.apache.stanbol.launchers.stable-0.10.0-incubating-
SNAPSHOT.jar -p 9090
```

Then build the Topic Engine:

```
$ cd ~/stanbol/Enhancer/engines/topic
$ mvn install -DskipTests -PinstallBundle
-Dsling.url=http://localhost:9090/system/console
```

...and the tool that enables the web interface (accessible from the Stanbol home page):

```
$ cd ~/stanbol/Enhancer/topic-web
$ mvn install -DskipTests -PinstallBundle
-Dsling.url=http://localhost:9090/system/console
```

Step 2 - Felix OSGi console configuration.

1) Open your browser and go to http://localhost:9090/system/console/configMgr

2) Create a new configuration for the Topic Classification Engine in the 'Configuration' tab. You have to specify:

- ***Name***: the Engine name (ex. "dbpedia-category-model");
- ***org.apache.stanbol.Enhancer.engine.topic.trainingSetId.name***: the name of the training set that you will specify to bind the documents to the related concepts. (ex. "dbpedia-category-trainingset")

3) Create a new configuration for the Topic Solr Training Set. You have to specify:

- ***Name***: the training set name (specify the same name you used for "*org.apache.stanbol.Enhancer.engine.topic.trainingSetId.name*" - see previous step");

[1] Apache Stanbol svn repository: http://svn.apache.org/repos/asf/incubator/stanbol/trunk/

[2] Topic Engine docs: http://www.iks-project.eu/resources/topic-classification-apache-stanbol-engine

- ***Solr Core***: the name of the Solr search platform for topic classification engine (document concepts's search engine). Insert "default" value if you have not created a new Solr server after installing Stanbol.

Step 3 - Concepts registration

In order to register the set of concepts perform a POST to the address of the newly created model.

- Single registration:

```
$ curl -X POST http://localhost:9090/topic/model/dbpedia-category-
model/concept?id=concept_1

$ curl -X POST http://localhost:9090/topic/model/dbpedia-category-
model/concept?id=concept_2

$ curl -X POST http://localhost:9090/topic/model/dbpedia-category-
model/concept?id=concept_3&broader=concept_1&broader=concept_2
```

- Registration from RDF/XML file:

```
$ curl -X POST --data @file_skos.rdf.xml
http://localhost:9090/topic/model/dbpedia-category-model
```

Step 4 - Training based on sample documents

This step determines the behavior of the Topic Engine within the calls chain of the Enhancer. It defines the relationships between entities in the document and the concepts (topics) associated with the document, and indexes these relations by using the Entity Hub.

First you have to POST the texts with the associated categories (in our project we have used the sample dataset provided by Olivier Grisel based on Wikipedia SKOS categories: 2563 SKOS Concepts[1] and 183.000 documents with associated category[2]).

The command to execute document registration is:

```
$ curl -X POST --data @file_1.txt
http://localhost:9090/topic/model/dbpedia-category-model/trainingset?
example_id=example_1&concept=concept_3&concept=concept_42
```

To automate this process, Olivier Grisel has provided a python script called 'dbpediacategories.py' (located in the folder 'stanbol/Enhancer/topic-web/tools') executable with the command:

```
$ python dbpediacategories.py /path/to/dbpediakit/dbpedia-
taxonomy.tsv /path/to/dbpediakit/dbpedia-examples.tsv.bz2
http://localhost:9090/topic/model/dbpedia-category-model/trainingset
```

In order to finish the training step, it is necessary to issue the command:

[1] https://dl.dropbox.com/u/5743203/IKS/dbpedia/dbpediakit-output/dbpedia-taxonomy.tsv

[2] https://dl.dropbox.com/u/5743203/IKS/dbpedia/dbpediakit-output/dbpedia-examples.tsv.bz2

```
$ curl -X POST http://localhost:9090/topic/model/dbpedia-category-model/trainer?incremental=true
```

And it should respond:

"Successfully updated the statistical model(s) of 200 concept(s)."

The Topic Engine should be correctly configured. Each call to the Stanbol Enhancer will correspond to a response including the topics identified by the Topic Engine, according to the performed training phase and according to the concepts stored in the Entity Hub.

Finally, we provide a summary diagram showing how the Topic Engine works (by Olivier Grisel's documentation).