GeoNames Extractor

1. Component Description

The GeoNames Extractor is a stand-alone component responsible for discovering GeoNames entities within text segments.

The component is distributed as an executable JAR file, thus it can be used in any platform.

GeoNames Extractor operates over a locally stored text file that contains the text to be analyzed for the presence of terms corresponding to entities defined within GeoNames.

Upon completion of its execution, it produces an XML document following the schema defined by a simple XSD schema, included with the distribution package of the component. Furthermore, it returns the RDF description as defined in GeoNames for each of the discovered entities.

2. Component Installation

2.1. Using the JAR Distribution

As an executable JAR file, the GeoNames Extractor requires solely the presence of a compatible Java Runtime Environment distribution in the host system. AgroVoc Extractor is compatible with JRE 7 [1: http://www.oracle.com/technetwork/java/javase/downloads/jre7-downloads-1880261.html].

2.2. Using the Source code

Download the source code of the component from the OpenMinted GitHub repository [2: https://github.com/openminted/uc-tdm-agriculture/tree/master/GeoNames%20Extractor] and run from the command line:

```
ant -buildfile build.xml
```

2.3. Testing the installation

GeoNames Extractor can be called without any arguments, using a default exemplary text input included in the distribution package.

3. Relevant Data Processing Scenarios

Extract GeoNames entities found in a document ~~~~~~~~~~~~~~ The user aims to retrieve the GeoNames entities found in a text document, named <filename>, found in the local <path> directory. To this end, she executes the GeoNames Extractor component from the command line:

```
java -jar AK_GeoNames.jar <path> <filename> <username> <maxrows>
```

The <username> parameter corresponds to the GeoNames account of the user. A registration for a GeoNames account is carried out via the GeoNames site [3: http://www.geonames.org/login].

The <maxrows> parameter defines the amount of lines that will be included in the response RDF delivered by the GeoNames Search web service used by the component [4: http://api.geonames.org/search].

Upon completion, the component produces the <filename>.GeoNames.xml file, which contains information on the presence and frequency of any GeoNames entity discovered, namely:

- The lexicalization of the entity that was found in the input document;
- The Lucene score (a variation of the tf-idf metric) of the term in the input document;
- The GeoNames ID for the entity.

Exemplary XML record in the output document

```
<term>
<text>Greece</text>
<geoname_id>390903</geoname_id>
<score>0.2628660500049591</score>
</term>
```

Furthermore, it produces a <filename>.<entity_name>.rdf file for each GeoNames entity discovered in the input text, containing the <maxrows> first lines of its RDF description in the GeoNames knowledge base.