Geopolitical Extractor

1. Component Description

The Geopolitical Extractor is a stand-alone component responsible for discovering terms defined in the FAO Geopolitical ontology [1: http://www.fao.org/countryprofiles/geoinfo/en/].

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

Geopolitical Extractor operates over a locally stored text file that contains the text to be analyzed for the presence of FAO Geopolitical terms.

Upon completion of its execution, it produces an XML document following the schema defined by a simple XSD file, included with the distribution package of the component. The XML presents the terms found in the document ranked according to a modification of the classic tf-idf metric.

2. Component Installation

2.1. Using the JAR Distribution

As an executable JAR file, the GeoPolitical Extractor requires solely the presence of a compatible Java Runtime Environment distribution in the host system. GeoPolitical Extractor is compatible with JRE 7 [2: 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 [3: https://github.com/openminted/uc-tdm-agriculture/tree/master/Geopolitical%20Extractor] and run from the command line:

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

3. Relevant Data Processing Scenarios

3.1. Extract Geopolitical terms found in a document

The user aims to retrieve the references Geopolitical terms found in a text document, named <filename>, found in the local <path> directory. To this end, she executes the GeoPolitical Extractor component from the command line:

```
java -jar AK_Geopolitical.jar <path> <filename> <ontology_path> <ontology_filename>
```

The components requires the existence of a local copy of the FAO Geopolitical ontology, named <ontology_filename> and located in the <ontology_path> directory of the host system.

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

- The lexicalization of the term (in English) that was found in the input document;
- The Lucene score (a variation of the tf-idf metric) of the term in the input document;

Exemplary XML record in the output document

```
<term>
  <text>Agiorgitiko</text>
  <oiv_id>50</oiv_id>
  <score>0.39380478858947754</score>
</term>
```