Grape Varieties Extractor

1. Component Description

The Grape Varieties Extractor is a stand-alone component responsible for discovering grape variety names, as defined by the OIV specification [1: http://www.oiv.int/en/technical-standards-and-documents/description-of-grape-varieties/international-list-of-vine-varieties-and-their-synonyms]. The OIV specification has been embedded programmatically into the component.

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

Grape Varieties Extractor operates over a locally stored text file that contains the text to be analyzed for the presence of grape variety names.

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 grape variety names found in the document ranked according to a modification of the classic tf-idf metric, along with their OIV ID.

2. Component Installation

2.1. Using the JAR Distribution

As an executable JAR file, the AgroVoc Extractor requires solely the presence of a compatible Java Runtime Environment distribution in the host system. AgroVoc 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/Grape%20Varieties%20Extractor] and run from the command line:

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

2.3. Testing the installation

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

3. Relevant Data Processing Scenarios

Extract Grape Variety names found in a document ~~~~~~~~~~~~~~ The user aims to retrieve the references to grape varieties found in a text document, named <filename>, found in the local <path> directory. To this end, she executes the AgroVoc Extractor component from the command line:

```
java -jar AK_Agrovoc.jar <path> <filename>
```

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

- The grape variety 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 OIV ID for the term.

Exemplary XML record in the output document

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