AgroVoc Extractor

1. Component Description

The AgroVoc Extractor is a stand-alone component responsible for discovering AgroVoc terms within text segments.

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

AgroVoc Extractor operates over a locally stored text file that contains the text to be analyzed for the presence of AgroVoc 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 AgroVoc 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 AgroVoc 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/AgroVoc%20Extractor] and run from the command line:

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

2.3. Testing the installation

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

3. Relevant Data Processing Scenarios

3.1. Extract AgroVoc terms found in a document

The user aims to retrieve the AgroVoc terms 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> <lang> (0|1) [ontology_path]
[ontology_filename]
```

The <lang> parameter defines the language of the input text document, expressed as an ISO 639-1 Alpha-2 code [3: http://www.iso.org/iso/language_codes].

The value of the last parameter specifies if the user wants to use the FAO service (http://web.archive.org/web/20160407175423/https://bitbucket.org/aims-fao/agrovoc-web-services/) for searching a given term (1), or if she wants to use a local copy of AgroVoc (0). The local AgroVoc copy should be located in the [ontology_path] directory and should be named [ontology_filename].

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

- The term 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 AgroVoc code for the term.

Exemplary XML record in the output document

```
<term>
  <text>agriculture</text>
  <code>203</code>
  <score>0.2628660500049591</score>
</term>
```