SC!FY
SCIENCE FOR YOU

openMINTED
Open Mining Infrastructure for Text & Data

OpenMinted Phase II
Call for TDM Software and Knowledge Resources

Deliverable T.2: Use case and example of usage
VineSum: A software component for vine/grape variety named entity extraction and clustering

Submitted via email: openminted.tenders@gmail.com

Table of Contents

Intro

The purpose of this document is to describe a use case and example of usage scenario of VineSum prototype. VineSum is an open source, OMTD integrated, software component that, given a collection of documents, it:

- Performs NER extraction, identifying four entity types: a) vine varieties b) persons c) locations d) dates.
- Clusters the documents by taking into account the extracted entities.

VineSum extends existing SciFY NLP components (NER and Clustering) in order to make them accessible via OMDT as a software component and achieve maximum interoperability with OMDT.
It utilizes the Grape varieties [1] agriculture and agronomy resource in order to be used in the NER Component. It is released with an Apache 2 licence, through Maven Central.

Related links:
Source code: http://bitbucket.scify.org/projects/OP/repos/vinesum/browse
Licence: http://bitbucket.scify.org/projects/OP/repos/vinesum/browse/LICENCE.md
Maven Central:
https://search.maven.org/#search%7Cga%7C1%7Ca%3A%22omtdvinesum%22

Use case

Rationale

The need for content categorization and summarization is large. Even though online services exist (eg. AGRIS FAO online service), these services lack the text and data mining modules that can help explore, identify patterns, visualize and understand large collections of resources. VineSum, being a generic NLP tool, can be used as a software component, in a wide range of applications: information extraction from online datasets, monitoring outbreak alerts, consumption occasions (e.g. places of consumption), analysis of trends in wine consumption etc.

Use case scenario

For the purpose of this deliverable the following use case scenario will be demonstrated: A set of documents (dataset) is extracted from AGRIS FAO online service [2]. This dataset is given as input to VineSum. The output of VineSum enriches the existing collection with keyword, entity (vine, persons, location and dates) and cluster information. This output is described in XMI format (UIMA CAS) and can be used in a mixed workflow, within OMTD, that allows the user to categorize/summarize the information helping her to discover any hidden and/or new knowledge.

Example of usage

A user extracts a set of documents related to a vine variety from AGRIS FAO online service [2]. This set of documents is given as input to VineSum in (XMI) format. The output of VineSum enriches the existing collection with keyword, entity (vine, persons, location and dates) and cluster information.

Input

For the purpose of this demo, we have selected 3 journal articles found at AGRIS FAO Online service and extracted their abstract.
File1 : Abstract of http://agris.fao.org/agris-search/search.do?recordID=IT2001060143
File2 : Abstract of http://agris.fao.org/agris-search/search.do?recordID=IT2006601663
File3 : Abstract of http://agris.fao.org/agris-search/search.do?recordID=IT2005600933
These files can be found at the VineSum's code repository "resources/example" directory (http://bitbucket.scify.org/projects/OP/repos/vinesum/browse/omtdvinesum/resources/example) with the following file names: test1.txt, test2.txt and test3.txt. All of these abstracts contain the vine type "Grechetto".

Experiment Execution / Pipeline Description

Following the input files definition, the next step is to execute the main method located in the **org.scify.vinesum.example.VineSumExperiment** class, containing an UIMA analysis engine pipeline using DKPro Core [3].
Inside the pipeline, a TextReader (provided by the DKPro community) is used which is able to transform a text file to UIMA CAS format. The language is set to **en**, the location of input documents is **resources/example** and only **.txt** files are used.
The VineSum component adds the annotation result to the UIMA CAS document representation. Finally, a XmiWriter is writing the final state of CAS to the **resources/example** directory. The produced file is an **.xmi** that has the same name as the file used for input, but with the **.xmi** suffix added to that.
For example, if a file named test.txt is found inside the **resources/example** directory, it will be used as an input to the VineSum component and the annotation result will be stored in the same directory with the file name test.txt.xmi.
More specifically, the experiment described above, will produce three **.xmi** files with the following names: test1.txt.xmi, test2.txt.xmi and test3.txt.xmi.

Output

Consider this abstract as the input of the experiment:

```
The clonal selection of main Umbrian grapevine varieties (Alicante n.,
locally known as Gamay perugino, Grechetto b., codified with the G 109
abbreviation, Sagrantino n. and Trebbiano spoletino b.) started in
1989. This work led to the identification of six biotypes and the
homologation was required for four of them. The differences inside the
group of varieties called Grechetto were confirmed once more and,
therefore, it is useful to distinguish them better and/or to revise
their denomination. For each biotype, the vegetative, production and
enological characteristics, supplemented by the organoleptic
evaluation, are reported
```

the output of the VineSum execution is described below at XMI format (displaying only the xml elements produced by the VineSum’s analysis):

```
   <types:Entity xmi:id="19" sofa="12" begin="372" end="381"
entityType="vine_variety" entityValue="Grechetto"/>
   <types:Entity xmi:id="25" sofa="12" begin="104" end="115"
entityType="vine_variety" entityValue="Grechetto b"/>
   <types:Entity xmi:id="31" sofa="12" begin="209" end="214"
entityType="date" entityValue="1989."/>
   <types:Entity xmi:id="37" sofa="12" begin="88" end="93"
entityType="vine_variety" entityValue="Gamay"/>
   <types:Entity xmi:id="43" sofa="12" begin="174" end="195"
entityType="vine_variety" entityValue="Trebbiano spoletino b"/>
   <types:Entity xmi:id="49" sofa="12" begin="156" end="168"
entityType="vine_variety" entityValue="Sagrantino n"/>
   <types:WineCluster xmi:id="55" sofa="12" begin="0" end="0"
wineClusterId="-656315663"/>
   <types:WineCluster xmi:id="60" sofa="12" begin="0" end="0"
wineClusterId="2100285929"/>
   <types:WineCluster xmi:id="65" sofa="12" begin="0" end="0"
wineClusterId="477311754"/>
   <types:WineCluster xmi:id="70" sofa="12" begin="0" end="0"
wineClusterId="-273635189"/>
   <types:WineCluster xmi:id="75" sofa="12" begin="0" end="0"
wineClusterId="384708963"/>
   <types:EventCluster xmi:id="80" sofa="12" begin="0" end="0"
eventClusterId="1989_1.0_Gamay_1.0_1.0_Grechetto_1.0_0_0_Sagrantino_1.0
```

```
_0_0_0_Trebbiano_1.0_0_0_1.0_1.0_b_1.0_1.0_0_2.0_0_0_n_0_1.0_0_1.0_0_0_
0_spoletino_1.0_0_0_1.0_0_0_1.0_0_1989_1.0_Gamay_1.0_0_Grechetto_1.0_0_
1.0_Sagrantino_1.0_0_0_0_Trebbiano_1.0_0_0_0_0_b_1.0_0_1.0_0_1.0_1.0_n_
0_0_0_0_0_0_0_spoletino_1.0_0_1.0_0_1.0_0_0_0_"/>
```

The **Entity** type describes all the named-entities recognised, which may be vine varieties, persons, locations or dates.

The **WineCluster** type contains a hash, which describes a vine variety found in the text provided. If the same variety appears more than once in a single file or in multiple files, the same hash will always describe the same variety.

The **EventCluster** type contains a string calculated using the n-gram of all the named entities recognised by VineSum’s NER extractor and can be used to discover similarities between different documents.

Sources

[1] International list of vine varieties and their synonyms
http://www.oiv.int/public/medias/2273/oiv-liste-publication-2013-complete.pdf

[2] AGRIS FAO online service http://agris.fao.org

[3] DKPro Core - https://dkpro.github.io/dkpro-core/