Applied Textmining

Forth Assignment

Please update the code from the GIT repository at *https://github.com/dimalabs/applied-textmining* by pulling the latest version of the project. to accomplish that simply type the following line into *GIT Bash*:

```
git pull
```

Then have a look at *de.tuberlin.dima.textmining.assignment4.OpenInformationExtractorTest*. There is a unit test that does nothing at the moment.

Also execute `mvn eclipse:eclipse` after the update. The Stanford NER library will be added to the project. This might be useful.

1 Mine Generic Relations

Have a look at *de.tuberlin.dima.textmining.assignment4.testExtractGenericRelations*. Here implement a method in the OpenInformationExtractor class that extracts generic relations from text. The OpenInformationExtractor finds both a pair of entities and generates a predicate that characterizes the nature of the relationship between these entities. Implement at least one of the open IE patterns defined in the paper "Open Information Extraction from the Web". The paper is linked on the ISIS page.

2 Project idea

As we have seen in the lecture and the last assignments, there is a need for a lightweight framework for information extraction pipelines. The idea is that the benefits of UIMA could be implemented in a framework that works without UIMA's huge overhead by building a lightweight framework upon well-established java technologies. For example, building upon the Spring framework for loose coupling, JUnit and Maven for build management and text-driven development. Such a lightweight framework would be simple to set up and allow for fast development. Depending on the size of the project group, some solution for conveniently defining IE patterns over annotated data may also need to be found, as well as additional features added to the framework.

Deadline

Because this assignment was posted late, you have until 16th of November 2011 to upload your solution as a patch in the ISIS system .