Applied Textmining

Third 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.assignment3.ShallowExtractorTest*. Here are two unit tests that at the moment do nothing.

1 Mine Quotes

Have a look at *de.tuberlin.dima.textmining.assignment3.testExtractQuotes*. Here implement a method in the ShallowExtractor class that extracts quotes from text. We're looking for the speaker-quote relation.

One example is the sentence *"We have reached an agreement, which I believe lets us give a credible and ambitious and overall response to the Greek crisis," French President Nicolas Sarkozy told reporters.* such a pattern would find "Nicolar Sarkosy" as the speaker and "We have reached an agreement, which I believe lets us give a credible and ambitious and overall response to the Greek crisis" as the quote.

Try to write some patterns that find at least 10 different speaker-quote pairs.

2 Mine apposition

Have a look at *de.tuberlin.dima.textmining.assignment3.testExtractApposition*. Here implement a method in the ShallowExtractor class that implements the mighty apposition pattern. The apposition pattern gives us a IS-A relation.

In the example sentence above, the apposition pattern would match *French President Nicolas Sarkozy* and give us *Nicolas Sarkosy* as the entity and *French President* as a term that can be used to describe the entity. Use this pattern on the entire text to find groups of mentions for entities.

Find at least 10 entity-apposition groups. Each group consists of an entity and a list of describing terms.

3 Optional: Find coreferences

It is possible to combine the results of both patterns in two ways:

3.1 Quotes to find coreferences

Firstly, the text that is analysed is from a cluster of news, meaning that it's the same piece of news from hundreds of different online news sites. There is a lot of redundancy with only small changes. The following quote for example is mentioned in different articles. The quote is the same, but the speaker mention is different. Here, this can be used in addition to the apposition pattern to find good coreferences for entities:

"We have done what needed doing," German chancellor ***Angela Merkel*** *said.*

and

"We have done what needed doing," ***she*** *said.*

In one sentence, the speaker is referred to as "Angela Merkel", in the next as "she". Such information might be intersting :)

3.2 Apposition to find more quotes

Another way of combining both patterns is to use the groups mined with the apposition clusters. In a quote that only appears once in the cluster, but with no named entity, apposition can help identify the speaker. Take the following sentence:

"Secondly, the fund itself could have been increased, though I don't think this is going to happen," the ***analyst*** *said.*

Using an apposition group, perhaps a candidate entity for this person can be found.

4 About the data

The data is from a crawl of a google news cluster about some developments in the EU. It is the same pieve of news on a large number of sites, so don't be surprised if some patterns return the same information over and over again.

5 Project idea

This homework could be expanded into a project. We envision a system that regularly crawls the most common news sites in German or English and builds a database of speakers and quotes. This could be an interesting search feature that some users might find useful. The project would include building a frontend where a person name, such as "Angela Merkel", and one or more keywords, such as "Mindeslohn", can be entered and all public statements for this topic retrieved.

Deadline

Please upload your solution as a patch in the ISIS system until 5th of November 2011.