Information Extraction: Lab

© 2015-04-18, Fabian M. Suchanek

Prerequisites

The purpose of this lab session is to extract structured information from a natural language text corpus. Our corpus will be the Simple English Wikipedia, a simpler and smaller encyclopedia than the regular English Wikipedia. For this labs, we provide the data in a preprocessed version containing only the first sentence of each article. The file consists of repeated lines of the form

- Title of the article
- First sentence of the article
- blank line

Download this file and unzip it.

It is suggested to do this work in Python, even though you can use any other programming language. We provide the following template for your code: Template. It reads the Wikipedia file line by line and matches a regular expression on each article. It prints a triple, in which the components are separated by "\t" (tabulator).

Tasks

In all of the following tasks, a higher precision gives more points (as long as recall is still reasonable).

1. Save the template as "ie_dates.py", and modify it so that it extracts the first date mentioned in the page. It should print triples of the form "*PageTitle* \t hasDate \t *Date*". For example, it should print "Elvis Presley \t hasDate \t 8 January 1935". The data may be the birth date, but not necessarily.

   Use named regular expressions. As reference, you can use the official Python tutorial about regular expressions. Try the extractor by running the program on the entire corpus. Measure the precision on the first 20 output triples manually.

   Hand in: the code, the first 20 output triples and the calculated precision in a file named `ie_date_evaluation.txt`.

2. **Optional:** if you are adventurous, try normalizing the dates you extract with your `DateExtractor` to the form `[-]YYYY-MM-DD`
3. Save the template as "ie_types.py", and modify it so that it extracts the type of the article entity. For example, from a page starting with "Leicester is a city", it should

print the triple "Leicester \t type \t city". The subject of the triple is simply the page title, the predicate is always "type", and the object is what you extract.

Adapt the code so that it excludes terms that are too abstract ("member of...", "way of..."). Try to extract only the noun(s).

Measure the precision on the first 20 output triples manually.

Hand in: the code, the first 20 output triples and the calculated precision in a file named `ie_type_evaluation.txt`.

4. **Optional:**Write an extractor that extracts the location of a place ("Hollywood is a district in **Los Angeles**"). The result should be the triples of the form "*PageTitle* \t locatedIn \t *Place*".

   Hint: It is OK to restrict the subjects and objects to only places.

   Measure the precision of the first 20 output triples manually.

   Hand in: the code, the first 20 output triples, and the calculated precision in a file named `ie_location_evaluation.txt`.

Send all files by email to fabian@suchanek.name .