CAP 6777 Web Mining

2016 Summer

**Information Extraction Homework [15 points, Due: 06/07]**

**Part 1: Questions and Answers** [1 pt/each]

1. What is Information Extraction? Please explain how to use information extraction to support information retrieval or search.
2. What is a “Wrapper” in information extraction? Please provide an example of a Wrapper for information extraction.
3. What is “Boosted Wrapper Induction (BWI)”? Please explain how does BWI determine whether a particular input field is a target field (e.g., a name of a speaker)?
4. What is Named Entity Extraction? Please briefly explain two methods for named entity extraction.
5. Please use Regular Expression to define a pattern to extract phone numbers from a webpage (your pattern must be able to detect phone numbers in the following format (xxx) xxx-xxxx, or xxx.yyy.xxx).
6. What is Named Entity Relation Extraction? What are the examples of relations? Please briefly explain two methods for Named Entity Relation Extraction.

**Part 2: Wrapper for Information Extraction** [3 pts]

Figure 1 shows an html page which lists the largest metropolitan areas by continent. Figure 2 shows the corresponding html code of the table (which encodes the content showing in Figure 1).

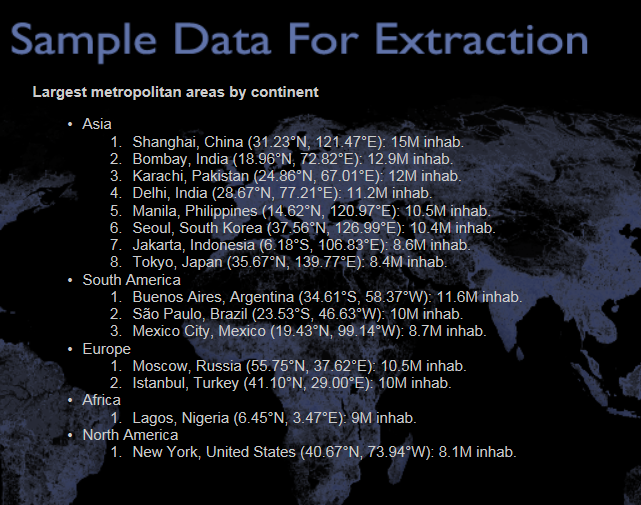
 

Figure 1 Figure 2

Please design a left-right wrapper to extract information as “City (including country)”, “Coordination”, and “Population”.

E.g., the extracted information will show as follows

|  |  |  |
| --- | --- | --- |
| City | Coordination | Population |
| Shanghai, China | 31.23°N, 121.47°E | 15M |

Please clearly mark the left-right wrapper for “City (including country)”, “Coordination”, and “Population”, respectively. (You do NOT need to carry out any implementation to extract the information, but only need to define the wrapper for each field).

**Part 3: Named Entity Recognition** [4 pts]

Stanford Named Entity Recognizer (NER), <http://nlp.stanford.edu/software/CRF-NER.shtml>, is a Java implementation of a Named Entity Recognizer developed by Stanford natural language processing group. The package can be used for named entity detection and named entity classification (classifying named entity into different categories).

The package can be downloaded from the following link:

<http://nlp.stanford.edu/software/stanford-ner-2015-12-09.zip>

The package also includes batch files for running under Windows or Unix/Linux/MacOSX.

Requirements:

1. Please download the package and report a screenshot of running the program on your computer. [1 pt]
2. Please explain what is Precision, Recall, and F-Score [1 pt]
3. Please copy the text from Jane Eyre (WikiPedia: <http://en.wikipedia.org/wiki/Jane_Eyre>) [using content from “Introduction”, “Jane's childhood”, and “Lowood”], and use NER’s “english.all.3class.distsim.crf.ser.gz” classifier to calculate the Precision, Recall, and the F-Score in identifying the 3 classes of named entities (Location, Person, and Organization) on the text collected from Jane Eyre. Please show a screenshot of the program, and also report the precision, recall and F-score in a table. [2 pts]

**Part 4: Designing Task [2 pts]**

Assume you were given a task to collect all Named Entities from Internet and find their relations to “Ebola” (e.g., the origin of Ebola, Ebola virus outbreak regions etc.). Please draw a flowchart (or diagrams) to elaborate the major steps of the project (1 pt). Please also explain the design of the experiments and the measurements to validate whether your method is working or not. [1.0 pt].