INF 558 – Spring 2019

Homework #4: Information Extraction

Due: 2/22, Friday
100 points
(TA handling this homework: Binh Vu binhlvu@usc.edu)

In this homework, you are asked to build an information extraction program using CRF to segment/chunk the course description of UCLA data set. For example, consider the following course description.

```
      Lecture, four hours; outside study, eight hours. Requisite:
course 180. Additional requisites for each offering announced in
advance by department. Selections from design, analysis, optimization,
and implementation of algorithms; computational complexity and general
theory of algorithms; algorithms for particular application areas.
Subtitles of some current sections: Principles of Design and Analysis
(280A); Distributed Algorithms (280D); Graphs and Networks (280G). May
be repeated for credit with consent of instructor and topic change.
Letter grading.
```

Your program should produce chunking of the description. You need to consider at least the following types of chunks: format, requisite, description, grading, and others (which means that they do not belong to any of previous types). Please use markup tag to indicate the type of chunk.

For example, here is the chunking of the above sample:

```
      <format> Lecture, four hours; outside study, eight hours.
</format> <requisite> Requisite: course 180. Additional requisites for
each offering announced in advance by department. </requisite>
<description> Selections from design, analysis, optimization, and
implementation of algorithms; computational complexity and general
theory of algorithms; algorithms for particular application areas.
</description> <others> Subtitles of some current sections: Principles
of Design and Analysis (280A); Distributed Algorithms (280D); Graphs
and Networks (280G). <others> <others> May be repeated for credit with
consent of instructor and topic change. </others> <grading> Letter
grading. </grading>
```

Requirements:

- The extractor component of the program should be trained using CRF suite.
- Train your program using the course description of all courses in the Computer Science department (*train-ucla.txt*).
- Prepare a test dataset using the courses from Chemical Engineering (at least 50 courses) and report the performance (precision, recall, and F1) on the test dataset. A chunk is

correct only if it is an exact match of the corresponding chunk in the data. In the above example: `<format> Lecture, four hours; outside study, eight hours. </format>` is a correct chunk, but `<format> Lecture, four hours; outside study, eight hours. Requisite</format>` is not.

- Wrap your extraction program in Python, named "extract.py", so that it can be executed as follow.
    - <program> <model> <input> <output>
    - Where <model> is the trained CRF model, <input> is a text file containing a list of course description, one per line; and <output> is a text file containing the marked up version of the description as described above.

Submission instructions:

You must submit the following files in a single .zip archive, which is named Firstname_Lastname_hw4.zip, to Blackboard.

- *extract.py*: your extraction program
- *ucla.model*: your trained model
- *test-ucla.txt*: the test dataset you have created
- *report.pdf*: A detailed report of your extraction program (describing your features and the performance on training and test dataset).