Homework 2: Information Extraction

Due Date: Wednesday, 2016/09/21 @11:59pm on Blackboard.

Summary

For this homework assignment, you will extract key information from unstructured text in webpages. You will first train a Conditional Random Field (CRF) classifier to extract the desired information, then use the classifier on a large amount of unstructured text to extract meaningful information you need. **Please read the entire homework before starting.**

Task 1 – Preliminary Work

Recall from lecture that there is usually a domain of interest when we are extracting information. The domain for this assignment will be RECENT 'disaster events', also called 'emergency incidents' (e.g. a fire in California). A disaster event does not need to involve massive loss of life.

The first step is to collect the data (from the Web) you are interested in. You can do this both manually or using a scraping/crawling tool. Recall that extractions operate at the level of word segments, so we'll define a unit of data in this assignment to be a sentence that contains at least one relevant label (see below). Your sources/sentences could cover more than one disaster event.

Furthermore, in the real-world, we see sources of varying quality. News articles are usually high quality, but blogs, people's comments, or twitter feeds may not be. As a preliminary step, identify at least **five sources** of information from the Web, where three are high quality and two are low quality (we will ask you to justify your choices in the deliverables). **Identify** at least a hundred units (sentences/sentence-like units) total from the sources.

We would like you to label each word in each unit with either the key information you want, or an “irrelevant” label (Task 2). Define the kinds of labels you will use. The information you want should be those that cannot be easily extracted from other parts of the webpage. You should define **at least three other labels** besides “irrelevant”. Examples include places, times, types of events etc. This is not as daunting as it sounds. Typically, a unit contains only a few relevant labels (by definition it must contain at least one); every other label is irrelevant.

Task 2 – Training the CRF

Your second task is to pick a CRF suite of your choice. You will not be implementing a CRF on your own. Our recommendation is the library from http://www.chokkan.org/software/crfsuite/.
Read the tutorial to understand how the library should be used. Note however that sometimes this library may not be suitable for your operating system. We do not require you to use this library, but we do expect you to find one that works on your OS. There are CRF packages available for almost every OS and programming language.

Prepare a set of training data for the CRF. To do so, you must:

- Figure out how to tokenize each sentence into a list of tokens/words.
- Decide what kind of features you would like to extract from each word/token. (E.g. part-of-speech, prefixes, suffixes). Write code or use existing tools to extract the features and append them to your data.
- Manually label tokens in **at least 100 units** for training. These are to be used as the training data.

Prepare a text file named `training` which contains the data you extracted and labelled with the following format for submission. Each row should contain N features and the label for one token in the format:

Token feature1 feature2 … featureN label

Separate each unit with an empty line.

Example of some rows:

```
California C n 1 0 0 Location
earthquake e e 1 0 0 Disaster
occurred o d 1 0 0 Irrelevant
in i n 1 0 0 Irrelevant
2008 2 8 0 1 0 Date
```

Finally, train a CRF model with your data by performing the following steps:

- Transform the format of your data to your chosen CRF suite's input format.
- Use your chosen CRF suite to train a CRF classifier with your training data

Task 3 – Questions

Answer the following questions in a report:

1. Succinctly describe the sources you identified by populating a table with 5 columns: name of source, URL, quality of source (high/low), number of (annotated) units, explanation of quality of source (at most 2 short sentences). Based on Task 1, you must have at least five sources, and a total of at least 100 annotated units.
2. Succinctly describe your features by populating a table with 3 columns: name of feature, a representative example of it, range of values (e.g. {0,1}, etc.) It's okay to describe the range (briefly) in words, especially if the feature is more open-ended.
3. Manually label **at least 20 more text** for evaluation (10 from a high quality source and 10 from a low quality source). You can use sources different from what you've listed in 1, but we prefer you just annotate more sentences from those sources. Calculate the precision, recall and F1 score of your classifier on each of the relevant categories for three different datasets (10 high quality sentences, 10 low quality sentences and for all 20 sentences). You should tabulate your results using 5 columns (precision, recall, F1, category and dataset) Did quality matter for the results? Describe in a short paragraph why or why not.

Submission Instructions

You must submit the following files/folders in a single .zip archive named `Firstname_Lastname_hw2.zip` and submit it via Blackboard:

- `Firstname_Lastname_hw2_report.pdf`: A pdf file containing your answers to **Task 3**.
- `raw`: A folder containing the (five or more) source data files, with each source in one file. It's okay to copy-paste the unstructured text into the file, rather than download the whole web-page.
- `training`: Your training data+annotations. They should be written in the format described in **Task 2**.
- `testing`: Has the same format as training but with the 20 sentences you used for testing.
- `source`: This folder includes all the code you wrote to accomplish **Task 2**

Ground Rules

This homework must be done individually. You can ask others for help with the tools, but the submitted homework needs to be your own work.