CE807 – Assignment 2 – Information Extraction

Spring 2018

School of Computer Science and Electronic Engineering - University of Essex

Assignment Due at 11:59:59am on Monday, April 23rd .

Electronic Submission URL:

https://www1.essex.ac.uk/e-learning/tools/faser2/

Please also see your student handbook for rules regarding the late submission of assignments

On Plagiarism

The work you submit must be your own. Any material you use, whether it is from textbooks, classmates, the web or any other source must be acknowledged in your work.

MOTIVATION: The lectures you attended and the labs you have done should have made it clear that Named Entity Recognition (NER) is an essential component of many text analytics applications. In the lectures you have seen how NER applications can be trained using supervised as well as semi-supervised approaches. A third approach, distant learning (or distant supervision) was only mentioned briefly. Now you have a chance to learn a lot more about it and to contextualise it with the above-mentioned alternatives.

OBJECTIVE: To train a Named Entity Recognizer using distant supervision.

SUBMISSION, ASSESSMENT AND RULES

- This assignment counts towards 50% of the overall mark for CE807.
- The assignment is to be done individually. **This is not a group assignment**.
- Be sure to put your name and registration number as a comment at the top of all code and other files.
- The assignment must be submitted in a single zipped archive containing the following subfolders:

CE807/Assignment2/	All files
CE807/Assignment2/Task1	The report produced for Task 1.
CE807/Assignment2/Task2	The code written to extract features and the files extracted in Task 2 (e.g., .arff if you use Weka)
CE807/Assignment2/Task3	The files extracted in Task 3 and the results of the evaluation.
CE807/Assignment2/Task4	The report produced in Task 4

Note: you are free to use any software you like for this assignment. Your software should run on your laptop or in one of the CSEE labs.

Using distant learning for Named Entity Recognition

Named Entity Recognition (NER) is one of the most widely used forms of information extraction.
The objective of the assignment is to develop a NER system to extract NEs from the WikiGold portion of the WikiNER corpus.

The Corpus

The WikiNER corpus:

http://schwa.org/projects/resources/wiki/Wikiner

(this page is currently unaccessible but see below) is a corpus of Wikipedia pages whose NEs have been *automatically* annotated using the distant learning methods presented in the paper:

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran (2013). **Learning multilingual named entity recognition from Wikipedia**. *Artificial Intelligence* 194:151–175.
(This highly-cited paper can easily be found online.)

In this assignment you will use the wp2 automatically annotated corpus:

```
aij-wikiner-en-wp2
```

to train your system. The file is **almost** in IOB format.

You will then evaluate your system by running it over the gold standard data:

```
wikigold.conll.txt
```

Both the training and the test data can be downloaded from the module pages.

Tasks

Your tasks will be as follows:

- In Task 1, you will **review the literature** to produce a concise critical discussion of the state-of-the-art approaches in NER with a primary focus on distant learning and the use of Wikipedia for training. We expect this to be no longer than two pages plus references.

- In Task 2, you will use any machine learning framework you wish to use (e.g., CRF++, SciKit Learn, LibSVM, etc) to **train a NER system** to identify and classify NEs in the WikiNER corpus that you will download from the CE807 Moodle page. This involves two main tasks:

    (a) download the corpus and go through the Nothman paper to understand the background
    (b) train a NER model. This in turn will involve extracting from the *training* part of the corpus the training items and their features, put them in the format required by your chosen ML framework, and create a model.

- In Task 3, you will run your trained model over the gold standard data and **evaluate its performance**. This will involve

    (a) extracting from the test corpus the test items, put them in the format required by your framework, and get the class
    (b) run the NER system on the test data (make sure the test data is in the right format)
    (c) compare your output with the gold standard.

- In Task 4, you will **write a report** explaining what you did and what you found in Tasks 2 and 3 and justify your approach. You will be asked to compare and contrast your results with what you found in Task 1.

____________________________ TASKS ________________________

TASK 1: NER literature review

Whenever you develop some text analytics software, you have to show how your system performs compared to the state of the art. The goal of this part of the assignment is to explore the landscape of approaches that have been applied to NER with a primary focus on distant learning and the use of Wikipedia for training. Discuss advantages and disadvantages of using distant learning when training a NER classifier.

TASK 2: Train a NER model

This task will only be marked if you have completed Task 1.

The goal of this part of the assignment is to demonstrate that you know how to build a NER system. *You will have to download the corpus files yourselves.* The Task2/ folder should contain the scripts you used to extract features from the training portion and any other file(s) you used to experiment.

You may use whatever software you wish for this task, including, for example, NLTK or GATE.

TASK 3: Evaluating your model

This task will only be marked if you have completed Tasks 1 and 2.

For this task, you will implement code to use the module trained in Task 2 to run over the gold standard. This work can be divided in two parts:

- Put the files in the evaluation set into your ML format, run your model, and convert back;
- Compare your output against the gold standard.

TASK 4: Report

This task will only be marked you have completed Tasks 1, 2 and 3.

Finally, you will write a report documenting what you did in Tasks 2 and 3 and comparing and contrasting your approach with what you discussed in Task 1. You should explain why you decided on the algorithm you used and how this compares to the state of the art. You should discuss the performance of your approach and reflect on what you have learned.

In your discussion point out particular issues you might have observed in your evaluation related to the fact that you applied distant learning (rather than a fully supervised algorithm).

MARKING BREAKDOWN (out of 100%)

Task 1. Literature review (NER + Distant Learning/Supervision) (20%)

○ Appropriate coverage and contextualisation:	up to 10%
○ Critical discussion:	up to 10%

Task 2. Train a NER system (30%) - Task 1 required

○ Choice of features:	up to 10%
○ Extracting the features:	up to 10%
○ Training the model:	up to 10%

Task 3. Evaluating your model (20%) - Tasks 1 and 2 required

○ Running the model on the evaluation data:	up to 10%
○ Evaluating your output:	up to 10%

Task 4. Report (30%) - Tasks 1, 2 and 3 required

○ Discussion of work carried out:	up to 10%
○ Contextualisation with related work / state of the art:	up to 10%
○ Lessons learned:	up to 10%