CE807 – Assignment 2 – Information Extraction

Spring 2017

School of Computer Science and Electronic Engineering - University of Essex

Assignment Due at 11:59:59am on Friday, March 17th .

Electronic Submission URL:

https://www.essex.ac.uk/e-learning/tools/faser/

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. Also, you are assumed to have read the booklet at http://www.essex.ac.uk/plagiarism/docs/Plagiarism_and_how_to_avoid_it.pdf

OBJECTIVE: To train a Named Entity Recognizer.

SUBMISSION, ASSESSMENT AND RULES

- This assignment counts towards 20% 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 code written to extract features and the files extracted in Task1 (e.g., .arff if you use Weka)
CE807/Assignment2/Task2	The files extracted in Task 2 and the results of the evaluation.
CE807/Assignment2/Task3	The report produced in Task3

Note: please use the filenames as indicated in the task descriptions below otherwise your work may not be marked.

Note: you are free to use any software you like for this assignment, but we cannot provide support for packages other than NLTK, Perl, and Weka. Your software should run on your laptop or in Lab 5.

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..
(The paper is available from the WikiNER site.)

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:

http://downloads.schwa.org/wikiner/wikigold.conll.txt

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

http://csee.essex.ac.uk/staff/poesio/Teach/807/Assignments/Ass2

Tasks

Your tasks will be as follows:

- In Task 1, you will use *any machine learning framework you wish to use* (e.g., CRF++, SciKit Learn, LibSVM, Vowpal Wabbit, etc) to train a NER system to identify and classify NEs in the WikiNER corpus that you will download from the CE807 Ass2 page. This involves two main tasks:
  (a) **download the corpus** and read the instructions ☺
  (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 2, 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) *put the results back in the format required by the script (IOB, read the instructions)* and (c) compare your output with the gold standard.
- In Task 3, you will **write a report** explaining what you did in Tasks 1 and 2 and why.

_____ TASKS _____

TASK 1: Train a NER model

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 Task1/ folder should contain the scripts you used to extract features from the training portion and any other file you used to experiment.

You may use whatever software you wish for this task, from Python to NLTK to GATE.

TASK 2: Evaluating your model

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

For this task, you will implement code to use the module trained in Task1 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 3: Report

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

Finally, you will write a report documenting what you did.

MARKING BREAKDOWN (out of 100%)

Task 1. Train a NER system (50%)

- Choice of features: up to 10%
- Extracting the features: up to 20%
- Training the model: up to 20%

Task 2. Evaluating your model (30%) - Task 1 required

- Running the model on the evaluation data: up to 15%
- Evaluating your output: up to 15%

Task 4. Report (20%) - Tasks 1 and 2 required

- Discussion of work carried out : up to 12.5%
- Analysis of your system: up to 7.5%