COMP4650/6490: Document Analysis

Assignment 4: Information Extraction

Main details:

Maximum marks:	10
Programming language:	Java (only)
Assignment questions:	Post to the Wattle Discussion forum
Deadline:	Q1: Lab on Wed 28 Oct and Lab on 29 Oct (automatic 0 for Q1 if fail to attend/demonstrate in the lab) Q2-Q4: Wed 28 Oct, 11.59pm (online via Wattle)

Marking scheme:

- *Written:* Full marks given for a formulation that provides a well-reasoned and succinct response to the question that addresses all requested points. There may be more than one answer for each question that achieves full marks.
- *Code:* Full marks given for working, readable, reasonably efficient, commented code that performs well on the test case given in lab.
- *Academic Misconduct Policy:* All submitted written work and code must be your own (except for any provided Java starter code, of course) — submitting work other than your own will lead to both a failure on the assignment and a referral of the case to the ANU academic misconduct review procedures:

  ANU Academic Misconduct Procedures

Electronic submission (only):

All written questions should be in a file `ANSWERS.pdf`. MS Word or other document formats are not accepted. LaTeX formatting is preferred.

Please submit `ANSWERS.pdf` and the files requested in Q1 zipped into a single file `assign4_yourname.zip` with a `README.txt` explaining the files/directories you've included.

Information Extraction: Programming

Q1 [5 pts]. NER with CRF++ tool (checked in the lab).

The task is to code a Named Entity Recognizer (NER) application in Java using the CRF++ tool and Conll2002 data sets. The CRF++ software and Conll2002 data sets are in *Topic 7, Information Extraction, Lab and Assignment resources* posted to Wattle.

First, built a NER classifier following the guidelines in the README.txt posted to Wattle. Second, write a Java application that use your NER classifier to label a new test set and output the recognized NE in the required format.

- Your Named Entity classifier must perform above 70 of F-Score (FB1) for each entity class, except the MISC class.
- Write a Java NER application that used your classifier. Prepare your application to be test by a new test set (provide in the grading lab). Note that before using your classifier to label the test set, you will need to apply a tokenizer and a POS-tagger (you can use the NLP code and the Spanish POS-tagger model in the Topic 7, Information Extraction, Lab and Assignment resources posted to Wattle).
- Your NER application must include a NE extractor that display the recognized entities organized by named entities categories, and display the frequency of each entity found (see the example/NE-ExtractorFormat.txt file).

Submit your code and the feature template. Grading will be based on your implementation (1 pt.), on the performance results of your classifier (1 pt.), on your feature template (1 pt.), your explanation about the feature template (1 pt.), and the correct output format (1 pt.).

In the lab you will be asked to use your NER application on the new test set provided at the grading lab.

Information Extraction: Written

For this section, simply submit your answers in your `ANSWERS.pdf` file.

Q2 [1 pt]. Measure the effect of the training size

Measure the effect of training size on NER accuracy, by building three NER classifier instances with 1/4, 2/4 and 3/4 of your training sentences. Evaluate the performance of each classifier using the conll-eval.pl script (the script is at: *Topic 7, Information Extraction - Lab and Assignment resources*). Summarize the results and conclude your outcomes.

Q3 [2 pt]. NER baselines

Think about two relevant baselines for the Named Entity Classification task in Q1. Remember that baselines are lower bounds of performance that can be either simple heuristics or based on simple machine learning techniques. Give a short description of them and write in pseudo-code how you would implement them.

Q4 [2 pt]. HMM for IE

Imagine you are developing an Information Extraction system using HMM that extract bibliographic information from scientific papers. The fields you want to extract are: (i) Title; (ii) Authors; and (iii) Publication date. What are the hidden states and the observation of the HMM model?