COMP4650/6490: Document Analysis

Assignment 4: Information Extraction

Main details:

Maximum marks:	10
Programming language:	Python (only)
Assignment questions:	Post to the Wattle Discussion forum
Deadline:	Q1: Graded in Lab 10
	Q2-Q4: Fri 27 Oct, 23.59 (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 the CRFsuite tool (graded in the lab).

The task is to code a Named Entity Recognizer (NER) application in Python using the CRFsuite library and the Conll2002 data sets.

To complete this task, follow the tutorial NamedEntityExtraction.ipynb and the Assignment.ipynb instructions posted in Wattle.

The following items summaries the assigment tasks:

- Built a NER classifier following the tutorial.
- Your NER model must perform above 80 of F-Score for each entity class, except the MISC class.
- Write a Python NER application that used your model. Prepare your application to be test by a new test set (provide in the grading lab). Note that before using your model to label the test set, you will need to apply a POS-Tagger.
- Your NER application must include a NE extractor that display the recognized entities organized by named entity categories.

Information Extraction: Written

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

Q2 [2 pts]. Named Entity Recognition 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.

Q3 [1 pt]. HMM

Imagine you are developing an extractive text summarization tool using HMM. What are the hidden states and the observations of the HMM model?

Q4 [2 pts]. Automatic summarization

Suppose you have to design a tool that automatically generates a summary that combines information from multiple news agencies. Describe an extractive summarization technique for this case. Your answer should specify the type of summary and the relevance method you would use with a short justification of each choice.