University of London
BSc Computing and Information Systems/Creative Computing
CO3354 Introduction to natural language processing
Coursework assignment 2 2018–19

Introduction

- Throughout this coursework assignment, '**NLTK**' refers to the Natural Language Toolkit version 3, and 'the **NLTK book**' refers to Bird, S, E. Klein and E. Loper *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. (Bejing: O'Reilly Media, 2009) available online at http://www.nltk.org/book. This version of the **NLTK book** is updated for Python 3 and NLTK 3. The first edition of the book, published by O'Reilly, is available at http://nltk.org/book_1ed/. (There are currently no plans for a second print edition of the book.)
- You should list all references at the end of your work, and they should be properly cited in-text whenever referred to. Answers that consist largely of quoted material are unlikely to get high marks, even if properly referenced.
- You should **explain** your answers and show working (where applicable) for full marks. Your main results should be given in the body of your answer, rather than relegated to appendices or additional files.
- Please submit your work as a **single PDF file**; this should include Appendices with any Python code you have written, and the results of running your code. If you have used Jupyter notebooks (recommended) you can additionally download your notebook in .ipynb format and submit it as a separate file (optional).
- Make sure your code is adequately commented – this can be done using the Markdown option in Jupyter. Comments should be grammatical, concise and avoid stating the obvious.
- **Do not upload zip files**.
- **Marks may be deducted if you do not submit your work in the required format**. There are 100 marks available for this coursework assignment.

Your coursework assignment should be submitted as a single PDF file, using the following file-naming conventions:

YourName_SRN_COxxxxcw#.pdf (e.g. MarkZuckerberg_920000000_CO3354cw2.pdf)

- ***YourName*** *is your full name as it appears on your student record (check your student portal);*
- ***SRN*** *is your Student Reference Number, for example 920000000;*
- ***COXXXX*** *is the course number, for example CO3354; and*
- ***cw#*** *is either cw1 (coursework 1) or cw2 (coursework 2).*

If you submit a Jupyter Notebook file, you should name it following the same convention, with .ipynb in place of .pdf.

REMINDER: It is important that your submitted coursework assignment is your own individual work and, for the most part, written in your own words. You must provide appropriate in-text citation for both paraphrase and quotation, with a detailed reference section at the end of your assignment (this should not be included in any word count). Copying, plagiarism and unaccredited and wholesale reproduction of material from books or from any online source is unacceptable, and will be penalised (see our guide on how to avoid plagiarism on the VLE).

Question 1: Part-of-speech tagging

Before tackling this question you should work through chapters 3 and 5 of the NLTK book, in particular the sections on "Accessing Text from the Web and from Disk" and "Automatic Tagging". You may also find it useful to read sections 8.1-8.3 of Jurafsky and Martin's draft 3rd edition of *Speech and Language Processing*.

a. For this question, you will be working with the Regular Expression (RE) tagger shown in Chapter 5, Section 4.2 of the **NLTK book**.

   i. Evaluate the tagger as it is against the tagged sentences in NLTK's sample treebank corpus and report your result.
   ii. Improve the accuracy of the tagger as much as you can by coding REs for parts of speech that are not covered in the book version, particularly closed-class words such as prepositions and conjunctions, and report your results. You should briefly explain any changes you make.

   [15 marks]

b. This question involves training and evaluating taggers on the treebank corpus.

   i. Find the most common POS tag in the corpus. Train and evaluate a unigram tagger and a bigram tagger individually, then evaluate them using **backoff** with a default tagger as the last resort.
   ii. Repeat the backoff tagging with your RE tagger in place of the default tagger, and compare the results.
   iii. A substantial number of items in the treebank corpus are tagged as -NONE-. Extract a random sample of 50 of these items and propose reasons why they have not been classified into any grammatical categories.

   [15 marks]

c. For this question you will be working with a sample of raw text. Prepare for the question by downloading the Plain Text UTF-8 version of the novel *Moby-Dick* from Project Gutenberg, at https://www.gutenberg.org/ebooks/2701

   i. Begin by segmenting the text into sentences, and use the NLTK POS tagger to create a list of tagged sentences.
   ii. Evaluate your backoff taggers from (b) against this new dataset and report the results.
   iii. Repeat (ii), but this time you should train the unigram and bigram taggers on the tagged sentences from *Moby-Dick*. Which combination of taggers gives the most accurate results?

   [20 marks]

d. Based on the recommended readings and any relevant sources from your independent reading, **briefly** (up to one page) describe any other techniques that could further improve the accuracy of your taggers.

   [10 marks]

Question 2: Information Extraction

Before tackling this question you should read through Chapter 17, "Information Extraction" (IE), in Jurafsky and Martin's draft 3rd edition of *Speech and Language Processing*.

a. **Briefly** describe the following IE tasks (no more than 100 words each):

   i. Named Entity Recognition
   ii. Relation Extraction
   iii. Event Extraction
   iv. Extracting Temporal Expressions
   v. Template Filling

[10 marks]

b. IE typically uses a combination of rule-based and machine learning techniques. Describe one rule-based, and one machine learning approach to **one** of the above tasks (these may be the same task or two different ones). Write no more than 800-1,000 words.

[20 marks]

c. Much progress in IE has been driven by formal evaluations with shared benchmark datasets, such as the Message Understanding Conferences initiated by DARPA in 1987. Give an account of **one** of these evaluations, drawing on the recommended readings for this course and your own independent reading. Write no more than 500 words.

[10 marks]

[Total: 100 marks]

[END OF COURSEWORK ASSIGNMENT 2]