University of London International Programmes

CO3354 Introduction to Natural Language Processing

Coursework assignment 2 2016–2017

Due date: 8th April 2017

IMPORTANT

Your Coursework assignment should be submitted using the following file-naming conventions:

FamilyName_SRN_COxxxxcw#.pdf (e.g. Zuckerberg_920000000_CO3354cw2.pdf)

- **FamilyName** is your family name (also known as last name or surname) as it appears in 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).

Notes

- Throughout this coursework assignment, 'NLTK' refers to the Natural Language Toolkit **version 3**, and 'the NLTK book' refers to *Natural Language Processing with Python* by Steven Bird, Ewan Klein and Edward Loper, which is 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 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) to gain full marks.
- Please submit your work as a PDF file, with an appendix including any Python code that you have written and the results of running your code. If you have used Jupyter (recommended), you can download your notebook in .ipynb format and submit it as a separate file. Please make sure your code is adequately commented – this can be done using the Markdown option in Jupyter. Marks may be deducted if you do not submit your work in the required format.
- There are 100 marks available for this coursework assignment.
- Any websites referenced in this coursework assignment were last visited on 18 January 2017.

Question 1: Classification [35 marks]

a) Suppose a corpus contains 300,000 word-tokens and 70,000 of these are tagged as N (common noun). The word-form *house* occurs 1,000 times in the corpus, and it is tagged either as N or V. Analysis shows that *house* accounts for 0.4% of all common noun tokens in the corpus. Use Bayes' formula to calculate the probability that a given occurrence of *house* is tagged as N. Explain your answer and show your working. Show your results to two significant figures.

b) Run the Naive Bayes movie review classifier described in Chapter 6, Section 1.3 of the NLTK book and generate a list of the 30 features the classifier found to be most informative. Discuss why these features might be informative and whether you find any of them surprising. NB: read through the Supplements to the subject guide before attempting this question.

c) Test the classifier on the following reviews of the film 'La La Land' (2016), which divided critical opinion. Discuss the results.

    i. A negative review by Rebecca Lewis at: http://metro.co.uk/2017/01/12/dont-believe-the-la-la-land-hype-6374470/

    ii. A positive review by Ian Freer at: http://www.empireonline.com/movies/la-la-land/review/

    iii. One positive and one negative user review of your choice from the Internet Movie Database (give the links in full, and include the **full text** as an appendix with title, author, date of the review and date you visited the site): http://www.imdb.com/title/tt3783958/reviews

Question 2: Information Extraction [35 marks]

Study the section on **Chunking** in Chapter 7 of the NLTK book, particularly the section on ‘Recursion in Linguistic Structure’.

Extend the chunk grammar from Example 2.3 in Chapter 7 to handle complex NPs, that include *e.g.* PPs or conjunction: *the cat in the hat*, *Rosenkrantz and Guildenstern*, *president and Chief Executive Officer*, *executive vice president of this electric-utility holding company*. You should use the Penn Treebank tagset, **not** the universal set. Explain the reasons for your extensions to the grammar.

Test your chunker on sentences 200–220 of the tagged WSJ corpus `nltk.corpus.treebank.tagged_sents()`.

Compare your outputs with the parsed versions of the sentences in this corpus, and discuss any cases where your chunker seems to have failed to identify complete NP structures.

Question 3: Evaluating Chatbots [30 marks]

The Loebner prize is an annual ‘Turing Test’ competition where computer systems are judged on their ability to pass as a human being in unrestricted conversation. The contest for 2015 was run under the aegis of the AISB (Society for the Study of Artificial Intelligence and Simulation of Behaviour). The contest included a preliminary round where each entrant was presented with a fixed set of 20 questions, and the four highest scoring entrants went on to compete in the final. Transcripts from this round can be found at:

http://www.aisb.org.uk/events/loebner-prize#finals2016

For this question, you should compare the performance of the four finalists.
Your answers should address such issues as:

- What can we learn from these examples about the challenges of simulating human interaction?
- What particular problems did the higher-scoring systems appear to have solved more effectively?
- Why did even the highest scorers still fail to convince the judges they were human?

Do not write more than about 600–800 words, (excluding reference section).

[TOTAL 100 MARKS]

END OF COURSEWORK ASSIGNMENT 2