University of London International Programmes
CO3354 Introduction to Natural Language Processing
Coursework assignment 2 (2015–16)

Due date: 8th April 2016

Notes

- Throughout this coursework assignment, ‘NLTK’ refers to the Natural Language Toolkit **version 2**, and ‘the NLTK book’ refers to *Natural language processing with Python* by Bird, Steven, Ewan Klein and Edward Loper (2009). Although version 3 of the NLTK is now available, students are **recommended** to continue using version 2 for compatibility with the course materials.
- All websites cited below were last visited on December 17th 2015.
- You should list all references at the end of your work, and they should be properly cited whenever referred to.
- Where you are asked to ‘explain your answer’, unless otherwise stated you should write no more than one or two sentences.
- Please submit your work as a single PDF file, with an appendix including any Python code you have written and the results of running your code. Any additional files will be disregarded.
- Follow instructions specified for electronic submission: ensure that you include your full name, student number, course code and coursework assignment number.
    - e.g. 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 CO3355, and
    - **cw#** is either cw1 (coursework 1) or cw2 (coursework 2).

There are 100 marks available for this coursework assignment.

Question 1 [25 marks]

a) Explain the following terms in your own words:

   i. tag
   ii. supervised learning
   iii. n-gram tagger
   iv. inter-annotator agreement
   v. backoff tagger.

b) Extend the RE tagger from section 5.4 of the NLTK book to cover closed-class words such as prepositions and determiners (See the UCL Internet Grammar of English for a definition of ‘closed-class’). Test it on the ‘news’ category of the Brown corpus and discuss whether and how much your changes improve the accuracy of the tagger.

Question 2 [30 marks]

a) Modify the gender features function from section 6.1 of the NLTK book to ‘build the best name gender classifier you can’, following the instructions in section 6.10 of the online version of the book. Report on the level of accuracy you achieve using either the Naïve Bayes, Decision Tree or MaxEnt classifier.

b) Test your new classification function using the remaining two classifiers. Discuss any differences in accuracy and execution time.

c) Finally, test your classifier on the lists of most popular names for newborn boys and girls in the UK: http://www.independent.co.uk/news/uk/home-news/baby-names-top-100-most-popular-boys-and-girls-names-10459074.html
Are there any names that do not occur in NLTK’s names corpus?

Question 3 [25 marks]

a) Modify the noun phrase chunker in Example 7.3 of the NLTK book by including up to four additional tags from the Penn Treebank tagset. Explain why you have chosen these tags. Run it against the NLTK Wall Street Journal Corpus. List the NP chunks from the first sentence, and the 10 longest chunks from the corpus.

b) Download the plain text UTF-8 version from Project Gutenberg of Chapter 1 of Gibbon’s *Decline and fall of the Roman Empire*. Write a Python program to extract NP chunks that refer to named Roman emperors.

Question 4 [20 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#Results2015.

For this question, you should compare the performance of the four finalists.

Your answers should address such issues as:

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

[Total 100 marks]

[END OF COURSEWORK ASSIGNMENT 2]