University of London International Programmes
CO3354 Introduction to Natural Language Processing

Coursework assignment 1 2015–16

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 Steven Bird, 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. The original version of the textbook is available at http://www.nltk.org/book_1ed.
- 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.
- 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_CO3354cw1.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 CO1108, and
    - **cw#** is either cw1 (coursework 1) or cw2 (coursework 2).

Part 1: Syntax and formal grammars [40 marks]

Question 1 [25/40]

a) Explain what is meant by **context-free** and **regular** grammars in the context of natural language processing, and the important differences between them. Which of these formalisms is generally considered to be more appropriate for expressing grammar rules for natural languages, and why? You should write no more than about 500 words.

b) Table 1 represents a non-deterministic finite state machine (FSM) where q1 is the starting state and q6 is the halting state.

i. How long is the shortest sentence that it will accept?
ii. How long is the longest sentence it will accept that does not repeat any of the personal names?

iii. Write out three more sentences that will be accepted by the FSM. Write two that will not be accepted, but are grammatical in ordinary English and use the same vocabulary.
iv. Write a formal grammar with equivalent coverage to the FSM, made up of rules of the form $X \rightarrow Y\ Z$ ...

Explain your answers.

c) These questions involve **regular expressions** (REs).

i. Write out five strings that match the RE **a+b***
ii. Write out five strings at least three characters long that match the RE **(a+|b)(c|d)***
iii. Write a regular expression that matches all English words that contain each of the letters 'q', 'x' and 'y' exactly once and in that order, and does not match any other English words. List any words that match from the wordlist in the NLTK 'words' corpus. An example that should occur in your results is '*quixotically*'.

You should explain your answers for full marks.

q1	John	q2
q1	Alice	q2
q2	and	q3
q3	Mary	q4
q3	Bob	q4
q4	often	q5
q4	and	q1
q4	or	q1
q5	argue	q6
q5	meet	q6

Table 1.

Question 2 [15/40]

a) Using the grammar rules on page 22 of the subject guide, construct syntax trees for the sentences:

i. The girl likes the cat
ii. If the boy eats burgers then either the girl likes cream or the girl eats a cake

You may show the trees either in graphical form or using labelled brackets as in the example on page 22 of the subject guide.

b) Modify the grammar so that it generates the following sentences. Show worked examples.

   i. If the girl likes sweet things then she will eat cake or candy
   ii. The cat will run if the dog chases it
   iii. If the boy is tired he will sleep

Part 2: Corpora and basic text analysis [60 marks]

Question 3 [30/60]

a) Explain the following terms **in your own words**, giving examples from the NLTK or elsewhere.

   i. Synonyms, antonyms, hyponyms and hypernyms
   ii. Lexicon
   iii. Parallel corpora
   iv. Overlapping corpus structure
   v. Concordances
   vi. Lexical diversity

b) Download the plain UTF-8 texts of the 19th century novels Dracula and Frankenstein from Project Gutenberg.

   i. Tabulate the number of times the following words occur in each text: *'good', 'evil', 'moral', 'science', 'monster', 'creature'*.
   ii. List the 50 most common words in each text, excluding stopwords and punctuation.
   iii. List the **collocations** for each text, as reported by the NLTK.
   iv. Do your results suggest any differences in the concerns and subject matter of these novels? If so, give details.

Question 4 [30/60]

a) Explain what is meant by **stemming**, with examples, and how it can be useful for various applications in text analysis. Describe some approaches to stemming with particular reference to the Porter, "Snowball" and Lancaster (Paice-Husk) algorithms. You should write no more than about 500 words.

   Some references (in addition to those listed in the subject guide):

Snowball: A language for stemming algorithms.
MF Porter, 2001.
http://snowball.tartarus.org/texts/introduction.html

A Comparative Study of Stemming Algorithms
Anjali Ganesh Jivani, 2011.
http://www.ijcta.com/documents/volumes/vol2issue6/ijcta2011020632.pdf

b) Run the quoted text below through the NLTK implementations of the Porter, Snowball and Lancaster stemmers and the WordNet Lemmatizer, and compare the results.

   i. See if you can reconstruct lists of rules that each of the programs appears to have applied. Explain your answers.
   ii. Discuss any cases where you believe the wrong decisions have been made.

Text:

Belting out a tune alongside hundreds of others increases your tolerance for pain, improves happiness, and makes you feel better-connected to your social group, psychologists from Goldsmiths, University of London and the University of Oxford have found. While it is already known that singing with others has positive effects, this new study is the first to confirm that the effects can be 'scaled up' to significantly larger groups. The researchers say their findings are consistent with the idea that music-making may have allowed our early ancestors to increase their community size beyond those found in other primates.

(Goldsmiths website)

[Total 100 marks]

[END OF COURSEWORK ASSIGNMENT 1]