University of London International Programmes

CO3354 Introduction to Natural Language Processing Coursework assignment 1 2016–2017

Due date: February 15th 2017

Notes

- Throughout this 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, 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 workings (where applicable) for full marks.
- Please submit your work as a PDF file. This should include 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.
- Please use the regular naming convention for submitted files:

  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 CO3354, and
  - **cw#** is either cw1 (coursework 1) or cw2 (coursework 2).

- There are 100 marks available for this coursework assignment.

Question 1: Syntax and formal grammars [50 marks]

a) Briefly explain what is meant by **context-free** and **regular** grammars in the context of natural language processing, and the important differences between them. Modify the sample grammar on page 22 of the subject guide so that it will generate examples (i–v) below but not the starred examples (vi–ix). Explain whether context-free or regular grammar rules are more appropriate for making these distinctions.

i. If the girl likes sweet things then she will eat either cake or candy
ii. If the girl likes sweet things she will eat cake or candy
iii. The cat will run if the dog chases it
iv. If the boy is tired he will sleep
v. The boy is tired so he will sleep
vi. *If the girl likes sweet things or she will eat cake
vii. *If the boy is tired if he will sleep
viii. *If the cat will run so the dog chases it
ix. *The boy will eat cake or or candy

a) Table 1 represents a non-deterministic finite state machine (FSM) where q1 is the starting state and q8 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 that 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 $X \rightarrow Z$ where X is a single non-terminal symbol and Z is a non-empty sequence of terminals and/or non-terminals.

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

i. Write out five strings that match the RE **a+b***.
ii. Write out five strings that match the RE **(a+b)***.
iii. Write out five strings that are at least three characters long that match the RE **(a|b+)(c|d)***.
iv. Write a regular expression that matches all English words that contain each of the letters ‘a’, ‘e’, ‘i’, ‘o’ and ‘u’ exactly once and in that order, and that 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 ‘*abstemious*’.

State	Input	New state
q1	Jordi	q2
q1	Elena	q2
q2	and	q3
q3	Fatima	q4
q3	Leon	q4
q4	will	q5
q4	often	q5
q4	and	q1
q4	or	q1
q5	argue	q6
q5	meet	q6
q6	,	q7
q6	.	q8
q7	while	q1

Table 1.

Question 2: Corpora and basic text analysis [50 marks]

a) Read Section 3.6 of the NLTK book, 'Normalising Text', and Chapter 2.3 of the draft 3rd edition of Jurafsky and Martin's *Speech and Language Processing* at http://web.stanford.edu/%7Ejurafsky/slp3/. Explain why text generally has to be normalised before any other language processing, and briefly describe, giving examples, the following: **word tokenisation**, **stemming**, **lemmatisation** and **sentence segmentation**. Describe some approaches to stemming with particular reference to the Porter, 'Snowball' and Lancaster (Paice-Husk) algorithms. You should write no more than about 500–800 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) Download the plain UTF-8 texts of the novels *The Virginian* https://www.gutenberg.org/ebooks/1298 and *The Hound of the Baskervilles* https://www.gutenberg.org/ebooks/2852, both first published in 1902.

   i. Tabulate the number of times the following words occur in each text: *'law','justice','death','pistol','fear','he','she','crime','cowboy','detective'*
   ii. List the 50 most common words in each text, excluding stop words and punctuation.
   iii. List the **collocations** for each text, as reported by the NLTK.
   iv. Do your results suggest any similarities or differences in the concerns, subject matter and style of these novels? If so, give details.

c) 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:

Tone deaf people may find it harder to read facial expressions or tell whether someone's laugh is real or fake, research from Goldsmiths, University of London suggests.

A study found that participants diagnosed with congenital amusia (an inherited defect in musical memory, recognition or processing pitch) were less likely to accurately identify silent facial expressions and emotional vocalisations than those with a typical music-processing ability.

And when processing laughter, participants with amusia showed a reduced sensitivity to whether or not the act was authentic. However, they found laughter to be as contagious as those with normal development do.

Goldsmiths psychologists,Professor Lauren Stewart and Professor Daniel Müllensiefen, with Dr César F. Lima and colleagues from Goldsmiths and UCL, believe their findings suggest a developmental music disorder is not just an impairment restricted to auditory information, and it can affect socio-emotional cognition in other subtle ways.

(Goldsmiths website)

[END OF COURSEWORK ASSIGNMENT 1]