University of London International Programmes

Computing and Information Systems/Creative Computing

CO3354 Introduction to natural language processing

Coursework assignment 1 2017–18

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, 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.
- 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 (recommended) you can download your notebook in .ipynb format and submit it as a separate file (optional). Please make sure your code is adequately commented -- this can be done using the Markdown option in Jupyter. **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_CO3354cw1.pdf)

- ***YourName** is your full name 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).*

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: Syntax and formal grammars

a)

i. Briefly explain what is meant by **context-free** and **regular** grammars in the context of natural language processing, and the important differences between them, using natural language examples.

ii. Modify the sample grammar on page 22 of the subject guide so that it will generate examples (1–5) below but not the starred examples (6–10). **Explain** whether context-free or regular grammar rules are more appropriate for making these distinctions.

1. If the dog likes sweet things then it eats candy.
2. If the dog is tired it sleeps.
3. The boy sleeps when he is tired.
4. Either the boy eats cake or he eats burgers.
5. The girl likes sweet things so she will eat cake or candy.
6. If the dog likes sweet things so it eats candy.
7. *If the dog is tired or it sleeps.
8. *Either the boy sleeps so he is tired.
9. *If if the cat is tired then it will sleep.
10. *The boy will will eat cake.

[20 marks]

b) Table 1 represents a non-deterministic finite state machine (FSM) where q1 is the starting state and q9 is the halting state. In the questions below, the length of a sentence concerns the number of words or punctuation symbols, rather than characters.

State	**Input**	**New state**
q1	Abdul	q2
q1	Beatrice	q2
q2	and	q3
q3	Carlos	q4
q3	Demetria	q4
q4	often	q5
q4	and	q1
q4	or	q1
q5	meet	q6
q6	for	q7
q6	.	q9
q7	tea	q8
q7	dinner	q8
q7	discussions	q8
q8	or	q7
q8	.	q9

Table 1

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 → Z where X is a single non-terminal symbol and Z is a non-empty sequence of terminals and/or non-terminals.

[15 marks]

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 match the RE ^(a+|b)(c|d*)$

iv. Write a regular expression that matches all English words that include the string ‘psych’, and does not match any other English words. List any words longer than 12 characters that match your RE from the wordlist in the NLTK ‘words’ corpus. An example that should occur in your results is ‘metempsychosis’. You should list your results in an Appendix.

[15 marks]

Question 2: Corpora and basic text analysis

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/.

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

i. Explain the differences between lemmatisation and stemming, as well as the different approaches implemented in the Porter and Lancaster (Paice-Husk) stemmers.

ii. Run the quoted text below through the NLTK implementations of the Porter, Snowball and Lancaster stemmers and the WordNet Lemmatizer. Discuss 10 examples that illustrate the functional differences between these tools.

iii. Are there any instances where you think one or more of the applications has made errors? Are these errors of commission or omission? Justify your answers.

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

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

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

Cristian Moral, Angélica de Antonio, Ricardo Imbert and Jaime Ramírez *A survey of stemming algorithms in information retrieval* (2014)
http://www.informationr.net/ir/19-1/paper605.html#.WeTBkcbTVXQ

[30 marks]

Text for stemming exercise:

Tests developed by Goldsmiths, University of London psychologists will help shape a special series of events exploring music's unique capacity to be remembered.

BBC Radio 3 and Wellcome Collection's 'Why Music? The Key to Memory' series takes place from Friday 13 to Sunday 15 October and will feature live performances, one-off broadcasts and wide-ranging discussions.

Ahead of the events, BBC Radio 3's Music Matters is asking listeners to complete three short assessments developed by Goldsmiths Reader in Psychology, Dr Daniel Müllensiefen and others in the Music, Mind and Brain research group. These are a melodic memory test, musical training questionnaire and a rhythm memory test.

The latter is based on a previous musical sequence transcription test initially developed by music teacher and researcher Paulo Estevao Andrade, who is currently studying the MSc Music, Mind and Brain course at Goldsmiths.

The findings and implications from the tests will then be discussed on a special edition of Music Matters with Tom Service on Saturday 14 October at 12.15pm.

Dr Müllensiefen said: 'We are very pleased to be working with the BBC on this project and it will be very exciting to see the findings. We expect to find substantial differences in the ability to remember music, and that some people are better at remembering rhythm than melodies and vice versa.'

'The results will tell us to what degree memory for melodies and rhythms are related and whether there are "rhythmic" and "melodic" listeners. We might also gain insight into the degree to which musical training benefits melodic and rhythmic abilities. This could provide some evidence for or against the common conception that some people happen to be musically gifted or just "have rhythm".'

The tests can serve as diagnostic tools and participants can train the skill where they feel they didn't perform well.

Furthermore, rhythmic memory has been linked to auditory working memory and literacy skills. Especially in children, simple rhythmic tests like those used in this study can be early indicators of being at risk for developing learning disabilities like dyslexia.

http://www.gold.ac.uk/news/mullensiefen-bbc-radio-3/

b) Download the plain UTF-8 texts of the novels *A Portrait of the Artist as a Young Man* by James Joyce (http://www.gutenberg.org/ebooks/4217)
and *The King of Ireland's Son* (Padraic Colum)
https://www.gutenberg.org/ebooks/3495; both were published in 1916.

i. Tabulate the number of times the following words occur in each text: 'King','Queen','Ireland','church','fear','he','she','sword','tree','hell', 'train'.

ii. List the 200 most common words in each text, excluding stop words and punctuation. Include the list in an appendix.

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.

[20 marks]

[Total 100 marks]

[END OF COURSEWORK ASSIGNMENT 1]