University of London International Programmes
Computing and Information Systems/Creative Computing

CO3354 Introduction to natural language processing

Coursework assignment 2 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.)
- A secondary reading for this coursework assignment is Jacob Perkins, *Python 3 Text Processing with NLTK 3 Cookbook.* (Packt Publishing, 2014).
- Any websites referenced in this coursework assignment were last visited on 15 February 2018.
- 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 gain high marks, even if properly referenced.
- In order to gain full marks, you should **explain** your answers and show working (where applicable).
- Please submit your answers as a **PDF file**. **If you have used Jupyter (recommended) you should download your notebook in .ipynb format and submit it as a separate file; otherwise your PDF should include an appendix showing your code and results**. In any case, you should include your main results in the body of your answers in the PDF 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.

Your coursework should be submitted as a single PDF file, using the following file-naming conventions:

YourName_SRN_COxxxxcw#.pdf (e.g. MarkZuckerberg_920000000_CO3354cw2.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).*

Important 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 coursework 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 penalized (see our guide on how to avoid plagiarism on the VLE).

Question 1: Classification

a) (Adapted from an exercise in the NLTK book.) Run the Naive Bayes movie review classifier described in Chapter 6, Section 1.3 of the NLTK book.

  i. Using the same training and test data, and the same feature extractor, build three classifiers for the task: a decision tree, a naive Bayes classifier, and a Maximum Entropy classifier. Compare the performance of the three classifiers on your selected task. Use these parameters for MaxEnt: max_iter=10,algorithm='gis'.

  ii. Experiment with changing the feature extractor for **one** of these classifiers. Can you improve its accuracy?

Note: read through the 'Supplements' to the subject guide before attempting this question.

b) For this question you should build a classifier which distinguishes fiction from non-fiction in the Brown corpus, roughly defined as follows:

- Fiction: 'adventure', 'fiction', 'mystery', 'romance', 'science fiction'
- Non-fiction: 'belles lettres', 'editorial', 'government', 'learned', 'news'.

  i. Compare the performance of decision tree, naïve Bayes and MaxEnt classifiers. List the 20 most informative features for the naïve Bayes classifier, and discuss why these features might be informative.

  ii. Test **one** of your classifiers on the following texts:

    1. 'Premium Harmony' by Stephen King:
       https://www.newyorker.com/magazine/2009/11/09/premium-harmony

    2. 'The Ones Who Walk Away From Omelas' by Ursula K. LeGuin:
       http://engl210-deykute.wikispaces.umb.edu/file/view/omelas.pdf

    3. The Constitution of the United States of America:
       https://www.gpo.gov/fdsys/pkg/CDOC-110hdoc50/pdf/CDOC-110hdoc50.pdf

    4. The first 10,000 characters of NLTK's Australian Broadcasting Corporation corpus.

[35 marks]

Question 2: Information extraction

Study the section on **Chunking** in Chapter 7 of the NLTK book, particularly the section on ‘Recursion in Linguistic Structure’, and optionally, Chapter 5 of the NLTK3 Cookbook by Jacob Perkins.

Evaluate the chunk grammar from Example 2.2 in Chapter 7 of NLTK against the conll2000 corpus using the following code:

```
from nltk.corpus import conll2000
score = chunker.evaluate(conll2000.chunked_sents())
score.accuracy()
```

where chunker is a chunk parser you have built using the grammar rules with RegexpChunkerParser.

You should expect to get a very low score. For this question, you are required to improve the accuracy and coverage of the chunk grammar by (among others):

- adding rules for non-NP categories such as VP, PP;
- dealing with complex NPs which include subordinate PPs, coordination, etc.;
- any other extensions of the grammar you find useful;
- using additional techniques such as chinking, merging and splitting (not obligatory, but may gain extra credit).

You should report the accuracy result for the final version of your chunker, as well as the following values:

- score.precision()
- score.recall()
- score.missed()
- score.incorrect()
- score.correct()
- score.guessed()

These values may also help you decide how your chunk grammar needs to be improved.

[35 marks]

Question 3: Evaluating Chatbots

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 2017 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.aomartin.co.uk/uploads/loebner_2017_finalist_selection_transcripts.pdf

.

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 involved in processing natural language communication beyond the level of the sentence?
- What particular problems did the higher-scoring systems appear to have solved more effectively? Which problems were hard even for the high-scoring entrants?
- What tricks or devices have the developers used to make conversations seem more ‘natural’?

Do not write more than about 800–1,000 words.

[30 marks]

[Total 100 marks]

[END OF COURSEWORK ASSIGNMENT 2]