
Coursework commentaries 2015–16

C03354 Introduction to natural language processing

Coursework assignment 2

General remarks

- Throughout this 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, you are recommended to continue using version 2 for compatibility with the course materials. The original version of the textbook is available at www.nltk.org/book_1ed
- All websites cited below were last visited on 17 December 2015.
- You should list all references at the end of your work; these should be properly cited in the text whenever referred to.
- Where you are asked to ‘explain your answer’, unless otherwise stated, you should write no more than one or two sentences. A number of students lost marks by failing to give explanations as required.
- Work should be submitted 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 were disregarded. All students submitted PDF files as instructed; however a number did not include their code, execution logs and results.

Comments on specific questions

Question 1

a. Bookwork question asking for explanations of terminology.

You should refer to the subject guide and/or the NLTK book. Given that this question essentially involved looking up definitions and correctly paraphrasing them, there were a surprising number of disappointing answers, though some were very clear and thorough.

b. Extending a regular expression (RE) tagger from the NLTK book and discussing how this can improve the accuracy of results.

Marks were awarded for extending the tagger, testing it properly and discussing the results. Some students lost marks by not showing any results and/or failing to **discuss** the effects of their changes, though there were also some very good answers. You should pay attention to the **ordering** of REs, as this makes a difference to the output. It is also important to be clear about what is meant by **closed-class** words. The question mentioned prepositions and determiners: students got extra marks by adding other categories such as pronouns and conjunctions.

Question 2

This practical question involved training a classifier using different machine learning (ML) techniques and applying it to novel data.

The instructions (in section 6.10 of the online version of the book) involve splitting the names corpus into three subsets: the test set, the dev-test set and the training set, and making incremental improvements which are checked against the dev-test set.

Marks were awarded for:

- following the specified procedure
- sensible, well-motivated improvements, incrementally tested
- clear reporting and discussion of results.

Not all students followed the instructions (e.g. on the size of dev-test or even using a dev-test set at all). ‘Incremental’ means testing and reporting one change at a time, which many did not do. The idea was to test incremental changes using error analysis on the dev-test set and then to run the final version against the test-set. Many students seem to have missed this point. Some answers were very thorough and well-presented, including all the required outputs, though several students gave only partial answers or omitted this question.

Some students simply referred to a code appendix without discussion or explanation – this is not acceptable, and cost marks.

Question 3

Again, in answering this question some students simply referred to a code appendix without discussion or explanation – you should always write to make it easy for your reader.

- a. This was a practical problem involving chunking. The ‘NLTK Wall Street Journal Corpus’ in the question referred to the excerpts from the Penn Treebank corpus. There was some confusion about this, and some students worked from the conll2000 corpus. These answers were accepted for marking.

Marks were awarded for:

- basic ability to load the corpus and run a chunker
- reasonable selection of additional tags
- reporting results as specified
- explanation and working.

There were several good answers but some students omitted this subquestion or lost marks by failing to show their code, providing inadequate explanation of their results, or poor presentation. Some only made minor changes to the original chunker and so obtained few marks.

- b. This question involved writing a Python program to process chapter 1 of Gibbon’s *Decline and fall of the Roman Empire*, and extracting noun phrase (NP) chunks that refer to named Roman emperors. One approach to this problem uses the `ie_preprocess()` function from the NLTK book, section 7.1, which connects the sentence segmenter, word tokenizer and POS tagger, in combination with the NLTK tree functions, to search for subtrees which have ‘emperor’ as the leftmost daughter followed by one or more proper names. Please note, this also picks up the emperors Frederick and Philip; not all students seemed aware that these are not names of Roman emperors (unless they were counting the Holy Roman Empire?) Another, more simple-minded technique was to start with a full list of Roman emperors to match against chunks including NNP (proper names). This probably had better recall, though

may have matched people with the same name who were not in fact emperors. Some students gave very good answers with acceptable results, while several others lost marks by skipping the subquestion or giving solutions that either failed to address the problem, got no results or had poor recall.

Question 4

This was an essay question addressing the performance of the four finalists in the 2015 Loebner Prize competition, discussing issues including the following.

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

Marks were awarded for sensible, well-motivated answers covering the above (or similar points) and for clarity, appropriate structure and persuasive argumentation. High-scoring answers backed up their arguments with reference to specific, appropriate examples from the dialogues rather than generalities. A number of students omitted this question, while others lost marks through answers that were too terse and/or lacking in well-chosen examples.