Coursework commentary 2018–2019

CO3354 Introduction to natural language processing

Coursework assignment 1

General remarks

Students were reminded that:

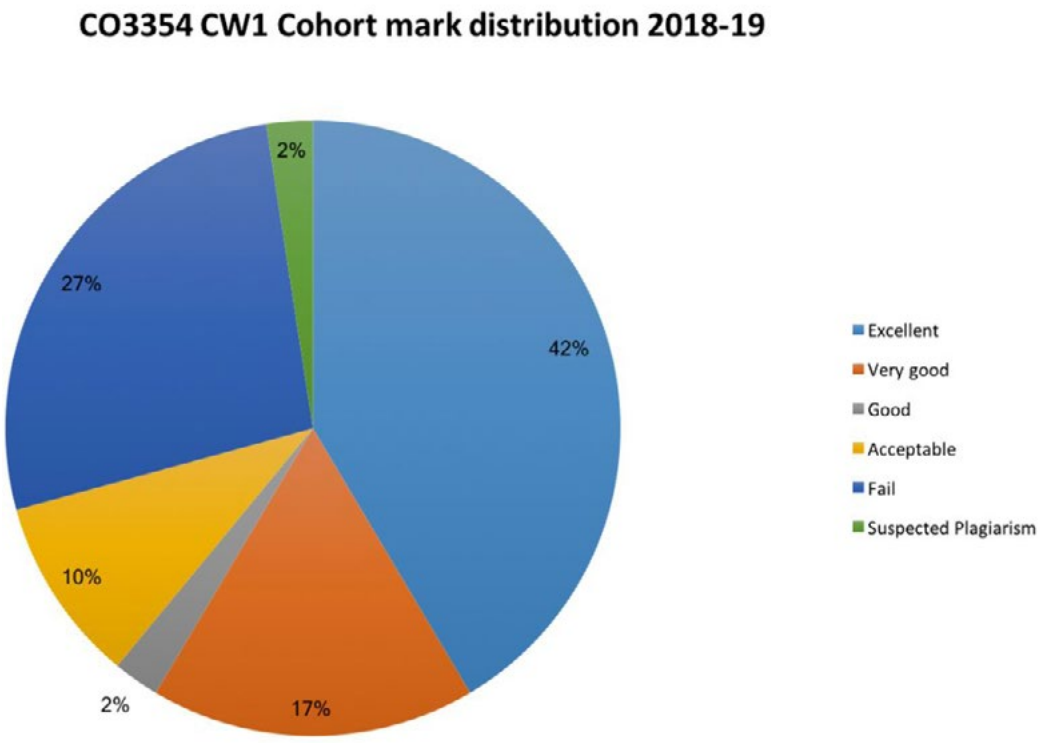
> *It is important that your submitted coursework assignment is your own individual work and, for the most part, written in your own words. You must provide appropriate in-text citation for both paraphrase and quotation, with a detailed reference section at the end of your assignment (this should not 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).*

This coursework assignment was intended to reinforce students' understanding of selected parts of the course content, including practical exposure to Wordnet.

Students were told that their PDF should '*include Appendices with any Python code you have written and the results of running your code*' – not everyone did this.

Some students simply presented their results as screenshots without explanation or commentary. It is important to explain the significance of what the examiners are looking at, in part to satisfy them that results are fully understood.

See the 2018–2019 CW1 cohort mark distribution below:



Comments on specific questions

Question 1

a. This question asked how we might assign unfamiliar words to a particular grammatical category, in this case **nonsense** words from Lewis Carroll's *Jabberwocky*. Obviously meaning does not come into play (though some words are glossed in the text) but sentence position and form (e.g. plural endings) give some clues. Some students only talked about the morphological form (e.g. suffix analysis), whereas sentence position is also critical. Students should say what it is about the sentence position that influences their answer. Answers must **explain** how categories are determined by the reader, not just list the different categories. Most answers were assessed as good, very good or excellent, though a few either showed weak understanding or were not attempted.

b. Marks were awarded for correctly running the tagger on the whole poem, for including results in the body of the submission, and for good selection and discussion of examples. Marks were deducted if the Penn tagset was not used or students did not show their results. Unfortunately, it turns out that Project Gutenberg is blocked in Germany following a copyright dispute, but fortunately the text is available from several other sites. Most students got good marks for this sub-question.

c. For this sub-question, students were essentially required to encode the categories and phrasal rules from the IGE grammar, using a similar rule format to the subject guide. Good answers to (i) would be shown to be consistent with the IGE framework, and the grammar rules should be optimally compact and general. Answers should give a single grammar covering all examples, not a separate one for each.

   On the whole, answers showed good general understanding of the use of formal grammar rules, though most did not reference the IGE scheme as required and syntax trees did not always match the proposed grammar rules. Answers to (ii) showed only partial understanding of issues of context-freeness and regular grammars: candidates tended to reproduce technical definitions without showing insight into how they applied to natural language grammar. About half the submissions were assessed as good, very good or excellent.

Question 2

Students were told that their results '*should be given in the body of your answer rather than relegated to appendices or additional files*' and that all Python code and results should be given in Appendices. Some students disregarded one or both of these instructions and thus lost marks.

a. This question was bookwork – answers could be found through careful reading of the subject guide and recommended texts, and most students obtained high marks. It is important when answering this kind of question to use your own words, and to avoid defining technical terms using other, unexplained technical terms such as 'superordinate'.

b. Students should have had little difficulty with this sub-question if they had read the subject guide and recommended texts carefully, as the requisite techniques are described and demonstrated in the NLTK book. Happily, the majority produced very good or excellent answers.

c. As with the previous question, the techniques needed to tackle this question are described and demonstrated in the NLTK book and the subject guide. Surprisingly, some students calculated two separate frequency distributions to obtain a CFD rather than using `nltk.ConditionalFreqDist()` as described in the subject guide. This

meant that they did not manage to 'tabulate' the results as required, and lost marks accordingly. Some eschewed NLTK functions altogether and calculated results from scratch, making extra work for themselves. Most answers were assessed as very good or excellent.

d. This question was meant in part to be exploratory, to encourage students to investigate the semantic relations captured by wordnet and assess how useful they might be in indicating the content and subject matter of a text.

   There are various different techniques for analysing the texts and manipulating WordNet, which may have resulted in slightly different outcomes. Good marks would be obtained by showing both technical competence and independent thinking, and for showing initiative in interpreting the question. Examiners paid attention to explanation, interpretation and discussion, and to presentation of code as well as to actual outputs.

   For parts (ii)–(iv), good answers would have discussed the extent to which shared vocabulary can indicate similar content, whether it is informative to use more general terms identified via Wordnet, and what difference the sample size makes. Good answers would also have included concrete references to particular words. Around half the submissions for this sub-question were assessed as good, very good or excellent.