Natural Language Engineering: Formative Assessment

Submission format: You should submit one file and that file must be an iPython notebook.

Due date: Submit your iPython notebook online on the module's Study Direct site before 4pm on Wednesday October 28th.

Marking and Feedback: You will be told your mark together and receive feedback via Sussex Direct on 17th November.

Weighting While this assessment contributes 0% of the mark for the module, it is very similar to the assessed coursework that is due in Week 10. Feedback on your submission will be useful when you are preparing your assessed coursework.

Overview

For this assessment you are asked to submit an iPython notebook that reports on what you discovered during your activities during the two labs on document classification: sessions 4 and 5.

During these lab sessions you were expected to have undertaken various experiments that enable you to address each of the following five questions:

1. How accurately do the various wordlist classifiers perform when compared to a Naïve Bayes classifier?
2. What are most informative features for a Naïve Bayes classifier and how do these compare with the words in your word lists?
3. For the methods that require training data, what is the impact on classifier accuracy when the amount of training data is varied?
4. What is the impact on accuracy of training a classifier with data in one domain, and testing the same classifier on data from a different domain?
5. What are the most effective feature extraction methods to use when designing a Naïve Bayes classifier?

Marking Criteria and Requirements

This coursework will be marked out of 100, based on the following criteria:

Overall quality of report: 15 marks

This concerns issues such as writing style, organisation of material, clarify of presentation, etc.

- Your report should be no more than 3000 words in length excluding the content of graphs and tables and any references. You must specify the length of your report. This is a strict limit.
- You **must** submit your report as an iPython notebook. **not as a Word document or a pdf**. If you do not submit an iPython notebook, your coursework will not be marked.
- You must use a formal writing style.
- All figures and tables should be clearly numbered, and have an appropriate caption.
- A good way to organise the report would be to have seven sections: an introductory section; a section for each of the five questions above; and a conclusions/summary section.
- Use subsections with meaningful headings.

Quality of code: 10 marks

This concerns issues such as coding style, appropriate use of comments, efficiency of algorithms.

- One of the advantages of using iPython notebooks is that when we are marking your submission we will be able to run your code in order to establish that it produces the outputs that you describe.
- The iPython notebook that you submit should include all of the Python code you have written or adapted, i.e. all code that you have used in your experiments except code in the NLTK and Sussex NLTK packages.

Quality of your experimental method: 15 marks

This concerns issues such as appropriate use of cross-validation, and division of labelled data for training and testing.

- Describe your methodology, i.e. how data was split, what size word lists were used, how the words were selected etc.
- It is also important to include appropriate baselines where appropriate.

Presentation of results: 10 marks

This concerns how well you present your empirical results.

- Graphs should be used when presenting your empirical findings.
- Be careful not to plot every experiment separately and then discuss comparisons without a graph that shows that comparison. In other words, when you are making a comparison between two or more results, you should display them together on the same graph so that the comparison can be seen directly.
- In cases where your empirical investigation is not complete, be explicit as to what you have not managed to achieve.

Technical understanding: 25 marks

This concerns how well you have explained each of the methods that you have used in your experiments.

- You should explain each of the methods in sufficient detail that a well-educated computer scientist who does not know about these particular Natural Language Processing methods will be able to understand what you have done.
- Do not just repeat details directly from the lecture notes or other sources. Your explanations must be expressed in your own words and be relate to the specific context of this report.

Quality of analysis: 25 marks (5 marks for each of the 5 questions)

This concerns the quality of your discussion as to what your empirical findings mean.

- Do not just present tables of numbers and graphs without any discussion. You need to put forward a reasonable explanation as to why you have observed the results that you have obtained. Describe an hypothesis as to what you expect your experimental results to be, with a justification as to why you make this prediction. After you have presented your results, discuss possible

explanations for any discrepancies between your predictions and what you have found. Where possible, describe additional experiments that you could have undertaken which could have verified your proposed explanations.

- Always back up claims with evidence. Be careful not to make bold conclusions from small-scale testing.