Natural Language Engineering: Assessed Coursework 1

Submission format: You should submit one file that should either be a Jupyter notebook or a zip file containing a Jupyter notebook and any other files (e.g., images or Python files) that you want to include in the notebook.

Due date: Your work should be submitted on the module's Canvas site before 4pm on Thursday 31st October. This is Thursday of week 5. The standard late penalties apply.

Return date: Marks and feedback will be provided on Canvas on Thursday November 21st for all submissions that are submitted by the due date.

Weighting This assessment contributes 25% of the mark for the module.

Overview

For this assignment you are asked to complete a Jupyter notebook ('a1.ipynb') which is provided with these guidelines. It is based on activities that you have already completed in labs during weeks 1-4 of the module. Any code you have developed during the labs can be submitted as part of your answers to the questions in the assignment. To score highly on this assignment you will need to demonstrate that you:

- understand the theory and your code;
- can write and document high quality python code;
- can develop code further to solve related problems;
- can carry out experiments and display results in a coherent way;
- can analyse and interpret results; and
- can draw conclusions and understand limitations of the technology.

For this report you should submit a single Jupyter notebook containing all of your answers to all of the questions in 'a1.ipynb'. You may import from standard libraries and the 'sussex_nltk' resources which you have been provided with. If you wish to import any other code, it must be included in a zip file with your Jupyter notebook. It **must** be possible for the assessors to run your Jupyter notebook.

Marking Criteria and Requirements

Your submission will be marked out of 50. Each question is worth 25 marks and all questions should be answered. Each question is broken down into 3 or 4 parts and the breakdown of marks between parts is specified in the notebook. General and question specific criteria are given below. Please read these guidelines carefully and ask if you have any questions.

General

- In order to avoid misconduct, you should not talk about these coursework questions with your peers. If you are not sure what a question is asking you to do or have any other questions, please ask me or one of the Teaching Assistants.
- Your report should be no more than 2000 words in length excluding code and the content of graphs, tables and any references.
- You should specify the length of your report. 2000 is a strict limit.
- You should use a formal writing style.
- All graphs should have a title and have each axis clearly labelled.
- Textual answers should be included in Markdown cells.
- Do not add external text (e.g. code, output) as images.
- You should submit your notebook with the code having been run (i.e., with the output displayed rather than cleared)

Question 1: Naïve Bayes Classification:

25 marks available

This question is designed to test your understanding of relevancy classification, Naive Bayes classification and the evaluation of classification models.

- There are 3 parts to this question worth 10, 5 and 10 marks respectively.
- In all parts, marks will be awarded for the quality of your written explanations.

  - Written explanations must be given in Markdown cells (not code cells).
  - Your explanations must refer to your unique set of examples generated by entering your candidate number at the top of the notebook. This must be your own candidate number.
  - Your explanations should define technical terms and include examples of calculations taken from your sample sets.
  - You may write or use code to calculate probabilities for your sample sets. However, you will not be marked for the quality of this code. Further, code on its own does not count as explanation.

Question 2: Training Data for Sentiment Analysis:

25 marks available

This question is designed to test your ability to investigate how the quality and quantity of training data affect the performance of a supervised learning method such as Naïve Bayes.

- There are 3 parts to this question worth 8, 8 and 9 marks respectively.
- In all parts, marks are available for quality of experimental design, quality and correctness of code, quality of presentation of results and quality of discussion of results and conclusions.
  - Your answers should include textual descriptions (in Markdown cells) with details of your experimental design and discussion of your results. Even well-commented code is no substitute for a proper description outlining what you have done and why.
  - Your training and testing sets will be unique to you, since your candidate number is used as a random seed. Therefore you may have different results and conclusions to other candidates. This will be accounted for when assessing your submission.