Applied Natural Language Processing: Assessed Coursework

Submission format: You should submit one file that should either be a Python notebook or a zip file containing a Python 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 26th November. This is Thursday of week 9. The standard late penalties apply.

Return date: Marks and feedback will be provided on Canvas on Thursday December 17th for all submissions that are submitted by the due date.

Weighting This assessment contributes 40% of the mark for the module.

Overview

For this assignment you are asked to complete a python notebook ('ANLPassignment.ipynb') which is provided with these guidelines. It is based on activities that you have already completed in labs during weeks 1-7 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 Python notebook containing all of your answers to all of the questions in 'ANLPassignment.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 notebook. It **must** be possible for the assessors to run your Python notebook.

Marking Criteria and Requirements

Your submission will be marked out of 100. The assignment comprises 2 separate (completely independent) questions, both questions should be answered and the breakdown of marks within questions is specified here and in the notebook. General and question specific criteria are given below. Please read these guidelines carefully and ask if you have any questions.

General: The following general guidelines must be followed when completing your assignment.

- 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 a Teaching Assistant.
- Your report should be no more than 4000 words in length excluding code and the content of graphs, tables and any references.
- You should specify the length of your report. 4000 is a strict limit.
- You should use a formal writing style.
- All graphs should have a title and have each axis clearly labelled.
- In all parts, marks will be awarded for the quality of your written answers as well as your code.
- Written / textual answers **MUST** be included in Markdown cells. Otherwise, you may score 0 for these answers.
- Code on its own does not count as an explanation or a discussion. Nor do code comments. Code should be commented but explanation and discussion MUST be given as text in Markdown cells (see previous point!).
- Do not add external text (e.g. code, output) as images.
- Your code must be applied to and your explanations must refer to the unique set of examples generated by entering your candidate number at the top of the notebook. This must be your own candidate number. Otherwise you may score 0.
- You should submit your notebook with the code having been run (i.e., with the output displayed rather than cleared)
- It **must** be possible for the assessors to run your Python notebook.

Question 1: Books vs DVDs 50 marks available

In this question, you will be investigating NLP methods for distinguishing reviews written about books from reviews written about DVDs.

Part a: **10 marks available**

Use your training data to find a) the top 20 words which occur more frequently in book reviews than in dvd reviews b) the top 20 words which occur more frequently in dvd reviews than book reviews. Discuss what pre-processing techniques you have applied (or not applied) in answering this question, and why. [15%]

The following breakdown of marks will be applied

- Clear and effective use of code in order to correctly find the top 20 words which occur more frequently in book reviews than in dvd reviews [3 marks]
- Clear and effective use of code in order to correctly finding the top 20 words which occur more frequently in dvd reviews than in book reviews [3 marks]

– Selection and justification of pre-processing techniques chosen / not chosen [4 marks]

Part b: **15 marks available**

Design, build and test a word list classifier to classify reviews as being from the book domain or from the dvd domain. Make sure you discuss 1) how you decide the lengths and contents of the word lists and ii) accuracy, precision and recall of your final classifier.[25%]

The following breakdown of marks will be applied

– Description of design of word list classifier [3 marks]
– Clear and effective use of code to build classifier [3 marks]
– Consideration of lengths and contents of word lists [3 marks]
– Testing including calculation of accuracy, precision and recall [3 marks]
– Discussion of results [3 marks]

Part c: **10 marks available**

Compare the performance of your word list classifier with a Naive Bayes classifier (e.g., from NLTK). Make sure you discuss the results. [15%]

The following breakdown of marks will be applied

– Clear and effective use of code to apply a NB classifier to the data [3 marks]
– Calculation of evaluation metrics for NB classifier [4 marks]
– Discussion of results [3 marks]

Part d: **15 marks available**

Design and carry out an experiment into the impact of the amount of training data on each of these classifiers. Make sure you describe design decisions in your experiment, include a graph of your results and discuss your conclusions. [25%]

The following breakdown of marks will be applied

– Description of design of experiment [3 marks]
– Clear and effective use of code to investigate effect of amount of training data on 1 of the classifiers [3 marks]
– Clear and effective use of code to investigate effect of amount of training data on 2nd classifier [3 marks]
– Presentation of results [3 marks]
– Discussion of conclusions [3 marks]

Question 2: Distributional Semantics 50 marks available

In this question, you will be investigating the *distributional hypothesis*: **words which appear in similar contexts tend to have similar meanings**.

Part a: **5 marks available**

Run generate_features(sentences[:5]). With reference to the code and the specific examples, explain how the output was generated.

The following breakdown of marks will be applied

- Correct general explanation [1 marks]
- Correct explanation which refers to examples in the output [2 marks]
- Correct explanation which refers to steps in the code [2 marks]

Part b: **5 marks available**

Write code and find the 1000 most frequently occurring words that are in your sample; AND have at least one noun sense according to WordNet

The following breakdown of marks will be applied

- Clear and effective use of code to find most frequently occurring words in sample [2 marks]
- Clear and effective use of code to identify words with at least one noun sense in WordNet [1 marks]
- Clear and effective use of code to combine the conditions and display the required words [2 marks]

Part c: **15 marks available**

Consider the code above which outputs the path similarity score, the Resnik similarity score and the Lin similarity score for a pair of concepts in WordNet. Answer the following questions

The following breakdown of marks will be applied

Part i: Clear explanation of each of the similarity scores and what the number calculated means [6 marks]

Part ii: Clear and effective use of code to find the semantic similarity of a pair of words [1 marks]

Part ii: Clear and effective use of code to find semantic similarity with a parameter to specify the measure of semantic similarity [1 mark]

Part ii: Explanation and justification of the strategy used for words which have multiple senses [2 marks]

Part ii: Clear and effective use of code to find semantic similarity of every pair of words [2 marks]

Part ii: Justification of choice of semantic similarity measure [1 mark]

Part iii: Clear and effective use of code to identify the 10 most similar words to the most frequent word in the corpus [2 marks]

Part d: **10 marks available**

The construction and use of distributional vector representations to find similar words

The following breakdown of marks will be applied

Part i: Clear and effective use of code to construct distributional vector representations of words in the corpus with a parameter to specify context size. [2 marks]

Part i: Clear and correct explanation of how you calculate the value of association between each word and each context feature [3 marks]

Part ii: Correct use of code to construct representations of the 1000 words identified in Q2 with a window size of 1 [2 marks]

Part iii: Clear and correct use of code and representations to find the 10 words which are distributionally most similar to the most frequent word in the corpus. [3 marks]

Part e: **15 marks available**

Plan and carry out an investigation into the correlation between semantic similarity according to WordNet and distributional similarity with different context window sizes. You should make sure that you include a graph of how correlation varies with context window size and that you discuss your results.

The following breakdown of marks will be applied

- Description of plan of how to carry out the investigation [3 marks]
- Clear and effective use of code to carry out the investigation [2 marks]
- Correct calculation of correlation between WordNet similarity and distributional similarity for at least one context window size [2 marks]
- Correct calculation of correlation between WordNet similarity and distributional similarity for different window sizes [2 marks]
- Presentation of results [3 marks]
- Discussion of results / conclusions [3 marks]