Assignment 2 - FAQ

1. I don't see any Jupyter notebook provided, so I assume we are to create one from scratch. Is this correct?

    - Please create a brand new jupyter notebook to house your submission. You are also free to define your own class and function names to implement your solutions.

2. How should we behave if we want to use specific 3rd party libraries? Are there limitations beyond those mentioned (Scikit learn and NLTK?), or can we use them freely? In the second case, is there something you suggest we do while preparing the notebook?

    - In the second graded assignment (car reviews), you are allowed to use third-party libraries to build your own original solution. That of course does not mean that you may download and submit a complete third-party solution to the task

3. Does the code produce some output to demonstrate that a vector has been created for each review, where each element in the vector represents EITHER a binary variable indicating the presence of a word/stem in a review OR the number of times that a word (or word stem) appears? Note that the output does not need to show the vector for all reviews, this only needs to contain a small sample of reviews.

    - The marking criterion just means that you need to clearly demonstrate in your submission that vectorisation of the textual data has taken place. You demonstrate that by printing some examples of such vectors. Presenting the vocabulary (or a relevant part of the vocabulary) may make the printed vector outputs clearer

to interpret. This criterion specifies that you can implement either lemmatisation or stemming and use either binary variables or occurrence counts in the vectorisation to prepare the input for Naive Bayes training.

However, even if you choose lemmatisation for input vectorisation, you still need to implement and demonstrate the stemming operation on the input data for the preceding criterion: "Does the submission demonstrate that words with the same stem have been appropriately recognised and treated as variations of the stem? This should be demonstrated for at least 3 different stems." The reason for it is that we would like you to get a feel for what the stemming operation does even if you opt for the alternative in the actual model training.

4. In the assessment criteria, we're asked to ensure punctuation is excluded from our sentiment classifier. As far as I can see, the reviews provided in car_reviews.csv don't contain any punctuation at all. Just checking this is as expected.

   - The car-reviews dataset you have been given does indeed seem to be preprocessed through punctuation removal. You are still asked to implement punctuation removal in your submission though because, if you ever decide to test your trained model on an unpreprocessed sentence (not from the dataset), that sentence would go through the same preprocessing pipeline you implement for the sentences of the dataset. So, you should keep it generic

5. I plan on submitting the Jupyter Notebook already executed with all output ready, but if the marker wants to rerun it, it will take some time. Is it ok? If not, what should I do?

   - Yes, definitely submit your notebook pre-executed as you say to support your discussion. Then write a visible comment to the marker warning them that re-running the code to generate the output from scratch would take about 6 minutes to complete.

6. I am installing a number of packages using the pip command line using Terminal on my Mac. When we submit do we have to make the code install all the packages we use by using system calls?

- Please don't include any package installation commands as part of your code and let the marker manage the python installation/virtual environment on their machine themselves. Instead, please provide a list of all neccessary libraries and their versions as part of your submission where it can be easily found. The marker will install any packages needed to run your code provided they are readily available via standard routes such as pip install

7. As assignment 2's specification mentions that we have to show that we handle punctuation, does this mean we still need to write code for it, even if it won't do anything?

    - Yes, implement punctuation removal in your submission. It is essential to maintain a generic preprocessing pipeline for the sentences in your dataset so that if you want to test your trained model on an **unprocessed sentence**, it goes through the same preprocessing pipeline as the dataset sentences. This will help ensure the accuracy of your model's predictions

8. Does this mean, word variations with the same stem where they share the same meaning, e.g., 'play', 'playing', 'played'? It does not refer to whether our stemmer avoids overstemming, such as 'skies' and 'skis' -¿ 'ski', where the words have different meanings?

    - To ensure that you receive full marks for this section, it would be beneficial if you included a segment in your notebook that demonstrates how stemming functions. It's worth noting that different stemmers will produce different results. For example, if you use the LancasterStemmer(), it will stem "reality", "really", and "realistically" all to "real". This is what you might expect from a linguistic perspective. On the other hand, if you use the PorterStemmer(), it will identify three stems: "realiti", "realli", and "realist". Ultimately, which stemmer you select depends on your specific use case.

9. The instructions for the assignment say "For Task 2, you are asked to identify, research and implement a way to improve on your solution to Task 1." (my bolding). Does this mean that you are looking for exactly one improvement to the original algorithm, or am I interpreting this too tightly and we can improve it in several different areas?

- For task 2, the requirement is that you implement an algorithm that can enhance the performance of your existing algorithm in task 1, e.g., accuracy. However, you can implement more features/improvements in the algorithm for task 2.

10. The marking scheme for Task 1 includes clearly demonstrating in the code things like "Does the submission demonstrate that words and punctuation, which are unlikely to affect sentiment, have been excluded from the sentiment classifier?". Can we leave this sort of demonstration out of Task 2 or should we include it there too? It does clutter up the code and the output somewhat.

    - No, you do not need to include the same features again in task 2, and you can use sklearn feature to generate vectors for train and test review. Once you successfully create two or three reviews, you can train multinominal naive bayes model and get a confusion matrix graph. Please write concisely the improvement that you made in task 2.

11. Just for the avoidance of doubt, can we use any function in scikit learn, including the methods for generating and manipulating feature matrices and confusion matrices etc?

    - Yes, you can use any function that can improve the performance of the sentiment classifier in task 2.

12. I understand in the coursework we need to tokenize and remove redundant formatting ect.. as part of our preprocessing. I suppose in our case, an outlier may be a review which says something like "How Bad,Bad,Bad,Bad Bad,Bad Bad,Bad my life was before this car!" and, how the sentiment is positive, but the underlying structure looks like it is negative...?

    - The car-reviews dataset has already been preprocessed, so no outlier detection

13. is there any word counts for the markdown we should be concerned about?

- There is no specific limit on the amount of text you can write. Please feel free to include any necessary information.

14. Do you want us to just put the code that would train the model and write about the results?

    - Please upload a single Jupyter file containing the results of the tasks to the system.