Instructions:

1. Kindly refrain from copying someone else's code or seeking help to complete this test.

2. The exam is very similar to your homework assignment.

3. You must submit your exam in an IPython notebook by 11:59 p.m. on Sunday, July 25.

4. The dataset for this exam is called "reviews.xlsx" and is available on Canvas under datasets. The file contains online clothing reviews.

Here are the things you have to do. Just follow these instructions and create an IPython notebook.

**Part A: Text Classification - Binary Classification**

1. Use Pandas to read the reviews.xlsx file.

2. Display the counts of 1,2,3,4, and 5 in the 'Rating' column.

3. Since this is a binary classification problem, all 4s and 5s will be treated as 1 (i.e., high rating) and 1 and 2 would be 0 (i.e., a low rating). Note that we are disregarding 3s. Create a column called "Target" that contains 1s (corresponding to ratings of 4 & 5) and 0s (for ratings of 1 and 2).

4. Display the proportion of 1s and 0s in the Target column.

5. Write a function to preprocess your text. It should do the following:

a) Convert text to lowercase;

b) remove punctuation and digits;

c) remove stopwords;

d) lemmatize the text and include only adjectives, adverbs, nouns, and verbs. You must also ensure that only words having 3 or more characters are included. Your function should take some text and return the preprocessed version of it.

6. Create a column called 'Clean Text' that is obtained by preprocessing the 'Review Text' column.

7. Use the 'Clean Text' column to create the corpus. Create labels using the 'Target' column.

8. Use train\_test\_split to create your training and testing sets.

8A. Create a vectorizer using TfidfVectorizer. Use the vectorizer to transform x\_train and x\_test (this will give you the document-term matrices).

9. You will build two predictive models, one that uses Naive Bayes and the other that uses LogisticRegression.

10. Display the confusion matrix and classification report. Note that you may have to try out a few things to ensure that the performance is reasonable. When you are satisfied with the performance of your model(s), you may move to the next step.

**Part B: Text Explainer**

1. Use LIME and ELI5 to see if you can explain what your model is doing. Test it on a couple of cases.

**Part C: Topic Modeling**

1. Perform Latent Dirichlet Allocation on the entire corpus. You may assume that there are 10 latent topics.

2. ***Display the top 3 reviews for your first topic***. Note that this will not be the same for all of you. **HINT: Use the probabilities from the document-topic matrix to determine the top three reviews for the first topic.**

**PLEASE NOTE THAT I WILL NOT PROVIDE ANY HELP OR ANSWER ANY QUESTIONS ABOUT THE EXAM.** Ensure that all these steps are included in your IPython notebook. Feel free to improvise to increase the accuracy of your model(s).