ASSIGNMENT LEVEL: 2

Predict if a particular sentence in an article should be included in the summary of the article or not (Incorporates NLP Along with ML)

Given a dataset comprising of articles and its summary, create a dataframe that enlists the sentences for each document, a list of relevant features and whether that sentence is present in the summary or not.

A sample of relevant features can be found here ->https://www.researchgate.net/publication/220974615_Automatic_Text_Summarization_Using_a_Machine_Learning_Approach

Do not confine your list of features to the ones mentioned above

Feature Engineering is important here.
Perform EDA and narrow down to a list of relevant features.

Apply Binary Classification techniques and Display the accuracy of predicting a sentence's occurence in the summary.

Dataset-:https://www.kaggle.com/pariza/bbc-news-summary
Duration- 1 day (8 hours)
Difficulty Level: Medium (Feature Engineering)

OUTPUT FORMAT:

1. Complete Code(Python Notebook)(The notebook should contain 2,3,4,6,7 automatically)
2. The master dataset created after parsing through each file and tokenizing the sentences, listing the relevant features and the predictor variable(i.e is the sentence a part of the summary or not)
3. List of Features engineered
4. Relevant Exploratory Data Analysis conducted to be recorded in the notebook
5. Training accuracy, Testing accuracy and Confusion Matrix
6. Excel Sheet of the Test Set with Doc details, Sentence ,Predicted Output and Output(1 or 0)
7. Function that allows a user to input a text document and returns a dataframe that contains all the sentences in the document and whether the make the summary or not.

Thus 1 notebook and 1 excel sheet.