**Overall Information**

* You must submit a report with code and interpretation within the same document. This report should be created within markdown or jupyter notebooks.

**Set up Text**

* Do preprocessing on your text to prepare it for the final machine learning model.
* Create a “label” column by using regular expressions to look for the ten most popular nouns in your text. This column should be true/false, 0/1, yes/no for includes the word or not.
* Delete the word from your sentence so we do not create a perfect prediction using re.sub.
* What items do you think will be important in your preprocessing to clean for your specific text? Write a summary of the different considerations you used for determining the cleaning for classification.

**Create Feature Extractions**

* Examine the balance of the data options.
* Create a balanced dataset and an imbalanced dataset (likely the default).
* Create **two** different encoding options for your text.
* For each of these two different encoding options: create **two** different preprocessing steps that you can examine to determine what may influence your final results.
* Therefore, you should end with **eight** (balance/imbalance by two preprocessing by two extractions)different feature extractions.
* Summarize your feature extractions. Which options did you use? Why do you think these are the best choices?