**Overall Information**

* We will complete this assignment **together** in class over the course of two weeks. You should submit the final assignment here.
* You will be scored as complete/incomplete for this assignment.
* You must submit a report with code and interpretation within the same document. This report should be created within markdown or jupyter notebooks.

**Find Text**

* As a class, we will find a text source to analyze. This text source usually will consist of a webpage or other dataset to examine and clean.
* Import the text into your report.
* If the text is one big long string, first break into sentence segments and store it in a Pandas DataFrame.

**Fix Errors**

* Examine the text for errors or problems by looking at the text.
* Use the “impurity” function from class to examine the text for potential issues.
* Remove the noise with the regex function.
* Re-examine the impurity to determine if the data has been mostly cleaned.
* Normalize the rest of the text by using textacy.
* Examine spelling errors in at least one row of the dataset.

**Pre-Processing**

* Using spacy and textacy, pre-process the text to end up with a list of tokenized lists.
* Create a frequency table of each of the tokens returned in this output. Below is some example code to get us started.

from itertools import chain

from collections import Counter

list\_of\_lists = OUROUTPUT

Counter(chain.from\_iterable(list\_of\_lists))

**Summary (on your own)**

* Write a paragraph explaining the process of cleaning data for an NLP pipeline. You should explain the errors you found in the dataset and how you fixed them. Explain the information that is gathered by using spacy and textacy and the final output. What did you learn from your frequency table? What is the text document about?