**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.

**Part of Speech Tagging**

* Tag your data with spacy’s part of speech tagger.
* Convert this data into a Pandas DataFrame.
* Use the dataframe to calculate the most common parts of speech.
* Use the dataframe to calculate if words are considered more than one part of speech (crosstabs or groupby).
* What is the most common part of speech?
* Do you see words that are multiple parts of speech?

**KPE**

* Use textacy to find the key phrases in your text.
* Use summa to find the key phrases in your text.
* What differences do you see in their outputs?
* Using textacy utilities, combine like key phrases.
* Do the outputs make sense given your text?

**NER**

* Use spacy to extract named entities.
* Create a summary of your named entities.
* Apply Snorkel to your data to show any relationship between names.
* What might you do to improve the default NER extraction?

**Knowledge Graphs**

* Based on the chosen text, add entities to a default spacy model.
* Add a norm\_entity, merge\_entity, and init\_coref pipelines.
* Update and add the alias lookup if necessary for the data.
* Add the name resolver pipeline.
* Create a co-occurrence graph of the entities linked together in your text.