DS 2500: Intermediate Programming with Data
Prof. Rachlin
Northeastern University

# Presidential Inaugural Addresses

## Problem Description:

In this assignment we'll analyze the political speeches for U.S. presidents, past and present. When Donald Trump gave his Inaugural Address in January 2017, it was widely regarded as being very dark in tone and very much unlike inaugural addresses from past presidents. Can data science reveal whether this is really true? And can we approach the problem scientifically without introducing biases reflecting our own personal political beliefs?

To carry out this analysis we will download data from at least six presidents along with sets of words with positive and negative associations. We'll look to quantify and visualize how the inaugural addresses of different presidents compared and perhaps evolved over time. What do these speeches tell us, if anything, about the mood of the country at different periods in our history?

This assignment lays out some minimal goals and guidelines for your analysis, but you should not feel constrained to carry out this work exactly as suggested below. Be creative. Outstanding analyses may earn extra credit and a chance to showcase your results to the class. Consider this homework statement a draft. Imagine, instead, that we are a collective consulting firm. Over the next couple of weeks our collective assignment is to discover what we can about this data.

## Some Recommended Data Sources:

The American Presidency Project: https://www.presidency.ucsb.edu/documents

Twitter sentiment analysis tutorial: https://github.com/jeffreybreen/twitter-sentiment-analysis-tutorial-201107/tree/master/data/opinion-lexicon-English

AFINN: http://www2.imm.dtu.dk/pubdb/pubs/6010-full.html

Another source of words with sentiment score:
https://github.com/hitesh915/sentimentstrength/blob/master/wordwithStrength.txt

## Broad Goals:

1. Include the Inaugural speeches of at least these six presidents:

    a. Donald Trump (2017)
    b. Barack Obama (2013)
    c. George W. Bush (2001)
    d. Ronald Reagan (1981)
    e. John F. Kennedy (1961)
    f. Franklin D. Roosevelt (1941)

2. Clean up your data by converting the speeches to lists of lower-case words, eliminating all punctuation.

3. Generate an array of word clouds, one for each speech.

4. Explore ways of scoring speeches in terms of how positive and negative the speech was and plot different presidents in a labeled scatter plot diagram. For example, the positive score might be the average number of positive words used per sentence, or per 100 words. We might also measure their overall positivity / negativity by combining the scores in some way. Are differences correlated with the political party of the speaker?

5. Identify the most common non-trivial words for each speech and determine whether the most commonly used words evolved over time. For example, it would be interesting to identify, for each speech, the top 100 words of length > n (for some n) and count the number of overlapping words for each pair of presidents, plotting the results as a heat map using Seaborn.

6. Develop other comparative measures of speech analysis, such as:

    a. Average sentence length
    b. Average number of words of length > n per sentence (or per 100 words)
    c. Number of unique words per 1000 words

    and generate some comparative figures.

## What to submit:

7. Your code in the form of a python file (.py) or a Jupyter notebook file (.ipynb) with embedded documentation, commentary, analysis, and visualizations.

8. A separate .PDF writeup if you aren't using Jupyter notebooks.