**Modeling Assignment – Executive Summary**

**Pooja Suri**

The model aims to create a relationship between congressmen’s speeches and their ideologies. Our aim is to parse and analyze the n-grams of congressmen’s speeches to see the top 100 bigrams and trigrams for both Republicans and Democrats.

Steps:

1. Downloaded Congressional speech data via Sunlight Foundation, using Gaurav Sood's script capitol speech fetch.py. Downloaded data from the 113th Congress.
2. Cleaned up data in excel in order to remove blanks and raw speech material not attributed to any party member.
3. Pre-processed Congressional speech data using own script using classroom slides. Retained stems of all words in the speech.
4. Downloaded DW-NOMINATE csv files for House and Senate; deleted all sessions that were not 113, combined senate and house files into one csv file and generated a concatenated term for State and Name.
5. Wrote a script to merge pre-processed Congressional speech data with DW-NOMINATE scores on State and name of member.
6. Extracted n-grams from Congressional speech data using attached script. This step extracts bigrams and trigrams from Republican and Democratic speakers.
7. Reduce terms with chi-square by creating a function to find chi-square values for each n-gram, testing the hypothesis that Democrats and Republicans are equally as likely to use that term.
8. Retaining a list of top 500 of the 500 most predictive bigrams and trigrams for Republicans and Democrats by sorting on chi-square value.
9. Built a ridge model about the relationship between congressional words and their ideology.

Results:

1. Top 500 bigrams and trigrams were generated and can be found listed in the repository.
2. Ridge Model: As shown in the last part of R script, top bigrams are checked in each congressman’s speech to see whether they have involved certain topic in their speeches. Based on this, a ridge model is created where the dependent variable is the ideology score and the independent variables are whether people’s speeches include the top 500 bigrams. Cross-validation is adopted to tune the model.