Stuff to do for mid-quarter presentation:

- **get data**
- **make similarity model**
- **allow user to enter name into model**

Machine Learn Project Proposal

Title and team members:
Identifying Language Patterns betweens Politician by Policy Area
Kyle Magida, Christine Cook

Description:
We will obtain public speeches from a variety of politicians (possibly including some presidential candidates) with the intention of identifying which politicians use similar rhetoric in talking about the same issue. We will compare politicians across policy areas to identify similarities in vocabulary and lexicographic diversity between candidates.

Data:
As described above, we will be using the public speeches and policy paper from politicians as our data source. These can be found easily as speeches are transcribed. Depending on the amount of data we require, we may need to scrape websites to pull all of this information.

Software, packages:
NLTK; to parse and clean texts, sklearn; to train and test models and evaluate classifiers, regular expressions.

Plan of Action:
We will choose the politicians we wish to study and the policy areas we're interested in. We will need to figure out how to abstract policy positions from texts for each politician, or perhaps just to classify each sentence in terms of the policy area which it refers to. This will be the largest challenge and I think this is a good indicator of progress for our Mid-Quarter presentation. We will then work on identifying patterns of speech for each politician on each issue and compare them in some way, either by a clustering algorithm or just a visual matrix depending on the number of candidates and issues we choose.

Mid-Quarter Presentation:
Demonstrate how to classify sentences in a speech by policy area.

Final Presentation:
Share our results! We'll show our comparisons between politicians and identify pairs or groups that have similar positions on different policy issues.

Evaluation Methods:

First we will need to decide which politicians we want to compare. To do this kind of analysis, it will be useful to have very large datasets, so we may want to limit our work to, say, the current presidential candidates. Once we've decided this, we will need to figure out how to collect the text. If we want to use text from debates, that will be easy to find in one or two locations, but speeches outside of debates will be harder to collect in bulk.

Once we have all the texts, we will need to identify key words for a number of different policy areas (e.g. immigration, tax reform, education) and break the speeches into different subject areas. With this broken down, we can compare the "important" words surround these policy issues between candidates (e.g. use of the word 'illegal' in the immigration sphere, 'cuts' in the tax reform sphere). We will measure the similarity of the candidates across different policy areas using different measures of lexicographic distance.

Work Division:

We will divide the work into two sections, one collecting the data and attempting to classify it by policy and the second as pairing and grouping the politicians based on similar policy positions. We will both work on each section but Kyle will take the lead in planning and researching the first task while Christine will focus on the second. We expect that we will write code in each section approximately equally but the general direction will be determined by the point person for that section.

We will plan on meeting at least weekly outside of class to check in on work and to collaborate electronically otherwise. We will set up a shared GitHub repository to track our changes and ensure that we are able to work on the most up-to-date code.