Political Speech Generator

Team Members:
Alex Lassalle
Tim Contois
Josh Pikovsky

Progress Report:

Our initial proposal was to create a functional speech generator that could take an input of political party and topic, and output a political speech that was readable and relevant to the subject matter. For the three week benchmark, we aimed to find useful datasets and tools for language generation and tagging, and then implement our findings in a simple text generator. The feedback for our proposal encouraged us to pare down our ultimate goals to creating a text generator that uses an n-gram model combined with some weighting of words given a specific topic to generate text. Based on the feedback, we have constructed a basic Markov text generator that trains on a database of President Obama's speeches, and constructs a sample text based on an inputted Markov chain length.

For our dataset, we have acquired a database of over 250 speeches by President Obama. The file is approximately 2.7 MB in size, and contains over 460,500 words and 25,000 sentences. This data comes from AmericanRhetoric.com, a site dedicated to compiling speeches by famous rhetoricians. The catalogue of Obama's speeches contains texts ranging from 2004 to 2015, and are typically between 2-3 pages in length. These are not tagged, but per the proposal feedback we have simplified our goal down to starting with just a simple generator from the position of one side. We excluded interviews and press conferences because they often involved Q&A, which would be tedious to sift through.

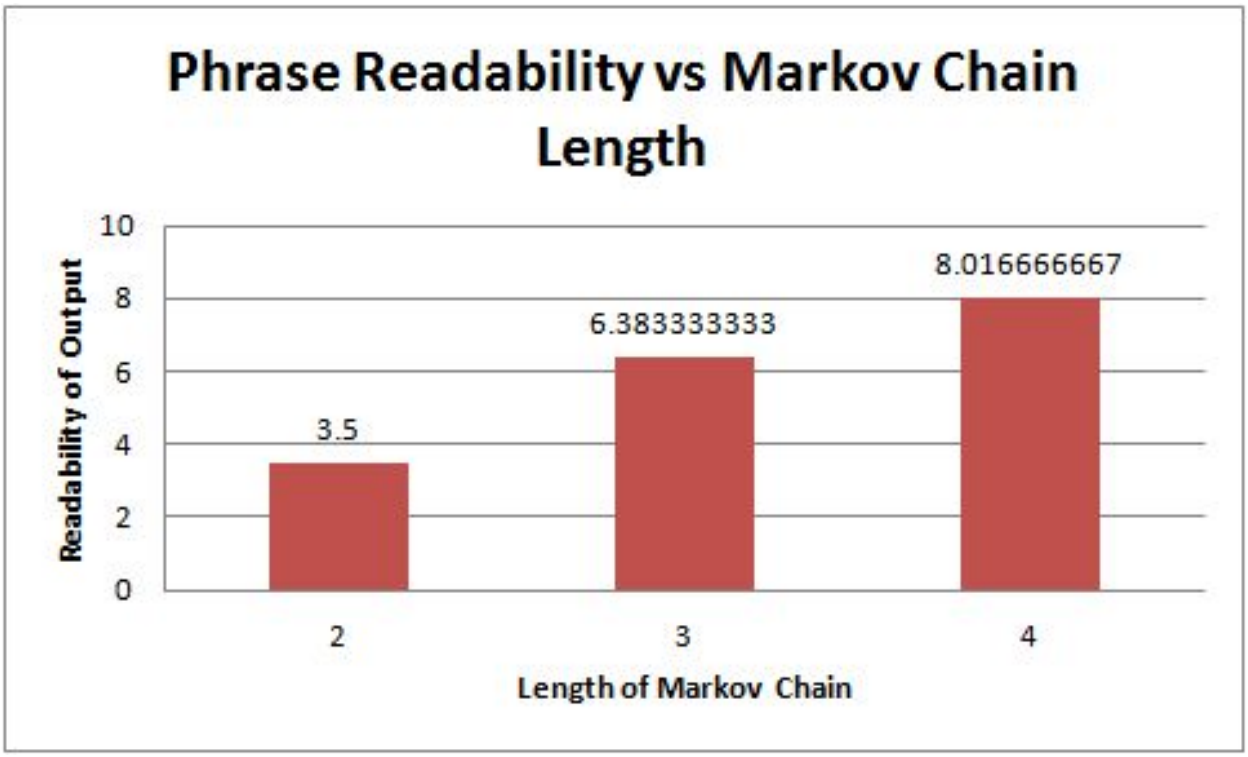
At this point in the semester, we have found and implemented several tools and acquired a dataset that allows us to generate very simple sentences trained on the language used in Obama's speeches. The sentences don't quite make sense with the model we have now, but one of our goals is to have speeches choose words based on their association with a certain topic. This will require us to find a way to annotate political speeches based on their subject matter.

Using the TextBlob text analysis tool, we will find the most prevalent noun-phrases in a speech, and use that to identify the primary topic of the speech. For example, one speech had the noun phrase ‘budget plan’ appearing very frequently. Given information like this, our generator will have a better sense of which words to weigh more heavily when building the speech. If the results of using the noun-phrase extractor are not great, we may have to annotate some speeches by hand.

Our coding progress began with finding a sample Markov text generator that outputted simple, pseudo-random sentences. This finding served as our jumping off point, as it provided us with a sample model on which we could build a more complex system. The first change we made was altering the initial word that the model used. The sample code began sentences at random points in the dataset, which yielded outputs that didn’t quite make sense, so we made changes to the code that would allow it to begin sentences at words that were likely to be the start of a sentence. By checking if a word came after a word with a period and had a capital first letter, we gave our model the ability to select words at the start of a sentence. To do this, the model just randomly selects a start word until the criteria is met. Another change we made was to the way the code handled sequence lengths. The starter code only ran sentences up to a certain length, and then ended the sequence regardless of its completion. This was changed so that it would continue choosing words until it came upon a word with a period, which would most likely symbolize the end of a sentence. After making a few modifications to create more clean output, we began experimenting with different length n-grams, and the text they produced.

Text Generator Example Output:

For our example output, we have compiled a variety of texts with varying Markov chain lengths. For the first set, we based predictions on two prior words, and then we continued up to four words. To evaluate our data, we individually graded the outputs on a 1-10 scale of readability, with 1 being gibberish and 10 being completely coherent. We calculated our average scores for the different n-gram lengths, which are included in the graph below.



To give a better sense of how the readability ratings correspond to the actual language, we have included a chart that shows sample outputs, and how we graded the texts.

Chain Length	Example Output	Josh's Rating	Tim's Rating	Alex's Rating
2	I came to do. And to do more than the universe expanding training to missile defense in with other definition of speech can respond quickly how we could not another country.	4	3	4
2	Because Congress returns so it is unreliable on the VA. And now, Jews from us. This was the end to make a free and research and to conduct nuclear program.	3.5	4	4
3	And Osama bin Laden, a man to lead again. Thanks to our own. That's my job. That's my job. That's not just in the halls of government.	6.5	6	5

3	We carry all that we would do. We can admit the intractability of depravation, and still strive for justice. We give thanks for that price. Every year, we freeze annual domestic spending, which represents a major commitment to comprehensive health care system.	6	8	7.5
4	I will not make that same mistake with health care. It was common ground, rooted in the cradle of civilization and a crucible of strife. For all the maps plastered across our TV screens today, and for their service and sacrifice.	9	6	8
4	John Crowder that will sustain the good people of all colors in the noble quest for freedom. Of course, King overcame in other ways as well. He took them camping and taught them to sail.	7.5	9	8

Looking at the results of the graph, it would appear as though a length four Markov chain produces the most readable result. The example confirms that while those texts are the most readable, they are also the least original, as they tend to reconstruct quotes from the text. For example, one output with length four Markov chain produced the sentence “You know, it's been 12 weeks now since my administration began. And I think it's important for members of Congress do, where -- we call it an "exchange," or you can face a growing challenge to its future.” This sentence is pretty readable, but the first line “You know, it's been 12 weeks now since my administration began” is a direct quote from an Obama speech. There are many examples of sentences being drawn directly from the text with a length four Markov chain.

Thus, the length three Markov chain produced the most readable output, while still retaining a degree of originality in the sentence structure and word choice. To create output with a more coherent topic, we will attempt to use a noun phrase extractor to bias our model towards certain combinations of words. Below is an example of the noun-phrase extractor applied to a chunk of the speeches.

```
er', u'brighter place', u'brazil', u'brazil', u'brazil', u'brazil', u'shows democracy delivers', u'braz
il', u'decades', u'cinelandia', u'brazil', u'political leaders', u'down', u'democratic aspirations', u'
brazilians', u'generations movement', u'own government', u'basic human rights', u'nations president', u
'dilma rousseff', u'ordinary people', u'extraordinary things', u'new world', u'new century', u'paul coe
lho', u'famous writers', u'muito', u'thank', u'god', u'thank', u'mr.', u'reid', u'mcconnell', u'pelosi'
, u'assistant leader', u'clyburn', u'rosa parks', u'america', u'rosa parks', u'formal seats', u'rightfu
l place', u'nations course', u'black caucus', u'childhood friend', u'parks', u'rosa', u'thats', u'alaba
ma', u'december', u'parks', u'front door', u'back door', u'winter evening', u'rosa parks', u'rosa parks
', u'little-known pastor', u'martin luther king', u'jr', u'montgomery', u'alabama', u'full day', u'sole
mn determination', u'god-given', u'eighty-five days', u'rosa parks', u'black', u'montgomery', u'entire
edifice', u'ancient walls', u'jericho', u'rosa parkss', u'activism didnt', u'criminal justice system',
u'local chapter', u'naacp', u'quiet leadership', u'civil rights movement', u'conyers', u'minds fasten',
 u'ms', u'parks', u'scripture', u'whether', u'simple lack', u'moral imagination', u'bus driver', u'way
things', u'entire neighborhoods', u'job loss', u'rosa parks', u'countless acts', u'anonymous courage',
u'fellow feeling', u'rosa parkss', u'singular act', u'dusty roads', u'montgomery', u'land truer', u'ros
a parks', u'may god', u'rosa parks', u'god', u'america', u'congress', u'jobs', u'new jobs', u'construct
ion workers', u'small business owner', u'america', u'jobs bill', u'congress', u'congress', u'jobs', u'n
ations debt', u'im', u'jobs bill', u'middle-class families', u'basic measure', u'economic security', u'
long run', u'massive debt weve', u'past decade', u'past decade', u'profligate spending', u'washington',
 u'tax cuts', u'record surplus', u'big pile', u'ious', u'dont act', u'childrens shoulders', u'dont act'
, u'medicare', u'washington', u'economic growth', u'long- term', u'deficit framework', u'april', u'hamp
er growth', u'small businesses', u'middle-class families', u'budget line-by-line', u'basic scientific r
esearch', u'road construction', u"n't balance", u'middle class', u'fair share', u'deficit reduction', u
'grand bargain', u'defense spending', u'everyone', u'special joint committee', u'congress', u'deficit r
eduction', u'im', u'specific proposals', u'ive', u'plan cuts $', u'new revenues', u'agricultural subsid
ies', u'large farms', u'modest adjustments', u'federal retirement programs', u'tax money', u'fannie mae
', u'freddie mac', u'financial firms', u'tax dollars', u'financial crisis', u'american people', u'addit
ional $', u'iraq', u'afghanistan', u'budget plan', u'republican', u'structural reforms', u'health care'
, u'medicare', u'medicaid', u'keep', u'health care law', u'long way', u'plan reduces', u'wasteful subsi
dies', u'erroneous payments', u'excessive health care costs', u'generic drugs', u'well', u'medicaid', u
'health care', u'such steps', u'medicare', u'medicaid', u'health care costs', u'im', u'medicare', u'vou
cher program', u'leaves seniors', u'insurance industry', u'health care', u'poor children', u'medicare',
 u'medicaid', u'fundamental commitment', u'social security', u'social security', u'separate track', u's
ocial security', u'government spending', u'modest adjustments', u'medicare', u'medicaid', u'fiscal prob
lems', u'such large budget deficits', u'fair share', u'john boehner', u'heres', u'additional revenues w
hatsoever', u'option relies', u'americas', u'technological edge', u'critical public assets', u'middle c
lass']
```

The noun phrase extractor had some interesting results in that it found many terms and word combinations that might register as buzzwords in the contemporary political vocabulary. Such terms as “financial crisis”, “excessive health care costs”, and “basic human rights” could help us create a more realistic sounding speech, and find a way to weigh certain words more heavily if they are strongly associated with certain topics of interest.

Conclusion:

Now that we can see what a simple Markov Model will create, we are looking to build a more complicated system that takes into account a topic, and generates more sensible outputs that are relevant to the given subject. We have narrowed our scope to training on speeches given by President Obama, and our current plan is to proceed with an n-gram model that selects words based on a Markov chain length of two or three. Our next step will be to implement a system by which we can have the model select words based on a given topic. We will accomplish this by using a noun-phrase extractor to weigh certain words more heavily based on association with the phrase, and failing that, we will annotate speeches for topic. If our n-gram model over words (i.e.

word1 and word2 generate word3) does not go as well as we would like, we will also look into having tag1 and tag2 generate tag3 and then emit a word based on the topic and the tag. We will evaluate our generated texts in a variety of ways. First, we must make sure our text generated is not just taking long excerpts from actual Obama speeches (as would potentially be the case with longer length n-grams). We will also do a simple coherence evaluation of the output on a scale from 1 to 10, like above. For our final evaluation, we plan to generate 50 speeches and also take 50 excerpts of similar length from real Obama speeches. We will then go through these speeches (which have been put together randomly) and give them a 1 or a 0 for whether we think it was an actual Obama speech or not.

Timeline:

11/13: Proposal Due

11/20: Collect all datasets necessary for topic association

Use noun-phrase extractor to associate words with topics.

Make necessary tags and annotations on datasets.

11/27: Use annotated datasets to construct simple sentences relevant to political topic

12/04: Begin evaluation of TimTextGen's outputs

12/8-12/10: Presentation

12/18: Final Paper Due

Sources

http://agiliq.com/blog/2009/06/generating-pseudo-random-text-with-markov-chains-u/

http://shiffman.net/teaching/a2z/generate/#ngrams

http://www.americanrhetoric.com/barackobamaspeeches.htm

https://textblob.readthedocs.org/en/dev/