Chris Castañeda-Barajas
Data Science 350: Methods for Data Analysis
Final Project Write up
June 2, 2016

Legislative Analysis using Natural Language Processing
Text Mining the Wisconsin State Legislature for ALEC Model Legislation

The American Legislative Exchange Council (ALEC) is a highly influential organization of conservative state legislators and corporate lobbyists. It's been quietly shaping American public policy for more than four decades now by operating in all 50 state legislative bodies. Its primary tactic has been to generate "model legislation" in private sessions between its member-legislators and its corporate funders, and then have these model bills introduced en-mass, often times word-for-word, across the country, year after year. While ALEC is by no means the only organization that produces model legislation, it is arguably the most (in)famous and secretive. Most of what is known about ALEC is due to the work of journalist and policy analysts painstakingly tracking legislation in their respective states. The challenge in developing a complete picture of the influence ALEC has over American public policy is that it practically requires analyzing every piece of legislation that has been introduced in each of the 50 state legislatures over the past 40+ years.

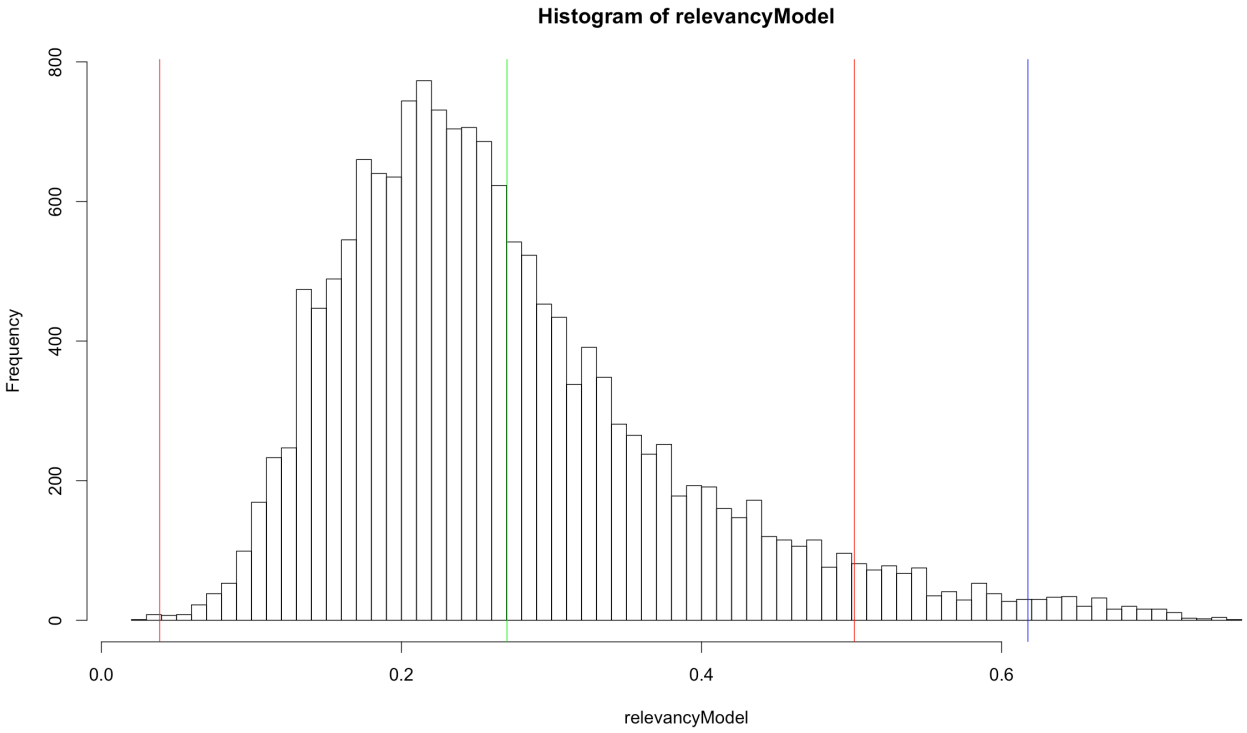
For my project, I used natural language processing (NLP) techniques to data mine the Wisconsin state legislature. I harvested the texts of every piece of legislation introduced into the Wisconsin state legislature from 1995 to 2014 (my scrapper bot had issues with extracting the texts from the 1999-2000 and 2015-16 bienniums, so those texts have been omitted from my analysis; legislation prior to 1995 were not easily accessible in a digital format). In total I harvested the texts of nearly 17,000 pieces of legislation. I also scrapped the text of known examples of ALEC Model legislation from the Center for Media and Democracy's ALECExposed.org website, specifically examples of bills related to Guns, Prisons, Crime and Immigration here: http://www.alecexposed.org/wiki/Bills_related_to_Guns,_Prisons,_Crime,_and_Immigration.

Methodology:
In short I used a cosine similarity model to help identify potential model legislation. I calculated the TF-IDF scores on of the set of known examples of ALEC legislation to then develop a target term vector representing this specific set of model legislation. I then calculated the cosine similarity between the target term vector and term vectors representing to each piece of legislation. I used the cosine similarity values as my relevancy score for each piece of legislation, those documents with a score close to 1 being the ones most similar to my example legislation.

Effectiveness:
I managed to whittle nearly 17,000 documents, introduced over two-decades down to a very manageable set of 246 documents (I simply took just the top outliers of the set).

It's clear from just a cursory examination of the top documents, that each of them relates to crime policy in one way or another. It's not clear whether these are actual ALEC model legislation at this point, but it's a starting point for more in-depth analysis. At the very least, these top documents are at likely representative of a very conservative policy perspective, similar to the one advocated for by ALEC.

Below is the histogram of my relevancy model (cosine similarity values); the green line marks the mean, the red lines mark two standard deviations from the norm, and the blue line marks three standard deviations. Everything to the right of the blue line represents the 246 best matches my algorithm identified.



Next Steps:
The obvious next steps would be to apply this analysis to other states as well as use other policy topic areas to model against. I also think this method would mesh well with a logistic classifier, if one were able to label a large enough set of know model legislation in the wild.