Chris S. Castañeda-Barajas

DS350 Final Project Proposal

04/21/2016

I’m interested in exploring solutions for identifying “similar” passages of text across large collections of documents and textual data. I’m interested in applying data science tools to legislative and public policy research. Specifically, I want to develop a model/tool for identifying “model legislation,” more specifically model legislation originating form the American Legislative Exchange Council ([www.alec.org](http://www.alec.org)). This idea developed from a personal interestes/side project I had prior to starting class (honestly, it was part of my motivation for signing up for this Data Science certificate program).

**Research Question:**

I realize this a big project, so for the purposes of this class I’d like scale down the project to this: given a set of known model legislation, can I take the text of those documents and develop a model using natural language processing and other statistical techniques that can be used to identify matches in new sets of legislative documents. I see this as supervised machine learning problem, similar to spam or fraud detection.

**Data Sources:**

I intend to mine the text of legislation introduced into a select set of state legislative bodies in the United States for instances of “model legislation” being used. To my knowledge, there’s no centralized datastore or API available for all state legislative data, so I’ve resorted to web scraping for this project. I’ve already built web scrapers for WA and WI state legislative websites; both of those states have the text of all introduce legislation easily accessible/scrapable over the web going back 25 and 21 years, respectively. My WA state dataset would comprise of about 37,000 individual documents; WI, about 21,000 documents. For this project I intended to add a few more states to the pool.

I will also utilize the text of known ALEC model legislation which I will gather from what’s been published on ALEC’s website ([www.alec.com](http://www.alec.com)) and from what’s been identified by various media and watchdog organizations (namely [www.alecexposed.com](http://www.alecexposed.com)). These documents will be used as my training data.

**Data Transformations & Analysis:**

I feel like this project is a natural fit for natural language processing (which I intend to learn a lot more about through the course of this project). From what I’ve read about NLP already, calculating the tf-idf (term frequency–inverse document frequency) for each document seems like a good first step in analysis. I’m still uncertain about exactly what other techniques I may use.