Springboard-Data Science Track
Oregon State Legislature Bill Clustering
Capstone Project 1 Proposal
3/13/2020

The proposal should address the following questions:

- **What is the problem you want to solve?**

Imagine a situation where you are working for a political party as an analyst and you have an extensive amount of jargon heavy legislative reading to get done. If that task wasn't enough on its own, you are working under a looming deadline as well. The time that you would spend manually reading is prohibitively large. What if you could run a program that would do this heavy reading and summarize the actual documents into a nice neat table (or story or whatever)?

One of the long-term goals of this project is to do just that. Any time you can take something that takes a long time and find a way to do it much faster is good, so long as quality is preserved. In this case, machine learning techniques would be applied to real data from the State of Oregon in order to analyze an incredible amount of recently passed legislature in a very short time and come away with a high level understanding of the material. The broader applications of this program are virtually endless, and the scenarios where having this tool might mean success or failure when time is short are many.

- **Who is your client and why do they care about this problem?**

Policy makers, government officials, city planners, lawyers, analysts, consultants, etc. These stakeholders care about this problem because being able to digest a large amount of difficult and cumbersome language in a short amount of time is priceless in this fast-paced world.

- **In other words, what will your client do or decide based on your analysis that they wouldn't have done otherwise?**

Prospective clients would be able to answer questions that would have been answered by reading the text manually. Of course it is important to understand that reading a summary is not exactly the same thing as reading a document from start to finish but it is useful to be able to understand the main ideas.

- **What data are you using? How will you acquire the data?**

This project will utilize data acquired from the city of Portland's open data website. That data can be found here. The dataset used in this project consists of bills signed by governors of the State of Oregon from 2013-2019. In this project we will focus on bills associated with Governor K. Brown for a window of time to be determined.The data will be acquired through an API that the State of Oregon uses called Socrata Open Data API or SODA for short.

- **Briefly outline how you’ll solve this problem.**

Ultimately, this problem will be addressed using an application of machine learning algorithms.

First the data acquisition will begin by using the API to access the multiple tables of bills signed by the chosen governor by year. Each table contains links to the documents but not the actual text of the document. Therefore wrangling in this case consists of two parts: obtaining PDFpdf documents from a given URL, and extracting text from PDF documents.

In this capstone project we will address the simpler question of characterizing similarities among the various bills per year for the chosen governor. More specifically, we envision using clustering algorithms to group the bills according to similarity metrics to be determined, to then examine the words associated with each cluster, hoping to find descriptive topics associated with them.

- **What are your deliverables? Typically, this includes code, a paper, or a slide deck.**

We will deliver all Jupyter Notebooks that will be developed, a written final report, and a presentation slide deck.