Regulatory Analytics

Jeremy Miller Deval Shah Helly Jain Jyothsna Krishnamurthy Duong Vo Aaron Roell

Motivation/Introduction

US federal rulemaking affects millions of people and with many groups interacting. Rule creation generates a large amount of data and continues generating during its legacy (e.g. court battles).

Analyzing the relationship between rules and their source law or public opinion is manual and retroactive. Public comments are considered, but no formal analysis is utilized when making these decisions.

There is no known application that links this data.

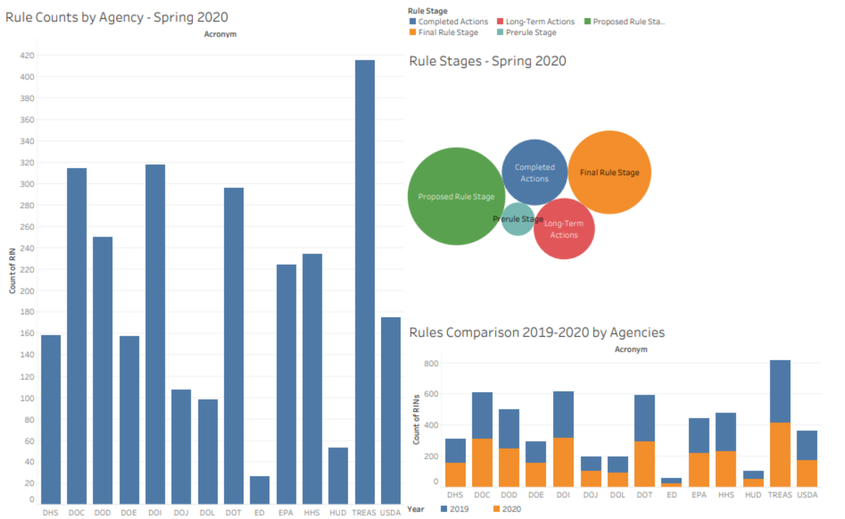


Tableau dashboard showing overall number of rules and rules process for 2020 and rule comparison between 2019 and 2020

Data

Data was obtained from:

1. Govinfo.gov
2. Regulations.gov
3. Reginfo.gov

Data is freely available via API or XML download resulting in 25,000 rules and regulations, 25,000 commnets (approximately 600 MB).
Initial data analysis and visualization was performed using data from 2018 through 2020

Approaches

Data obtained from API was obtained using Python
Data was stored using a SQLite database and was associated between sources using SQL
Data visualization was performed using D3 JavaScript
Sentiment Analysis was perfomed using Word2Vec neural network model trained on the Large Movie Review Dataset (Maas, 2011)

The combination of these approaches produces a network graph that automatically shows the relationship between a regulation and its source law (rather than manual association) and perform a sentiment analysis to summarize the overall public opinion as the law is being proposed.

Experiment and Results

An initial visualization of 10,000 rows from the dataset resulted in a sparse network graph. This helped us iteratively refine the visualization by filterting to only include the data with connections, adding tool tips, and aggregating the sentiment score.

A Keras model was trained using a 1 Convolutional layer with relu activation and a fully connected hidden layer with 250 neurons for the results. A dropout of 0.2 was used in hidden layer to introduce some bias in the model and avoid overfitting. Adam optimizer and binary crossentropy loss were used fit the model on imdb movie review data. The results aligned well with the sentiments validated manually

A small subset of comments were manually analyzed for sentiment. We observed a disparity between the movie training set (where sentiment was more clearly divisive) and our data (which was spread throughout the entire spectrum of sentiment). Our code also had to exclude comments that referred to an attached pdf for the comment on the rule.

The results of the sentiment analysis were also used to generate word clouds for negative comments, neutral comments, and positive comments. This kind of analysis can be used to validate our sentiment model in a holistic way, and see the commonly used phrasing in the comments.

Negative

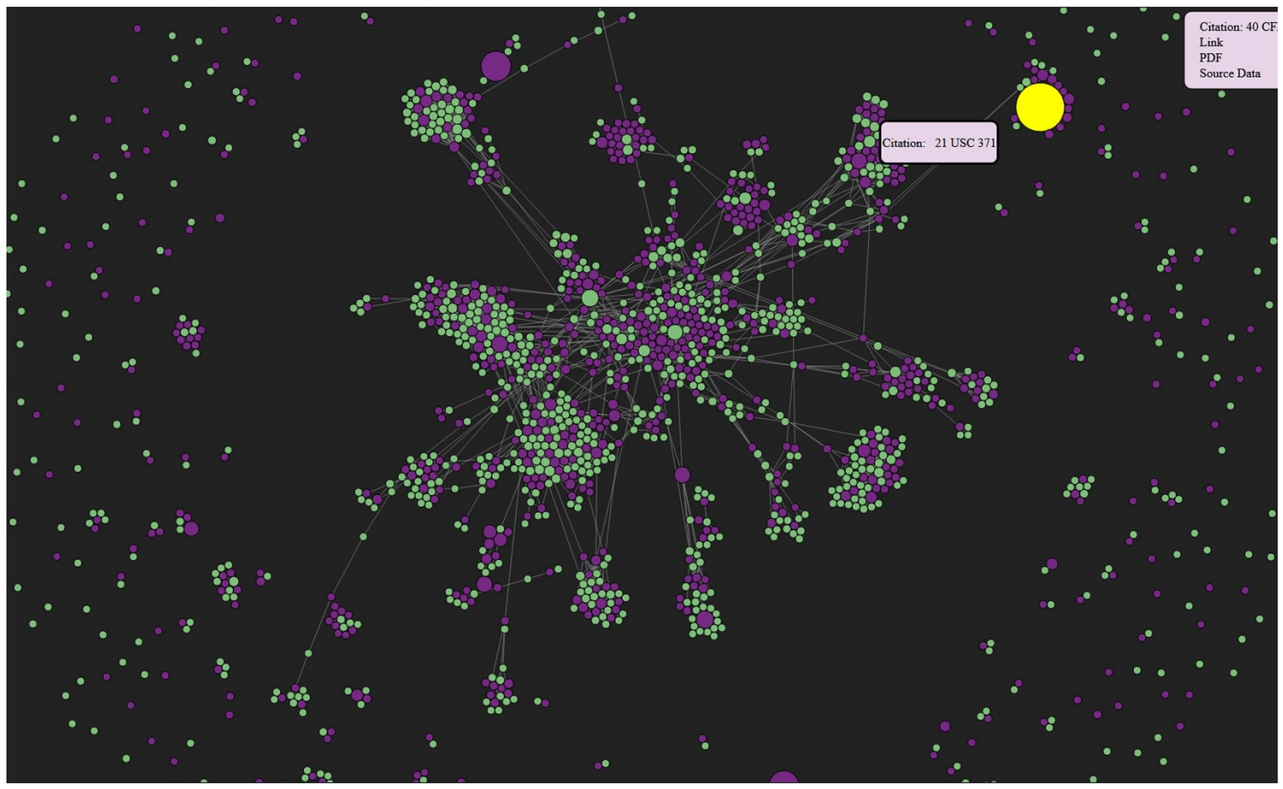
Neutral

Positive

Discussion

The visualization is a force directed graph representing the interconnections between top 2000 US Code (USC), Code of Federal Regulations (CFR) and Federal Register (FR) citations by Page Rank related to the public comments analyzed.

The full graph is approximately 150,000 nodes across 4 node types. Size represents the node's Page Rank; color represents the node type. Green nodes are statutes from the US Code. Purple nodes are regulations from the CFR. No FR nodes had a high enough page rank to be displayed. The Yellow node is the selected node and its details are displayed in the box on the top right which links to the web, PDF and XML versions of the node. The is also a tool tip for the hovered node.
Forces include a many-body force to repel nodes from each other and both center and link forces to pull related nodes. A collision force keeps nodes from overlapping one another.



Recommendations and Future Improvements

There are several opportunities in the area , we have only scratched the surface using NLP, sentiment analysis, and data integration using network graphs. Apart from these, our project can be expanded to find:
1. Outdated regulations
2. Using text analytics and relatedness score one can determine sections of CFRs that are similar copies of CFRs in other regulations
3. Regulations that are similar or contradict one another.
4. An exploratory visualization where the user selects a regulatory entity (statute, regulation, comment, etc.) and can hop recursively to related entities
5. Additional filtering withiin the visualization so the user can focus on the data they are interested in
6. Reading information from a pdf document and attachments to the public comments
7. Automate the data acquisition; currently the code used to amass the final data set requires significant labor in order to obtain all of the necessary source files

As an example, in 2017, there was an executive order to reduce cost of regulations “One-in Two out” , implying for every new regulation two regulations will be eliminated and the net additional costs of the new and old regulations must be 0. Here is where our project would play a major role. Instead of eliminating regulations that protect people , using NLP processing we could identify regulations with similar impact. Our initial sentiment analysis on the comments can be extended to entity extraction and regulations that are impactful or non-impactful. Identifying those in an automated way would optimize time and cost.