Open Project Proposal

Thomas Bagley, Sachet Bangia, Kushal Byatnal, and Jason Novack

Our team will be working on the suggested quantitative fact-checking project through the data on GovTrack.us. We would hope to provide an application that would allow users to explore and visualize the rich dataset available. After our discussion with Professor Yang, we have decided that we will join his project team developing similar capabilities on the same data set. In preparation for our work with his team, we have considered a variety of features that we think would be in public interest.

First is the ability to view which representatives and senators vote most similarly to each other. In its most basic form, this would take the form of a ranked list of votes on all bills in the last several congresses, where the user can specify a start date. However, this also allows for several expansions. One would be a visualization of these data, allowing for a more engaging comparison of voting similarity. Another would be to filter votes by “controversial” bills, as defined by the difference in the fraction of Republicans and Democrats who voted for each bill.

Second, we would allow users to select any two representatives in congress to compare similarity in voting records, both in the most recent Congress and over time. This also allows for extensions, including the ability to compare several representatives simultaneously (though likely not more than 5). Also, when selecting two representatives (call them A and B) we could also return a list, or just the number, of representatives who are more similar to A than A is to B for a more qualitative comparison of their similarity.

For both types of comparison above, we also propose the ability to organize bills by category, such as health care and gun control to compare similarity on specific issues. The automated categorization of these bills may be difficult, but this would be incredibly useful so we would like to figure out at least some implementation.

One final feature we have considered is summary statistics and visualizations for Congress over time. This would allow for analysis of overarching questions concerning political polarization and prediction of future voting trends.

This project is interesting because it allows for quantitative study of political issues affecting the entire country. These issues are often discussed in qualitative ways and through hypotheticals, but this will allow for concrete analysis and exciting visualizations of the reality of these issues.

A previous attempt to visualize polarization of the US Congress over time can be found [here](http://www.washingtonpost.com/news/wonkblog/wp/2015/04/23/a-stunning-visualization-of-our-divided-congress/). This type of network graphs could be one possible way we would display polarization, but in its current form, this visualization fails to convey the nuances of what may be happening. For instance, there is no mention of the number of bills brought to a vote in each Congress, and which party was in control of each chamber. This may be relevant as Republicans have voted to repeal the ACA several times since gaining control of the House in 2010, but more bills with bipartisan support, such as immigration reform, have not been brought to a vote. Thus the increased polarization may be caused because inherently controversial bills are voted on more frequently.

One example of political analysis software is NOMINATE (Nominal Three-Step Estimation), developed by Keith Poole and Howard Rosenthal. The software developed a multidimensional scaling model for comparing Congressmen’s voting records across issues, and also allows for comparison of voting patterns. These calculations can be demonstrated in two and three-dimensional graphs. Starting with the methods they use, it would be possible to create a variety of analyses and visualizations for voting records. Similar methods may be able to be used for other probabilistic comparisons relevant to this project.

Developing automated visualization techniques to quantify the “interestingness” of claims relies heavily on programs being able to search and identify relevant parts of the query. One notable instance of this is the “Truth Teller”, an app developed by The Washington Post that fact-checks live political speech. To identify any relevant parts of the query, a fuzzy string matching algorithm based on the Levenshtein distance model is utilized. The Levenshtein distance model is a metric of measuring the difference between two queries based on the minimum number of single-character edits (i.e. insertions, deletions, or substitutions) required for transforming one query into the other. This is the basis for how a large portion of fact-checking programs operate when identifying facts and is an important step leading towards effective automated analysis and visualization.

We will know more about the building of the application on Friday, when we meet with Professor Yang’s team. We will of course be following their lead on system architecture, platform, and other application-level issues.

Our discussion with Professor Yang occurred over email on October 14, 2015, where he responded positively to these ideas and suggested we work with his research team studying the GovTrack data set.