Midterm Report -- Estimating Senatorial Influence

Gregory Adams

Where I started

My initial goal was to build a meta-classifier that combined network analysis and natural language processing to develop a measurement of the influence of individual Senators. I still want to understand influence, but outside of the network analysis, it's difficult to directly measure. So I've refined my goal to finding the senators that *add the most value*, i.e. increase the probability that the bills they cosponsor will pass the most.

The data I started with gave me a solid start for the project, but was somewhat limited. The main thing I have been using from the initial data are the cosponsorship networks, which is pretty much just two json files. Other than that, though I spent some time trying to recover other aspects of the data--such as the names of the bills each person sponsored, or the sqlite databases--I was unsuccessful. It seems that, because of the size of the data, at some point I deleted them.

My initial milestones for this time in the quarter were the following: first, to recover the data from my previous project and make it useful. Second, to build a web scraper and pull down all of the bill texts. And third, to begin building and tuning a TFIDF classifier, trying to predict whether or not a bill would pass.

Where I am

Of the three goals I initially set for myself to try to accomplish by the midpoint of the quarter, I have accomplished the first two, and made significant progress on the third. I've (1) recovered my old data (insofar as it's possible), (2) pulled down my new data, and (3) begun building my NLP classifier. In addition, I have begun looking into different network analysis methods that I can use for the second part of the project.

With regard to the first goal, this mostly involved combing through my old repository and recreating my previously-completed analysis. The visualizations still

work (even independently of the now-deployed version), but I was unable to gather some of the secondary data we had gathered, namely, the names of bills that each Senator had sponsored/cosponsored. One of my old project partners mentioned he may still have a copy of the sqlite database in which we stored that data; I am in the process of getting that from him.

The second goal has been my biggest focus so far. Congress.gov, while a well-structured website, does not make it particularly easy to scrape data. With much of the necessary information wrapped in javascript, I used selenium to build the scraper and download the pdfs of the bills, converting them to text files as I go. I plan to use regular expressions to pull out the sponsors from the text data; if that does not work, I can go back and modify my scraper to pull out cosponsorship data as well.

Finally, I've begun building and tuning my TFIDF classifier. Far from finished, I'm currently focusing on cleaning the data as well as possible (to avoid redundancies, identify stopwords, tokenize effectively, etc.), as well as creating the feature generation pipeline. I plan to use my model loop from ML for Public Policy last spring in order to do the modelling.

While I'm focusing in on the text right now, I've come across datasets online that give different data points for each Senator (e.g. seniority, money raised, etc.). I may integrate those factors into my text analysis model in order to increase predictive accuracy. Alternatively, I may create a separate model out of these factors, combining that model with the other two in my final meta-classifier.

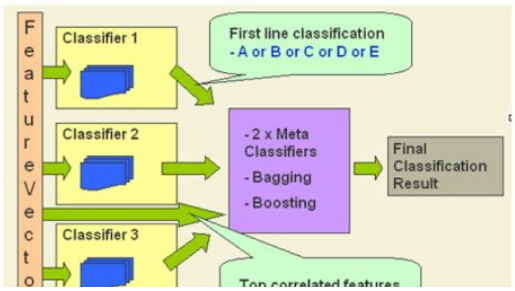
The final thing I have been doing is reading up on the literature surrounding this topic, especially with regard to network analysis and meta classifiers. There are three sources from which I may draw methods which discuss one of each network analysis for congress, combining NLP and network analysis, and meta-classifiers more generally.

The congressional network analysis not only argues persuasively that the method I'm using to measure leadership is valid--it cites five different sources arguing that cosponsorship is an effective signal of leadership--but it gives alternative

measurements of network centrality and influence that I could use.[1] Though PageRank is probably the most reliable method, building a model of different measures of centrality (e.g. a Logistic Regression combining pagerank output, raw number of co-sponsorships, etc.) could vastly increase predictive power.

The meta-classifier paper, though not peer-reviewed helps me design my algorithm: I plan to use a very similar structure to the design in figure 1 (from the paper).[2] And the final paper develops a unique method for integrating the two different analysis techniques that contrasts with the parallelized approach from the previous source, allowing me to compare two different models for integration.[3]

Different from bagging and boosting, we use ***Decision tree (c4.5)***, ***Neural Network*** and ***Naïve Bayes*** as meta-classifiers respectively. Additionally, rather than rely solely on output scores or on the set of domain-level features employed in ***bagging*** and ***boosting***, we introduce the use a set of features that provide a low-dimensional abstraction on the original feature set.



Where I'm Going

Once the data is gathered, the network built and the NLP classifier is up and running, I need to move onto two more main goals: building my network analysis model, and building my meta-classifier. I also need to work on joining the cosponsorship data from each dataset, so I can build my classifiers from an integrated source. I'm roughly on track with my initial schedule, and I plan to have most of analysis finished by the end of 10th week.

7th Week:	Continue building text analysis models (Goal 2 from initial report). Experiment

[1] James, Fowler H. "Connecting the Congress: A Study of Cosponsorship Networks." Fowler.ucsd.edu. Accessed February 7, 2017. http://fowler.ucsd.edu/best_connected_congressperson.pdf.

[2] Siyang, Gu et al. "Meta-classifier in Text classification ." Imada.sdu.dk. Accessed February 7, 2017. http://imada.sdu.dk/~zhou/papers/cs5228project.pdf.

[3] "Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media." Pnas.org. Accessed February 07, 2017. http://www.pnas.org/content/113/42/11823.abstract.

	with different types of models and optimize classifier.
8th and 9th Week:	Work on Goal 3 from the initial report: run the network centrality analyses and experiment with different types of models for predicting passage.
10th Week:	Integrate text and network models. Optimize and validate final model.
Finals Week:	Buffer Period. Prepare final presentation slides and polish project.