Project Proposal
Who Are Trump Supporters?

CS 7930 Social Media Mining
Spring 2016

Gopal Menon
Computer Science Department
Utah State University
Logan, UT
gopal.menon@aggiemail.usu.edu

1. BACKGROUND

The current front runners in the 2016 Presidential Election are Donald Trump from the Republican party and Hillary Clinton from the Democratic party.

The Republican party base has been steadily moving away from traditional candidates as they reportedly feel that they do not have a say in matters that are of importance to them. The rise of the Tea Party and the influence of Talk Radio, has reportedly made the Republican base disillusioned with Congress. In the last few elections, traditionally safe Republican candidates have been unseated in the primaries by Tea Party candidates.There has been growing discontent with President Obama's perceived Liberal agenda and people in the Republican base seem to want to start afresh with new candidates as they seem to have lost trust in career politicians.

Trump is not a career politician and has no experience in government. He started out as an outside candidate who was not taken seriously. However he has succeeded in becoming the front-runner by appealing to the base of the republican party. He has been accused by many of being racist, fear mongering and wrong on things that he has put forward as facts. He is accused of being neither religious nor conservative and has still emerged as the likely nominee from his party. He is reportedly disliked by members of his own party, but despite this, has emerged as the leading candidate.

The Democratic party traditionally has its strong base in the East and West coasts, Liberals and minority communities. Unlike the Republican party, the Democrats have not been going through any internal turmoils. However, their candidate has had her share of problems. Clinton has been associated with scandals from the past like her husband's infidelities and the Whitewater investigations. During her term as Secretary of State, she was held responsible for the death of the American ambassador along with three other persons, during the attack on the consulate in Bengazhi. After her term, she was faced with the email scandal. Outside of her party she is reportedly disliked and is accused of being corrupt and is a very divisive figure.

The Democrats, unlike the Republicans, started out with just four candidates and were left with only candidates - Hillary Clinton and Bernie Sanders. At the time of writing this proposal, Clinton seems to be the likely nominee from the party.

2. RESEARCH FOCUS

My research focus will be on using Tweets to find the profile of the supporters of the two candidates. I am primarily interested in finding out what issues matter to the supporters.

2.1 Why is it interesting

I find this interesting because of the success of the Trump campaign. He had been written off initially by news analysts and they did not expect him to last beyond the initial phase. His continued success has taken many people by surprise. By finding the issues that matter to his supporters, I can find out what are the key factors that drive the apparent demand for change among the base. Since the career politicians have not been successful, I can find out how the Republican party is out of touch with the issues that matter most to its base. The reasons could be perceived loss of employment opportunities due to globalization and illegal immigration, changes in the demographics of the voter population, perceived Liberal policies of President Obama, opposition to gay marriage and the Affordable Care Act (also known as Obamacare), fear of terrorism or it could be something that I have not thought of. Whatever the reasons are, they will be interesting as the emergence of Trump is an unforeseen phenomenon that maybe even he cannot explain.

The identification of issues important to the supporters of Clinton will be a bonus, but may not be as interesting as the issues that motivate Trump supporters.

2.2 Who will be interested

Social scientists, people who follow politics and key members of the two parties will potentially be interested in knowing the results of the investigation.

The Republican party will be interested as they appar-

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

ently need to change their focus since all their traditional presidential candidates have been rejected by the primary voters. After the election of President Obama in 2008, many news analysts said that the Republican party needs to change its attitude towards that Latinos, who consist of the largest minority, in order to be viable in the future. They said the party may need to change in order to survive.

The Democrats on the other hand, have not had any issues with their traditional base. They would be interested in knowing what issues are important to their opponents and would want to use this information in order to to better compete.

3. PREVIOUS RESEARCH

Previous research has been done by Kloumann and Kleinberg [1] on identifying members of a community starting from a small seed set. The seed set was expanded using PageRank and other algorithms. The PageRank algorithm reportedly had good results in identifying key members of a community given the seed set.

4. APPROACH

I plan to use the tweets collected as part of Assignment 1 as the seed set. If required, I will retrieve more tweets to use as part of the seed set. The PageRank algorithm will be used to identify important members of the community by sorting the community by PageRank and selecting the ones with the higher values. Once the important community members are available, I plan to use Latent Dirichlet allocation (LDA) topic modeling on their tweets using the R package for finding the topics important to them.

4.1 High Level Design

I plan to expand the seed set using the PageRank algorithm. Each member of the seed set will be a Twitter account. The other accounts followed by the seed account will constitute the equivalent of the forward links on a web page. In addition to the forward links, the list of accounts that are followers of an account will be the incoming links. This is where my approach differs from PageRank. The equivalent of the PageRank case where the graph walker is transported to a random web page, will be replaced by random transport to a member of the seed set. I plan to evaluate the output both with and without the incoming links. Use of PageRank will expand the seed set and give me a big community that is linked together.

When the community is sorted by decreasing order of page rank, I will obtain the thought leaders in the community. These accounts with higher page rank will be thought leaders as they will be the ones that are most important due to the fact that more people follow them, and because more important people follow them.

Once the community though leaders have been obtained, I plan to use topic modeling on their Tweets in order to find out what issues and topics interest them and their followers.

At the end of the project, I will have obtained the thought leaders in the Twitter community for each of the two presidential candidates and the topics and issues that matter to them.

4.2 Tools and Techniques

I plan to use Twitter APIs through Python for retrieval of Tweets and for Twitter account crawling. There are some repositories on GitHub that claim to be able to crawl Twitter. I would need to investigate these. The ability to recover from errors and restart along with the capability to crawl the Twitter graph in parallel will be an important criteria to consider.

The Twitter graph of users along with followers will need to be stored in a graph database. For this purpose, I would need to evaluate and use a graph database.

I would need to code or use a function that would compute a left eigen vector for computing the PageRank as described in [2].

The R package has topic mining libraries that I plan to use for finding the major topics in the profiles and Tweets of the thought leaders of the community for each of the candidates.

4.3 Risks

- As was seen in assignment 1 for retrieving Tweets, the process may take a long time and may not return many results. As a result, the process of retrieving the seed set or the Twitter graph of users and followers may take a lot of time.
- The Tweet retrieval process would fail intermittently.
- Twitter has limits on the retrieval rate that would limit data collection for the project.
- The seed set may not be of the right size or the right mix of accounts. This would limit the ability of the crawling process to uncover the important members of the community.

4.4 Mitigation of Risks

- In order to mitigate the risk of not getting enough data due to slow retrieval, I would need to start the process early.
- The retrieval process would need to be able to recover from failure and restart automatically.
- I would need to create multiple Twitter accounts in order to get around the limits on the retrieval rate. Retrieval would need to be interspersed with delays so that maximum limits are not reached. In order to further speed up the retrieval process, I would need to do parallel retrieval using the multiple CPU cores available.
- I would need to do further research in order to find out the optimal seed set composition for the project.

5. MILESTONES

- High Level Design complete - mid March
- Seed set identification complete - 3^{rd} week of March
- Twitter account crawling complete - 1^{st} week of April
- Topic Modeling complete - 3^{rd} week of April
- Project Report complete - end of April

6. EVALUATION CRITERIA FOR RESULTS

I plan to compare the set of thought leaders obtained by following only forward links versus following both outgoing and incoming links. I expect the seed set expansion to be more complete in the latter case.

I expect to get a list of Twitter community thought leaders for each of the two candidates and their main topics and issues.

There is no way to predict what results I am going to get, although I am sure that they will be interesting. The interesting for me will be the answer to the question *who would want to support Donald Trump.*

7. REFERENCES

[1] Kloumann, I.M. and Kleinberg, J.M., 2014, August. Community membership identification from small seed sets. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1366-1375). ACM.

[2] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval/Christopher D.