Linguistic Features of Website Text for Presidential Candidates (Computational Linguistics, Project 1)

Jack Bandy

1. Choosing Text for a Corpus

Several different corpus sources would give different insights about the presidential candidates. I considered speech and debate transcripts, tweets, or blog posts. I ended up using text from the candidates' websites under the "issues" pages, because these sites, unlike the other mediums, serve a common purpose for each candidate. Each "issue" page is crafted to clearly communicate the candidate's position and explain his or her platform. There exists some variation, for example, Donald Trump essentially has three essays that communicate his positions and platform for different issues, whereas Marco Rubio provides over thirty different pages. Carly Florina's "issues" page only provides links to videos, so I could not use her for comparison in my project.

2. Website Scraping and Clean-Up

I used a recursive `wget` command to scrape issue pages from the following websites:

https://www.hillaryclinton.com/issues/
https://berniesanders.com/issues/
https://marcorubio.com/issues/
https://www.donaldjtrump.com/positions
https://jeb2016.com/rollouts?lang=en#
https://www.bencarson.com/issues

Once I had clones of each website, I used a python script to extract all the text in paragraph sections of the website, snipping out headers and footers.

3. Investigating the Text

I wanted to investigate two broad questions: (1) what kind of "political archetypes" do candidates tap into, and can archetypes be traced to the candidate's party affiliation? Then, (2) how does each each candidate present themselves linguistically? Essentially, some calculations with word and bigram frequencies can answer both questions.

4. Features

	Clinton	Sanders	Rubio	Trump	Carson	Bush
Length	8870	9172	17422	4913	1854	13146
Vocab	1916	2303	3124	1367	579	2943
Lex. Diversity	4.6	3.98	5.57	3.5	3.2	4.6
Top Bigrams	criminal justice	sen. sanders	president obama	united states	united states	united states
	health care	social security	american dream	tax rate	founding fathers	middle east
	minimum wage	united states	united states	concealed carry	innocent life	president obama
	Hillary believes	health care	supreme court	middle class	middle east	border patrol
	clean energy	nuclear weapon	second amendment	tax plan	american people	energy revolution
Noteable bigrams	human rights	billionaire class, middle class	21st century	america great, make america	dangerously belligerant	Ronald Reagan
Top Words	Hillary, America, care, women, health, plan, access, work, need, new	must, people, sanders, country, americans, war, time, veterans	american, president, america, world, life, must, people, need	tax, immigration, plan, america, americans	must, america, american, states, time, united, israel	veterans, president, energy, american, federal, must, border
First Name	139	3	16	4	0	18
Last Name	3	38	7	17	0	1
Self-Ref Pct.	1.6%	0.4%	0.1%	0.4%	0%	0.1%

5. Feature Intuition

Length.
(total words from issue pages) Ben Carson's site can be considered an outlier, each of his issue pages contains about three sentences. Rubio's site is considerably larger than others.

Lexical Diversity.
(total words used divided by unique words used) As expected, lexical diversity generally increases according to length of the respective text.

Self-references.
(first name or last name of candidate) No site used personal pronouns.

Bigrams and words.
(most frequently occurring word pairs and individual words) I used nltk's built-in stopword set for english to eliminate words like "a," "the," "it," and more from the analysis.

6. Observations and Future Work

Political Archetypes.
Although these pages are supposed to declare specific positions on issues, the frequently occurring bigrams and words do not serve that purpose. References such as "human rights," "american dream," "second amendment," "founding fathers," "ronald reagan," "21st century," "make america great," and more tap into political archetypes about which all of us have preconceived notions, while doing little to clarify a candidate's position. Clinton's frequent use of "Hillary believes" may be considered an exception.
As far as tracing features to party affiliation, the table reveals a few shallow insights. Because of the variance in length, it is not appropriate to say that one party has a stronger lexical diversity than the other, however, Clinton's site achieved the same lexical diversity as Bush's site with about 4,000 fewer total words. Health care made the top bigrams for both Democratic candidates, but not for any Republican candidates. President Obama made top bigrams for two Republican candidates, but not for any Democratic candidates. While these features are interesting, they are somewhat unsurprising.

Self-references.
As previously noted, Bush and Clinton refer to themselves almost exclusively by last name. Although I can only speculate the intention is a disassociation from dynastic politics, the choice is clearly deliberate. This can be seen even in Bush's website address, which is "jeb2016.com." Other candidates using their last name is not entirely surprising, however it is interesting that Ben Carson did not reference either.

Future work.
Although different candidates use twitter and blogging for different purposes that might make comparisons complicated, a corpus built from tweets and/or blogs may carry additional insight beyond what is found on the candidates' issue pages.