IST 736 - Final Project

Thomas Bahng | Matthew Kruse | Lauren Lawless | Alex Manso

30 May, 2020

Table of Contents

# Introduction

Lorum ipsum.

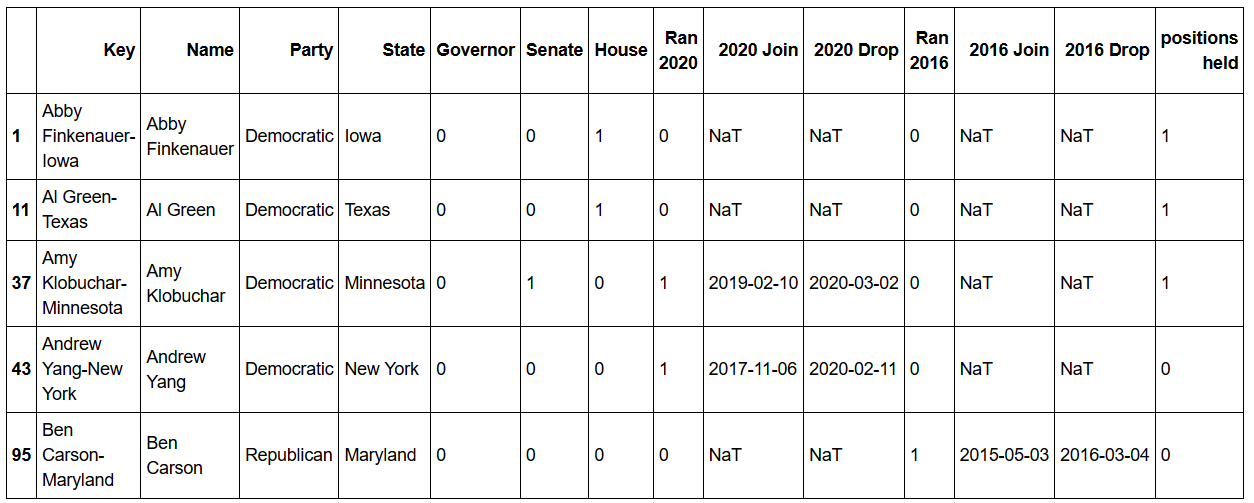
# Analysis

Lorum ipsum

### About the Data

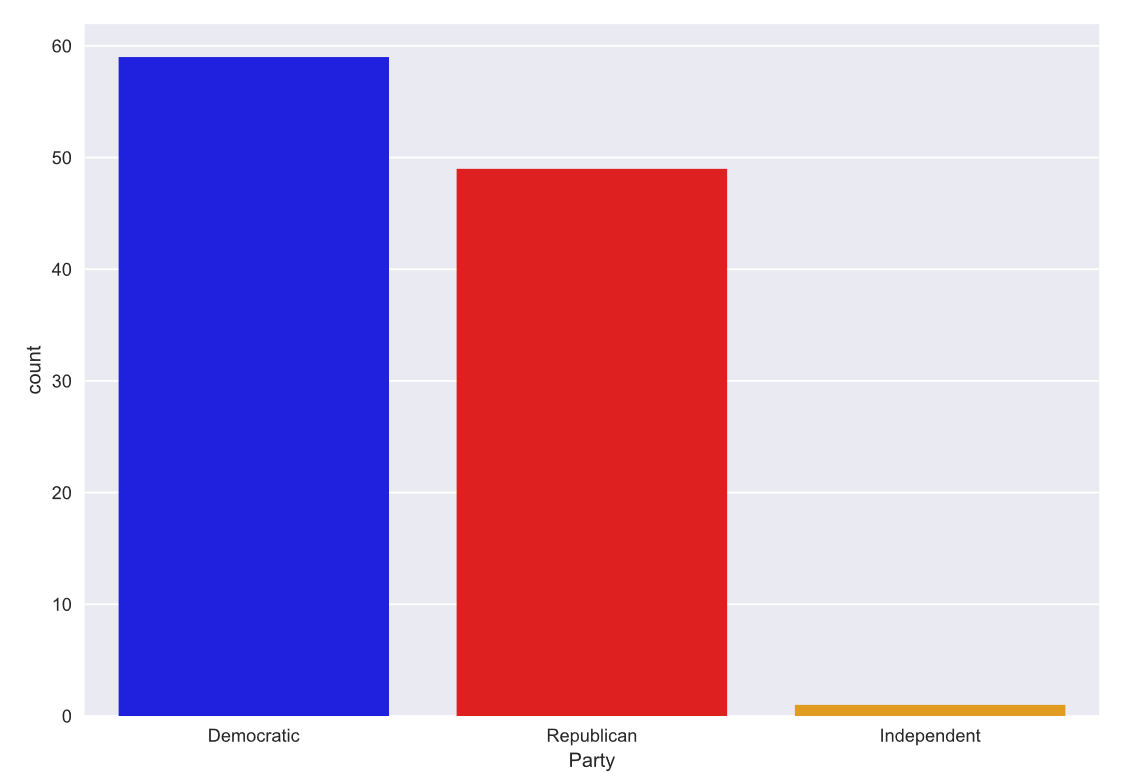
On May 24th, over 300,000 tweets respective to approximately 110 Twitter handles were extracted using the Tweepy library API. These were handles of pre-identified presidential candidates and sampling of non-candidates, mostly consisting of well-known political figures. This data was collected manually from publically available online sources such as <https://ballotpedia.org>.

#### Fig.1 Sample of People



This data was used to determine which individuals announced their candidacy to run for President of the United States, the date of the announcement, and other attributes. Each observation represents dimensional data specific to a person; data includes both **candidates** and **non-candidates**. Among these attributes are included their political affiliations as shown in proportions below.

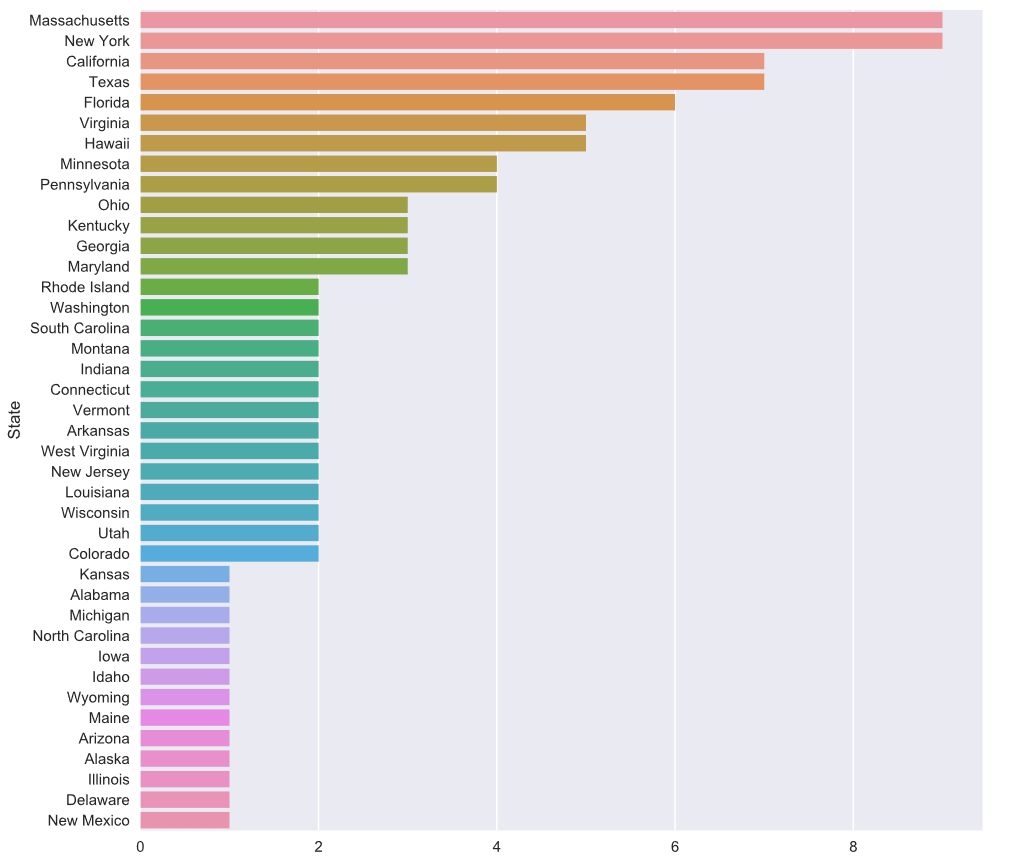
#### Fig.2 Party Proportion



The majority of individuals in the data set either favored or were part of the Democratic Party followed by the Republican Party.

In addition to their party affiliation, other information includes the states in which they were primarily active.

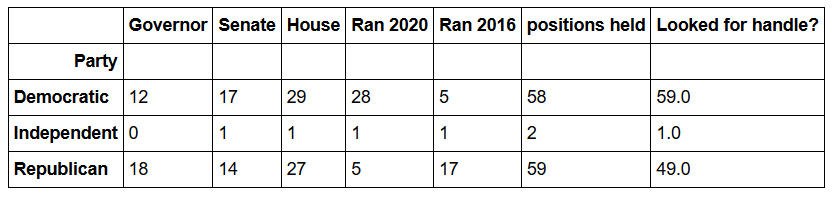
#### Fig.3 States Proportion



The top 5 states include Massachusetts, New York, California, Texas, and Florida.

The following table shows an aggregation of key attributes by political party.

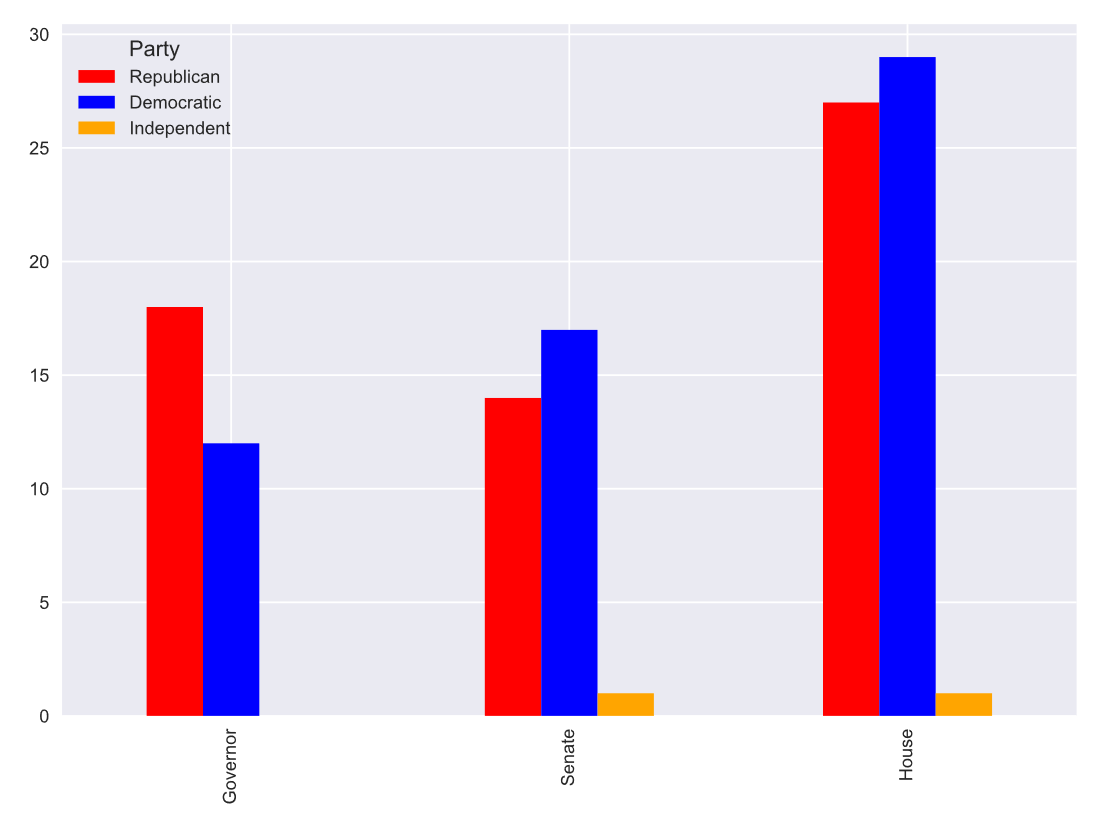
#### Fig.4 Key Attributes by Party



The majority of those who ran in 2020 were Democrats, contrasting with a Republican majority in the 2016 campaign.

The following plot compares party affiliation to their primary government role (i.e. governer, senate, house).

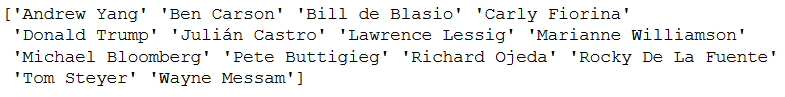
#### Fig.5 Party Affiliation by Government Role



Most of the Democrats come from House or Senate roles, whereas Republicans come from House or Governor roles.

There are 14 individuals in the data that did not hold a government role such as House, Senate, or Governor. Some of these have held municipal leadership roles and others came from the business roles.

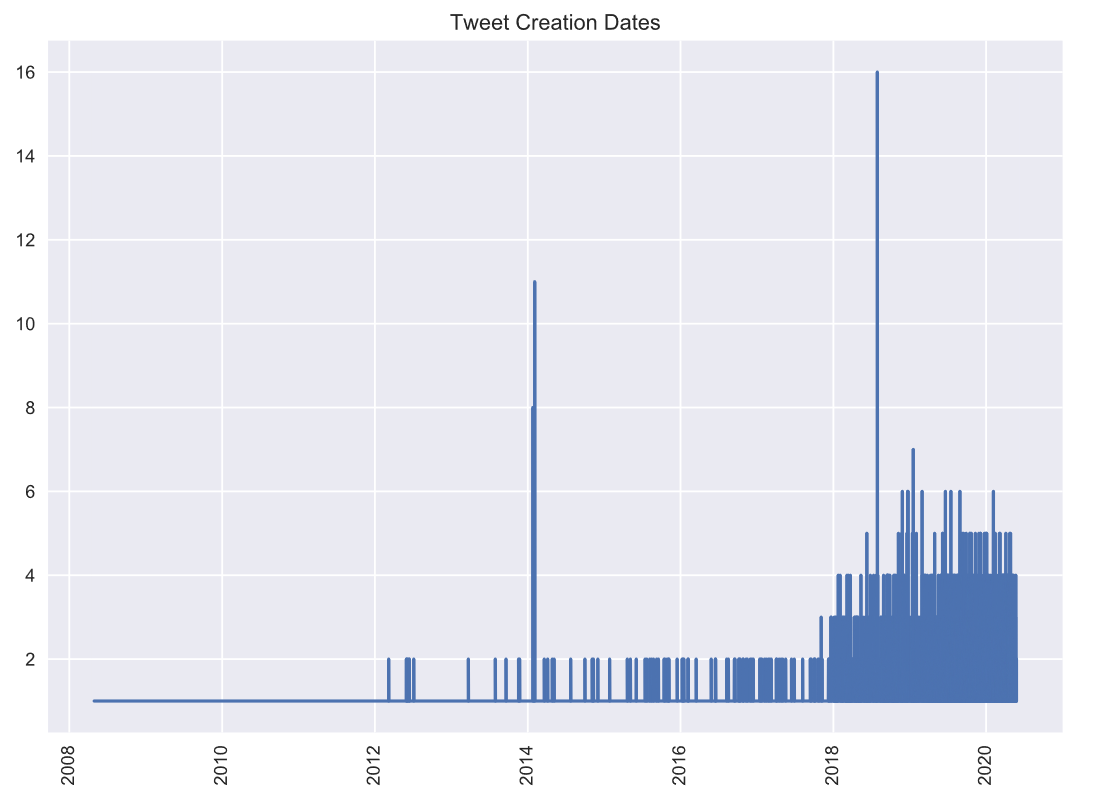
#### Fig.6 Individuals from Other Government / Business Roles



The original tweets collected from Twitter were formatted as JSON, filtered for english language, and subset to include just the text and creation dates. A total of 308,423 tweets were collected across 109 Twitter handles.

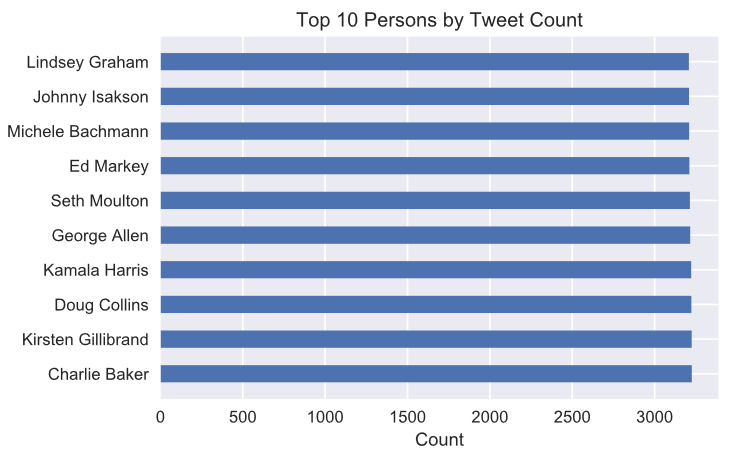
Altogether the tweets were created over a period of several years. The following shows this distribution.

#### Fig.7 Tweets Created Over Time



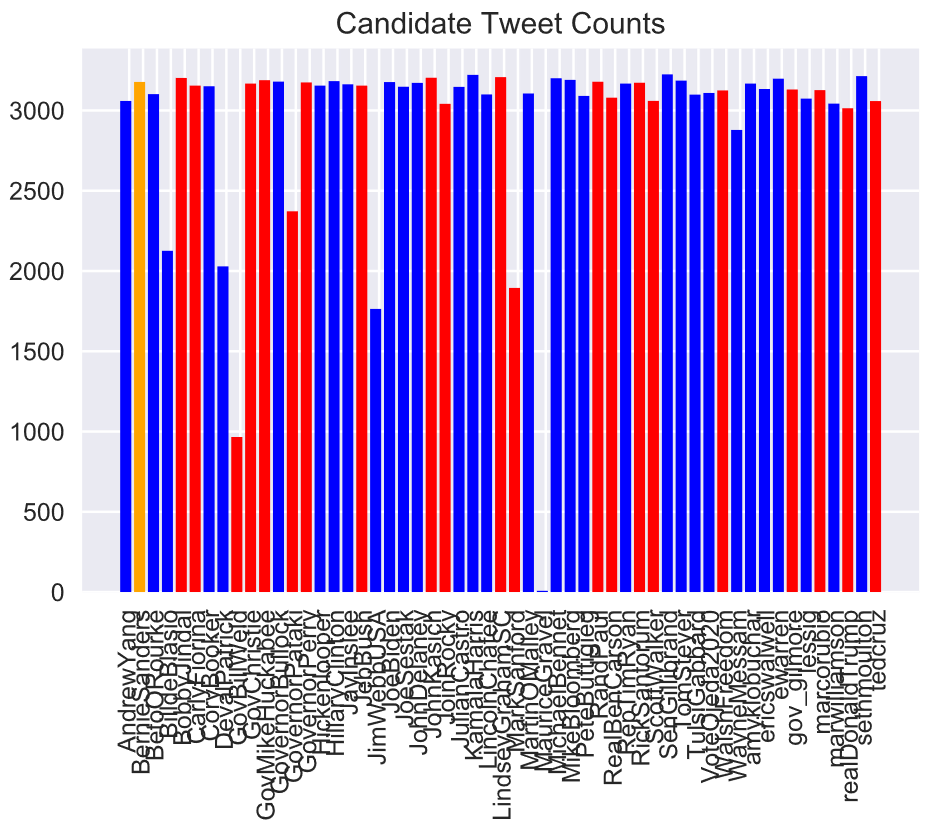
The tweets in this data set were mostly created between 2018 and 2020.

#### Fig.8 Top 10 Candidates by Number of Tweets

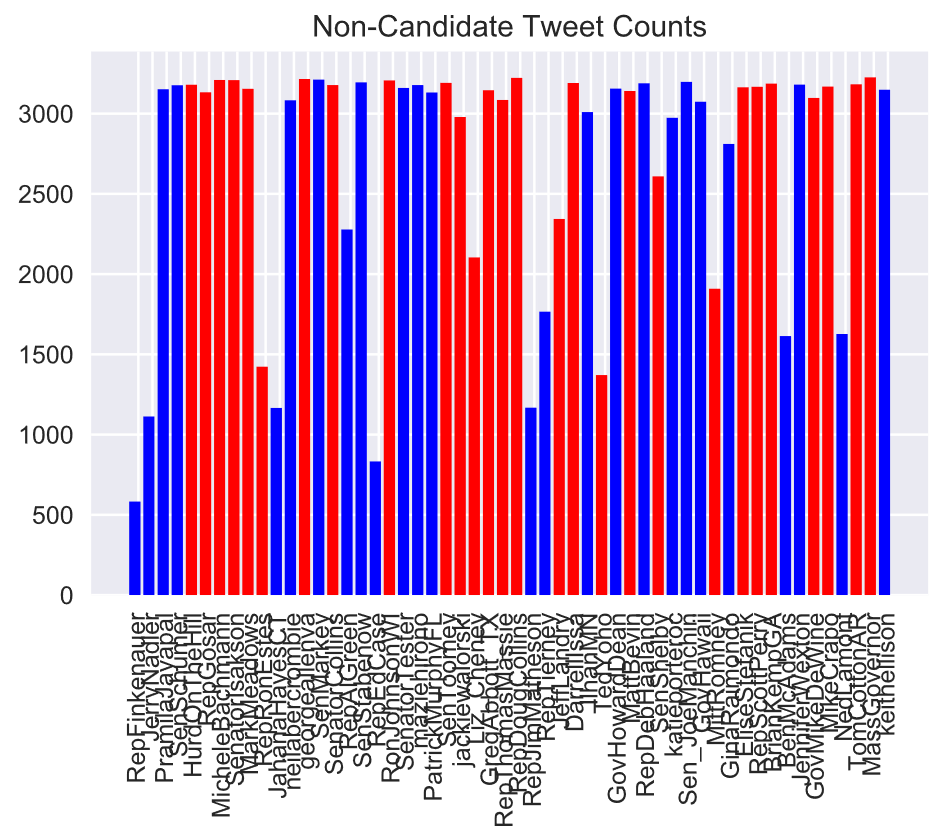


As a result of querying for all available tweets from each user’s Twitter timeline, there are approximately 3000 tweets per individual. A minimum of 8 tweets were collected by individual and a maximum of 3226.

#### Fig.9 All Presidential Candidate Tweet Counts



#### Fig.10 All Non-Candidate Tweet Counts



There are more observations of candidate tweets than non-candidate tweets.

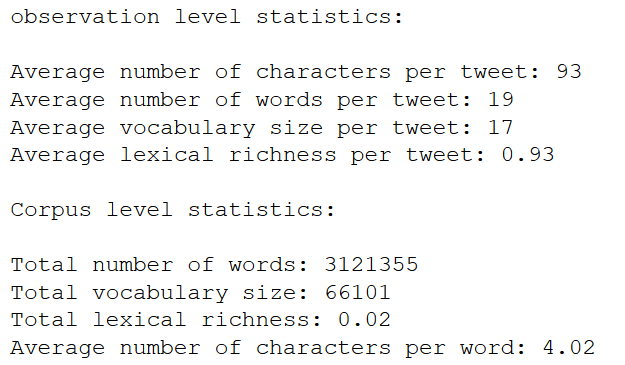
The tweet text was analyzed by candidates and non-candidates to assess for any noticeable differences. Methods for comparison included:

* corpus statistics of the tweet text
* term frequency
* sentiment
* part-of-speech

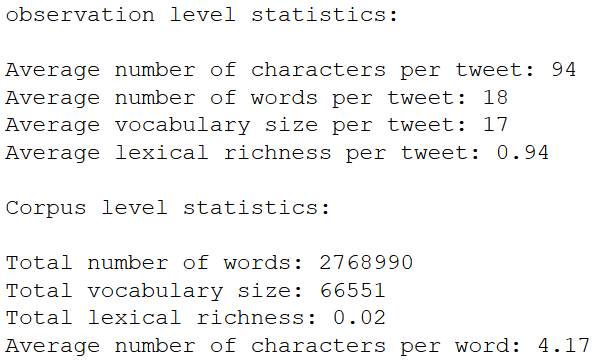
The following transformations / cleaning were applied to the tweet text:

* stopwords removed
* removed links, hashtags, mentions, and converted to lowercase
* alphabetic words only

#### Fig.11 Candidate Tweets: Corpus Statistics

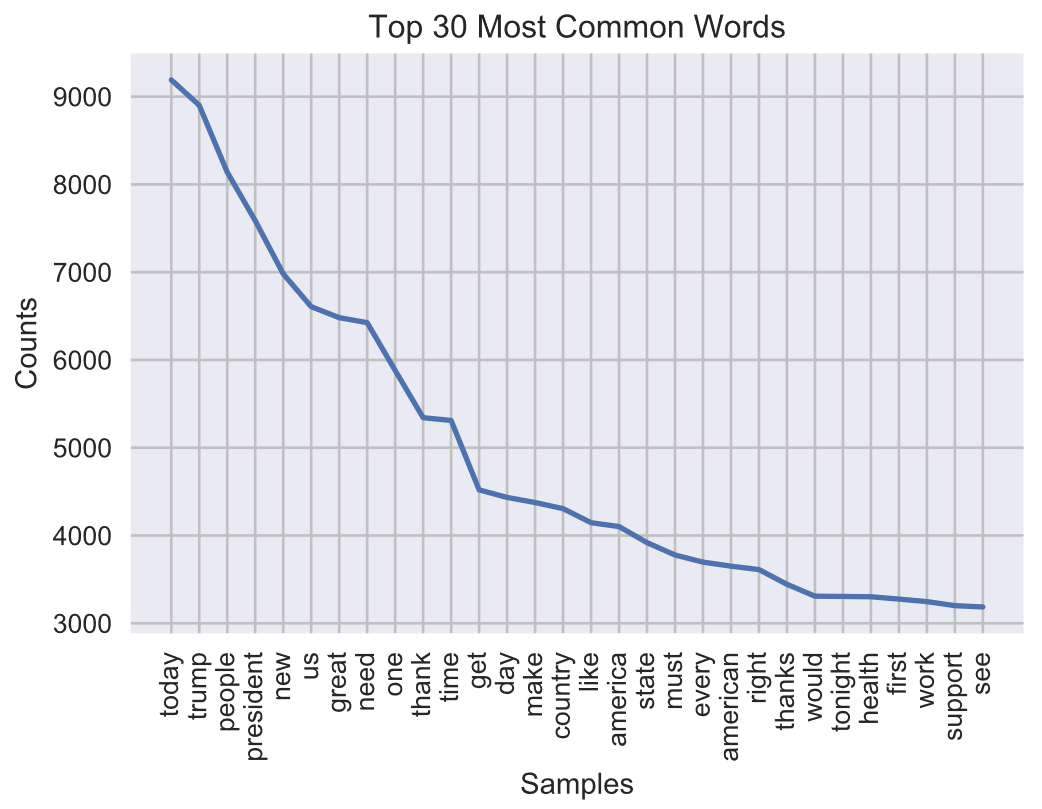


#### Fig.12 Non-Candidate Tweets: Corpus Statistics

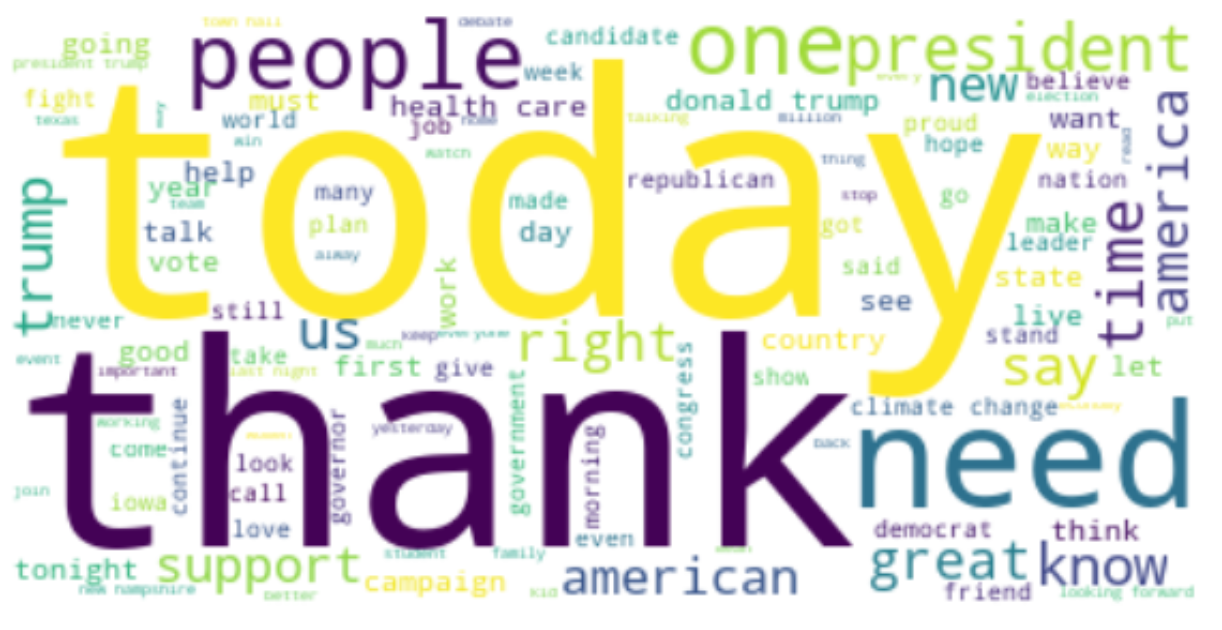


*There were no substantial differences between candidate and non-candidate tweets in terms of overall text statistics in both the observation (i.e. tweet) and corpus level.*

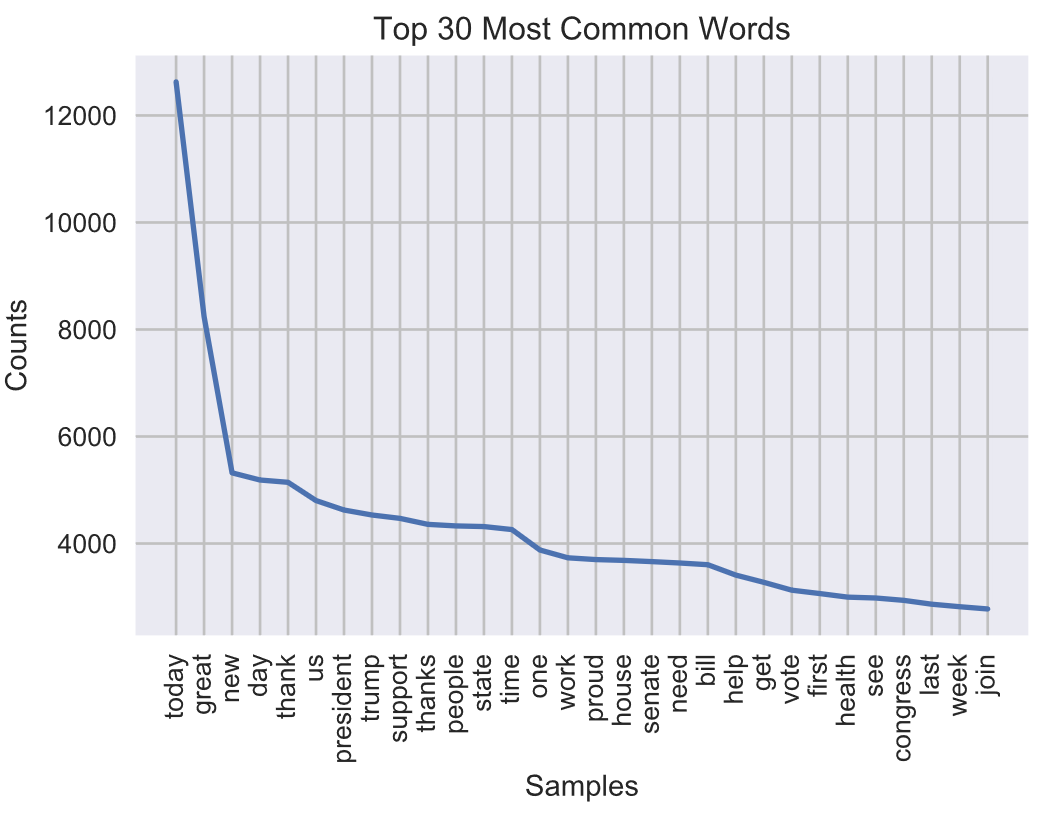
#### Fig.13 Candidate Tweets: Term Frequency



#### Fig.14 Candidate Tweets: Word Cloud



#### Fig.15 Non-Candidate Tweets: Term Frequency



#### Fig.16 Non-Candidate Tweets: Word Cloud



*Term frequency distributions show a stark contrast between candidate and non-candidate tweets. The former shows a far smoother inflection in its term frequency distribution, whereas the latter exhibits a steep drop off in counts within the top three words.*

*Candidate tweets contain the following high-frequency terms in higher proportions than non-candidate tweets:*

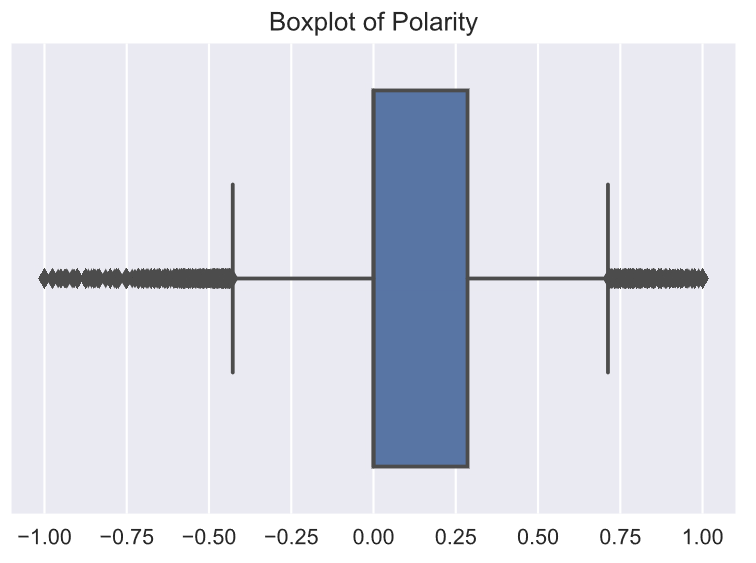
* today | trump | people | president | new | us | need | great

*Non-candidate tweets contain the following words in greater proportion:*

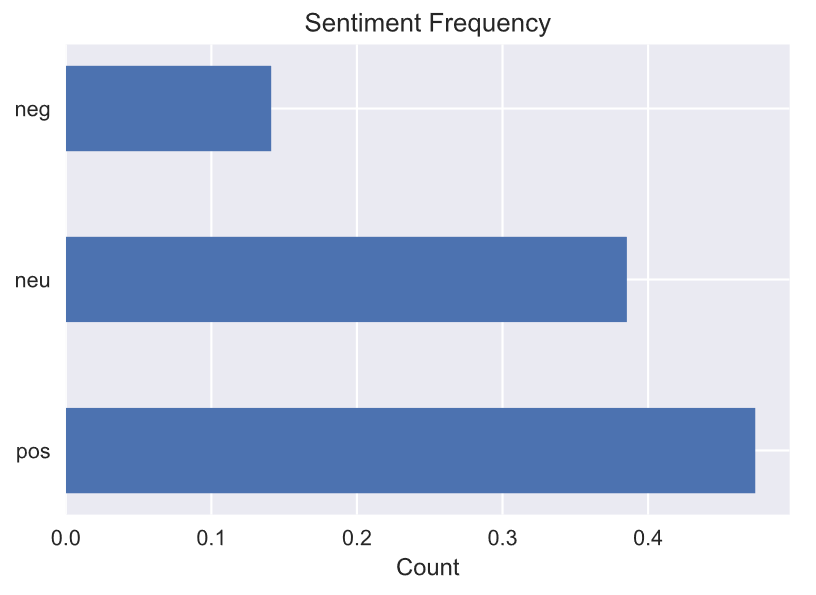
* support | vote | proud

*Very few top terms really distinguished candidates and non-candidates. Similar top terms are used by both. Distinction between classes may come via common function words (i.e. some stopwords) and patterns of less frequent words.*

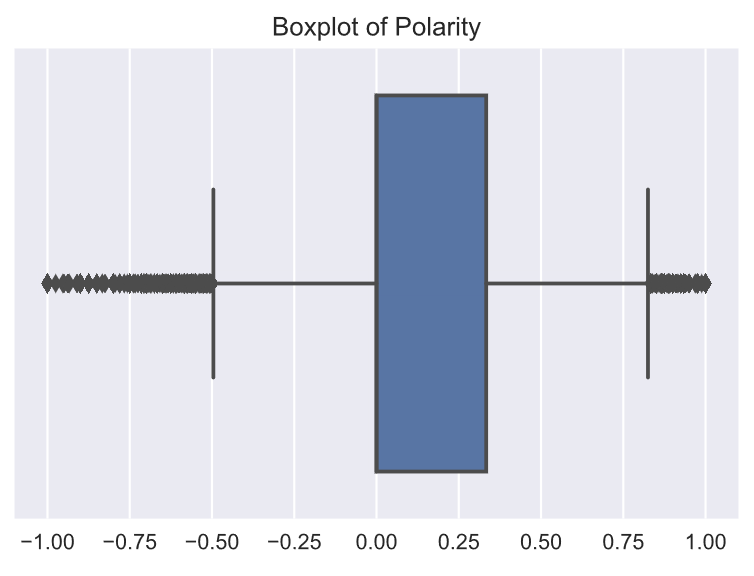
#### Fig.17 Candidate Tweets: Sentiment Polarity



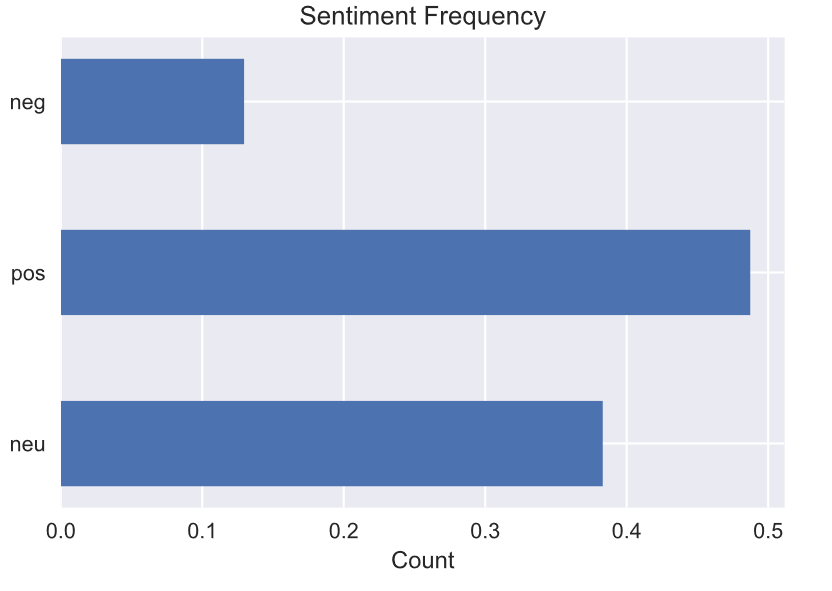
#### Fig.18 Candidate Tweets: Sentiment Proportion



#### Fig.19 Non-Candidate Tweets: Sentiment Polarity

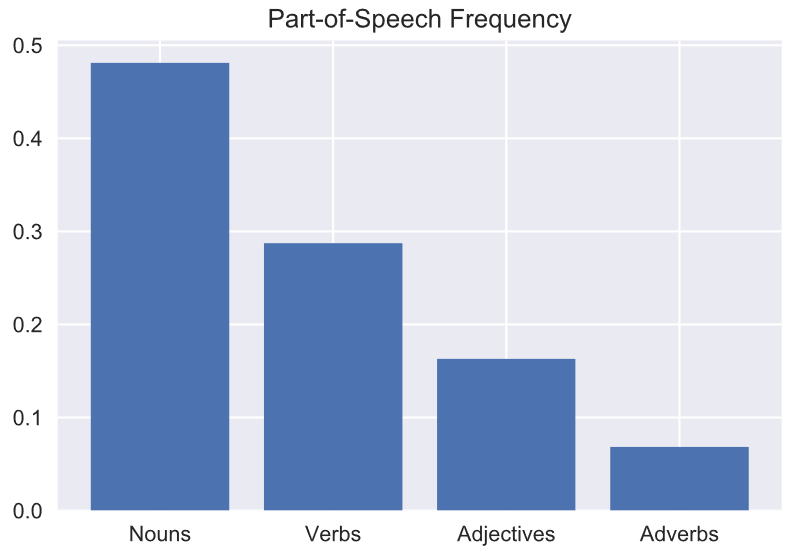


#### Fig.20 Non-Candidate Tweets: Sentiment Proportion

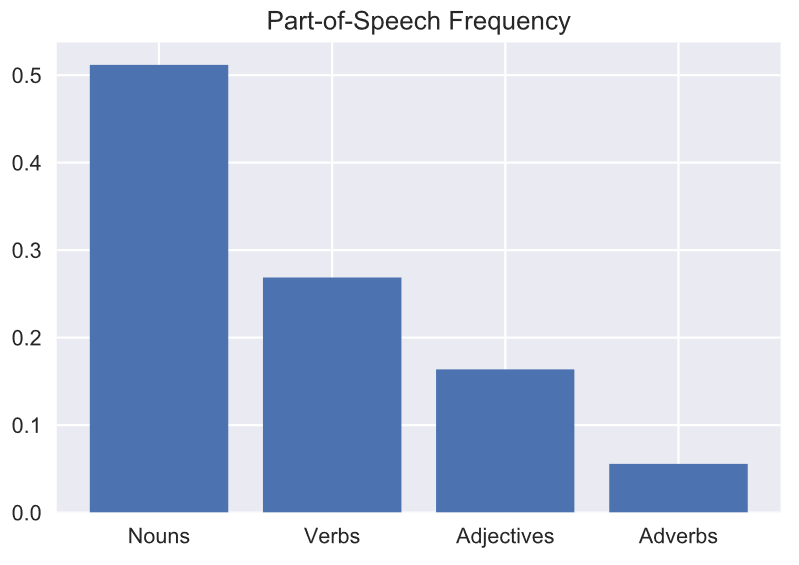


*The distribution of polarity and proportion of sentiment between candidates and non-candidates were similar with little difference. Non-candidates wrote slightly more positive tweets, and candidates wrote slightly more negative tweets. Overall, there was not much difference.*

#### Fig.21 Candidate Tweets: Part-of-Speech



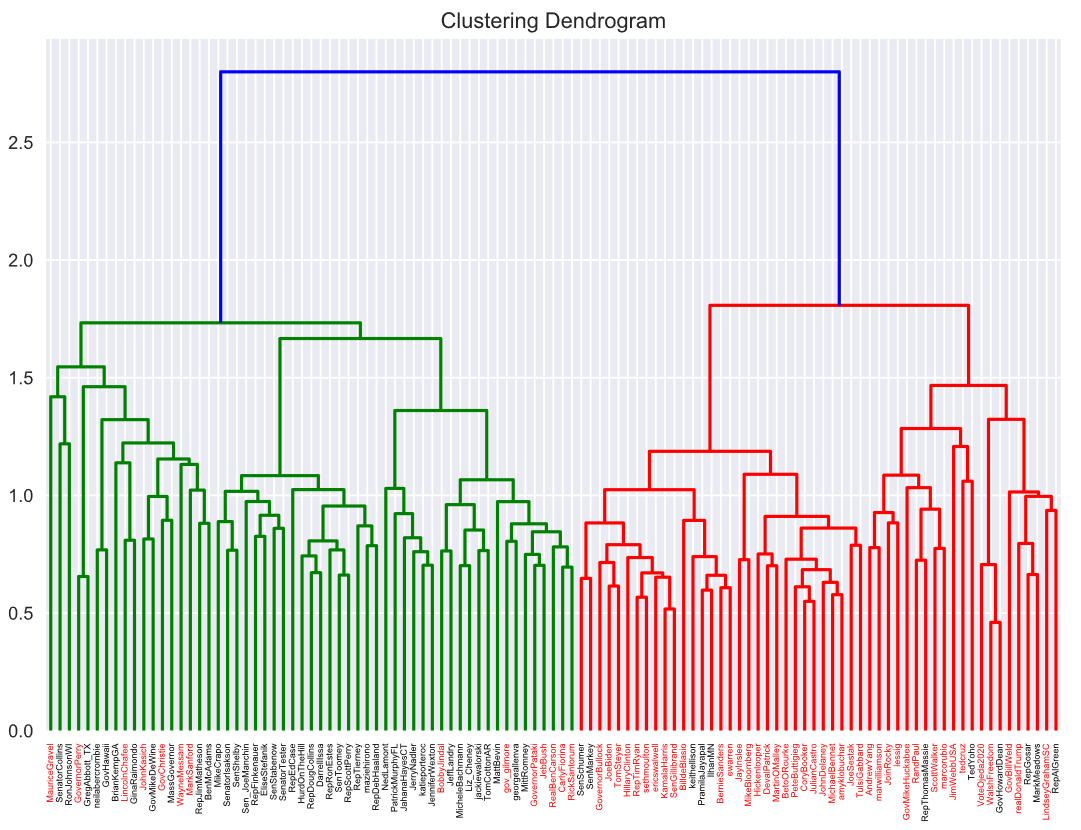
#### Fig.22 Non-Candidate Tweets: Part-of-Speech



*Overall differences in part-of-speech between candidate and non-candidate tweets appeared negligible. Non-candidate tweets contained more nouns and adjectives, and candidate tweets contained more verbs and adverbs.*

At an aggregate and without concern for when tweets were created, do the tweets distinguish candidates and non-candidates?

#### Fig.23 Dendrogram of Users



*From an aggregate perspective, two clusters were distinctly identifiable and tweets from candidates (red labels) were mostly clustered together. Likewise, tweets from non-candidates (black labels) clustered with other non-candidates for the most part. There were a number of exceptions:*

*Non-candidates (i.e. black labels) in the candidates (red) cluster:*

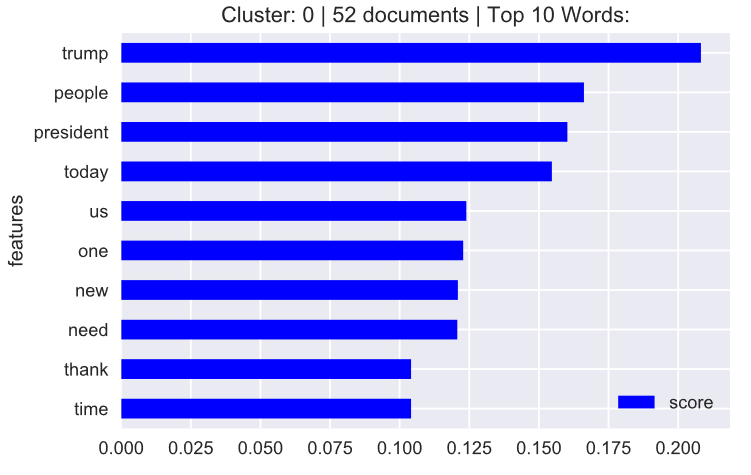
* 1. SenSchumer | SenMarky | keithellison | PramilaJayapal | IlhanMN | RepThomasMassie | TedYoho | GovHowardDean | RepGosar | MarkMeadows | RepAlGreen

*Candidates (i.e. red labels) in the non-candidates (green) cluster:*

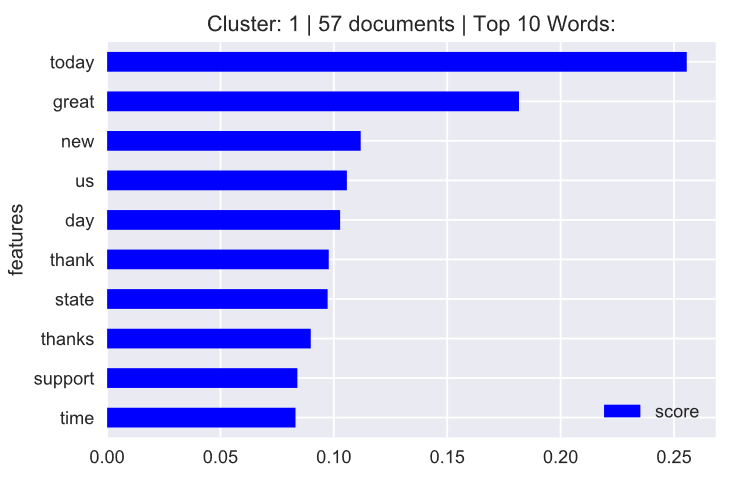
* 1. MauriceGravel | GovernorPerry | LincolnChafee | JohnKasich | GovChristie | WayneMessam | MarkSanford | BobbyJindal | gov\_gilmore | GovernorPataki | JebBush | RealBenCarson | CarlyFiorina | RickSantorum

Top terms in each cluster were identified using mean term frequency.

#### Fig.24 Top Features: Candidates Cluster

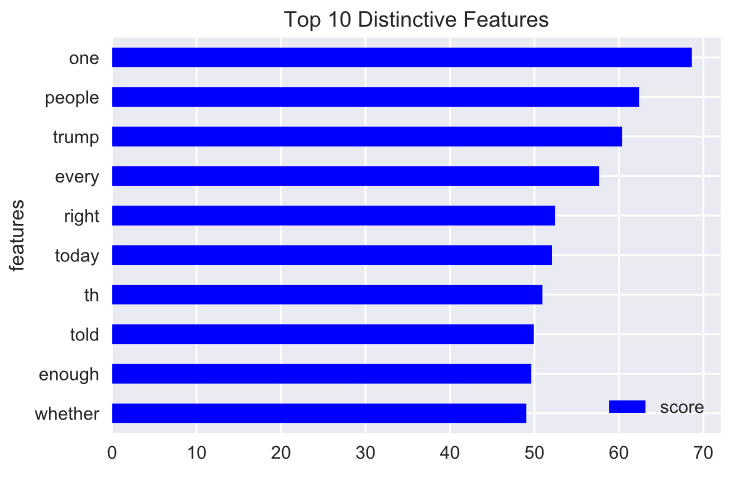


#### Fig.25 Top Features: Non-Candidates Cluster



Top terms that best distinguish clusters were identified using the F-statistic on the mean difference in groups.

#### Fig.26 Top Distinctive Terms for Clustering



### Models

Lorum ipsum.

# Results

Lorum ipsum

# Conclusion

Lorum ipsum.