Twitter Sentiment Clustering of 2016 Presidential Race

Rafael Zamora, Justin Murphey

November 30, 2016

Introduction

This notebook is used to view and analyze Twitter data on the 2016 United States presidential race. Our goal is to find different classes of tweets using the sentiment of the tweet and how much they reference either the Republican or Democrat candidates. These classes of tweets and their sizes will then be used to analyse the change in sentiment over the last few weeks of the election. We hope to see how specific events during the race influence Twitters sentiment to either candidate.

Data was gathered for the 3 weeks prior to the election and 1 week after the election.The data was pulled from Twitter using Python with the following parameters:

Keywords: @hillaryclinton OR #hillaryclinton OR Hillary Clinton OR Hillary OR @RealDonaldTrump OR #donaldtrump OR Donald Trump OR Trump

Start Date: 2016-10-16

End Date: 2016-11-14

The following values were gathered for each tweet:

Author-ID

Date with Time

Text

The data is stored by day located in the data/raw/ folder. It was then processed to find the sentiment value and candidate reference value for each tweet using NLTK on Python. The processed data is stored by day located in the data/process/ folder.

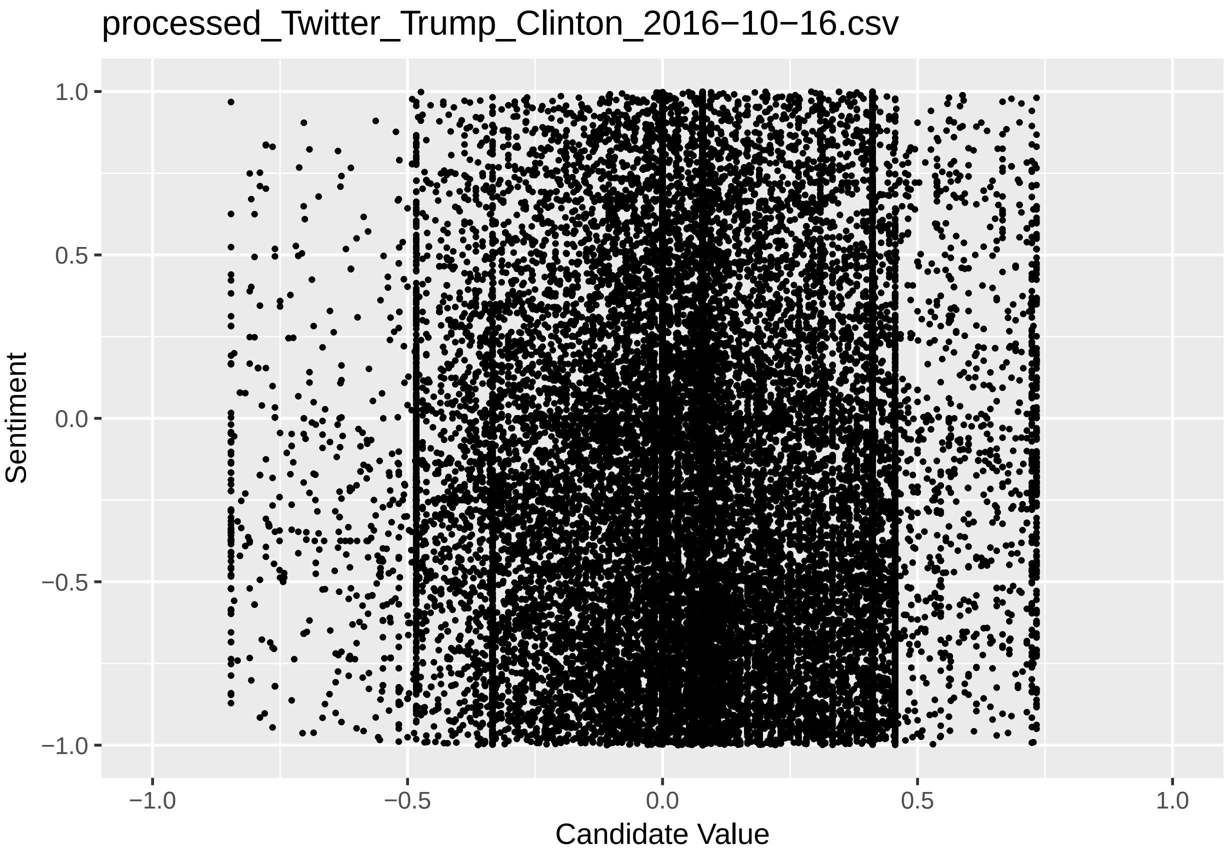
Graphs created by this notebook can be found in the doc/figures folder.

Processed Data

The following is used to display the processed Twitter data. To view a specific data file change filename to desired file located in data/processed/ folder. Uncomment for Pdf version of graph will be saved with desired pdf_filename.

This example will be showing the processed Twitter data from October 16, 2016.

```
filename = "processed_Twitter_Trump_Clinton_2016-10-16.csv"
pdf_filename = "figures/processed/processed_Twitter_Trump_Clinton_2016-10-16.pdf"
data = read.csv(file=paste(path_to_data,filename,sep=""), head=FALSE, sep=",")
colnames(data)[1] = 'Sentiment'
colnames(data)[2] = 'Candidate_Value'
#pdf(paste(path_to_doc, pdf_filename,sep=""),
#width=11,height=8.5,paper='special')
qplot(data$Candidate_Value, data$Sentiment, main=filename, xlab="Candidate Value",
ylab="Sentiment", size=I(.5), xlim = c(-1,1),ylim = c(-1,1))
```



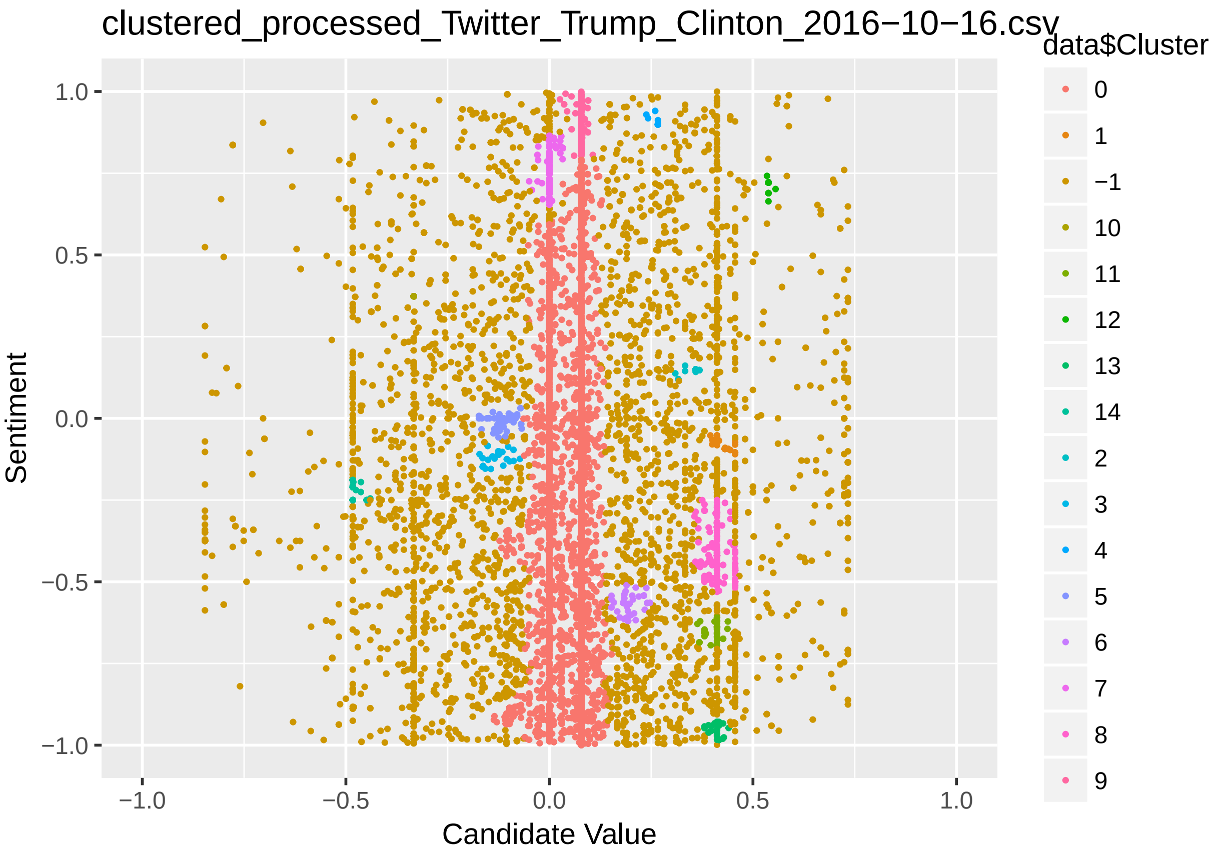
```
#dev.off()
```

Clustered Data

The following is used to display the clustered Twitter data. To view a specific data file change filename to desired file located in results/ folder. In the results folder, results.txt includes the eps and min sample values used for the DBSCAN clustering. Uncomment for Pdf version of graph will be saved with desired pdf_filename.

This example will be showing the clustered Twitter data from October 16, 2016.

```
filename = "clustered_processed_Twitter_Trump_Clinton_2016-10-16.csv"
pdf_filename = "figures/clustered/clustered_processed_Twitter_Trump_Clinton_2016-10-16.pdf"
data = read.csv(file=paste(path_to_results,filename,sep=""), head=FALSE, sep=",")
colnames(data)[1] = 'Sentiment'
colnames(data)[2] = 'Candidate_Value'
colnames(data)[3] = 'Cluster'
data$Cluster = as.character(data$Cluster)
#pdf(paste(path_to_doc, pdf_filename,sep=""),
#width=11,height=8.5,paper='special')
qplot(data$Candidate_Value, data$Sentiment, colour=data$Cluster, main=filename,
xlab="Candidate Value", ylab="Sentiment", size=I(.5), xlim = c(-1,1),ylim = c(-1,1))
```



```
#dev.off()
tapply(data$Cluster,data$Cluster,length)
```

```
##    0    1   -1   10   11   12   13   14    2    3    4    5    6    7    8
## 5181   42 3456   43   65   69   36   51   53  154   37   86   69  298  233
##    9
##  127
```

Analysis