

Text Analysis of Political Subreddits: The Trump/Clinton Dichotomy

Senan Hogan-H., Shirley Jiang

14 December 2017

Introduction

Reddit is an online website for discussion, content aggregation and rating. The platform is divided to thousands of subreddits, which are communities built by users for a specific topic. Users have created subreddits for uncountably many diverse topics, from following professional sports, such as the NFL, to pictures of puppies. Over the last two years, reddit has come under scrutiny from the news media over its communities that have support for (now President) Donald Trump and the alt-right, prompting the Reddit staff to track down and delete communities devoted to the alt-right.

The site has gone through a tough process of finding the middle ground between the free expression of its users' with their community building and the condemnation of racist or hateful sentiments associated with the alt-right and surrounding political figures. We're interested in comparing the use of words in comments in the official candidate subreddit for Trump to those for the Clinton subreddit, to see whether there are differences in word usage, and differences by sentiment analysis. Lastly, we create a classification model to classify comments as belonging in the Trump or Clinton subreddit, to be applied to comments from a neutral subreddit, seeing the relative prevalence of political support in that neutral subreddit.

The Combined Dataset

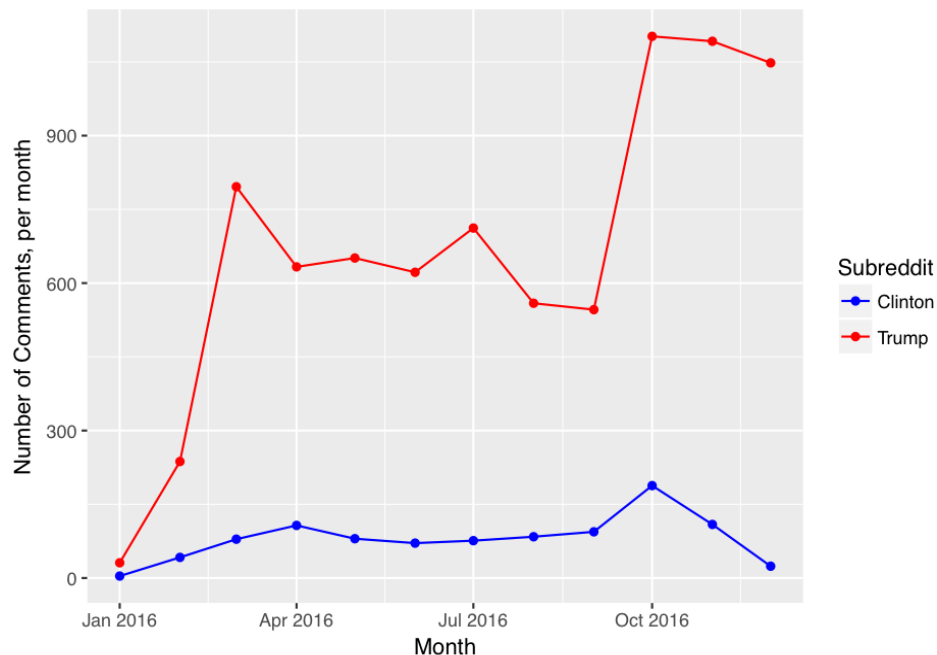
We chose to work with all the text data from every single comment on the official 2016 presidential candidate subreddits. The official Donald Trump support subreddit is called "r/The_Donald" and was created in 2015 following Trump's plan to run in the 2016 election; the official Hillary Clinton is called "r/hillaryclinton" was also created in 2015 following Clinton's plan to run in the 2016 election. Originally, we used a Python scraper to pull comments using the Reddit API; however, the Reddit API can only pull 100 comments at a time. We decided to find an alternative and more efficient way to collect the dataset.

Google Bigquery hosts a full repository of data about Reddit, including a near complete history of over 3 billion comments across the entire site. We pulled every single comment for the respective subreddits using SQL commands, by filtering on the relevant subreddit titles in the year 2016. The resulting monthly files were merged together, to create a huge .csv file storing every comment for both the subreddits. The data set's observations are each comments posted on the website, with variables 'body' for the text of the comment, 'score' for users' ratings of the comment, 'controversiality' for whether reddit counts the comment as controversial, 'month' for month the comment was posted in 2016 and 'subreddit' for whether it was posted in the Trump or Clinton subreddit.

Note: the .csv files used are hosted at the Google cloud links used in the code, for anyone to access.

First load the dataset, and apply some routine cleaning steps.

The dataset is extremely large, with around 9 million comments. Since the data set is extremely large, we ran the analysis on a random subset of 1 million observations. It should be noted that the Trump subreddit was around 5 times as large as the Clinton subreddit (by comment frequency). We can see this in the following graph which shows the amount of comments per month. We see a sharp spike in comment frequency for Hillary and Trump starting in October, the month right before the November 2 election date. The Trump comments increased by about 33% in October compared to the previous month and continued to stay at the same level for the next two months.



Below is one of the most ‘upvoted’ post on the Clinton campaign subreddit, to show an example of one of the most popular posts.

24

Did Trump win because of racism and xenophobia?

(self.hillaryclinton)

submitted 1 year ago by vekul1

Black Lives Matter

Hillary Clinton won all three debates handily and the "grab them by the pussy" tape should have lost Trump the votes of women and religious conservatives. The Trump camp also did some idiotic things, such as plagiarizing Michelle Obama's speech for the RNC. The only way I can rationalize Trump's victory is that all that mattered to his supporters was that Trump would stick it to brown and black people.

35 comments source share save hide give gold report crosspost hide all child comments

All 35 Comments

Subscribe

sorted by: top

navigate by: submitter | moderator | friend | me | admin | highlighted | gilded | IAmA | images | videos | popular | new

Thy_Lord_Castiel

26 points

1 year ago


The thought that anyone who disliked Obama was racist, and if you didn't vote for Hillary you hated women, has pushed people away. I live in rural america. We don't care your skin color or your sexual choice. We care that for 8 years you left wingers called us racist, xenophobic hicks.


Then you called Trump the same thing. You created the link when there wasn't one.

permalink source embed save save-RES give gold hide child comments

Figure 1: A top post on r/hillaryclinton, November 2016

Below is the most ‘upvoted’ post on the Trump campaign subreddit, where the man himself (or perhaps some of his campaign staff) answered some questions for interested parties. There was some controversy surrounding this event on Reddit at the time, and as far as we are aware only posts pre-vetted by the site’s moderators were answered by Trump (or his staff).


42.1k



I'm Donald J. Trump and I'm Your Next President of the United States. (self:The_Donald)

submitted 1 year ago · (last edited 1 year ago) by the-realDonaldTrump · 45th · 115

Hello The_Donald readers and the entire Reddit community -- this is going to be SO huge and I'm looking forward to answering your questions. I'm doing this in flight to visit the great people of Toledo, OH, so Internet connection might be spotty -- I promise you, I'll answer all the questions I can. I want to do BIG things for America and as your President, I WILL Make America Great Again! Be back in 30 -- 7 pm ET!

UPDATE: Proof: <https://www.facebook.com/DonaldTrump/posts/10157382886305725>^[1]

Looking forward to answering your questions! We are still in the air on our way to Ohio.

Such a great time answering your questions. Thank You!



<https://www.facebook.com/DonaldTrump/photos/a.488852220724.393301.153080620724/10157383302255725/?type=3&theater>^[2]

21267 comments · source · share · save · hide · give gold · deport · crosspost · hide all child comments

top 200 comments · show 500

sorted by: q&a (suggested) ▼

navigate by: submitter | moderator | friend | me | admin | highlighted | gilded | IAmA | images | videos | popular | new

[-] trexroar · PA · 2414 points · 1 year ago

Hello Mr. Trump. I'm 29 years old, registered Libertarian and have voted that way my entire life. I feel closely aligned with what the party stands for, and am passionate about it. However, this election I am starting to think that, while I think a Libertarian candidate is what's best for America long term, that you may be what America needs right now. I also feel like you care about this country so, so much, and you want what's best for it. This is a difficult decision for me, because on the issues I seem to either firmly believe or strongly disagree with what you have to say. I think you may be the best possible candidate for our economy, but I do fear some of your international policies.

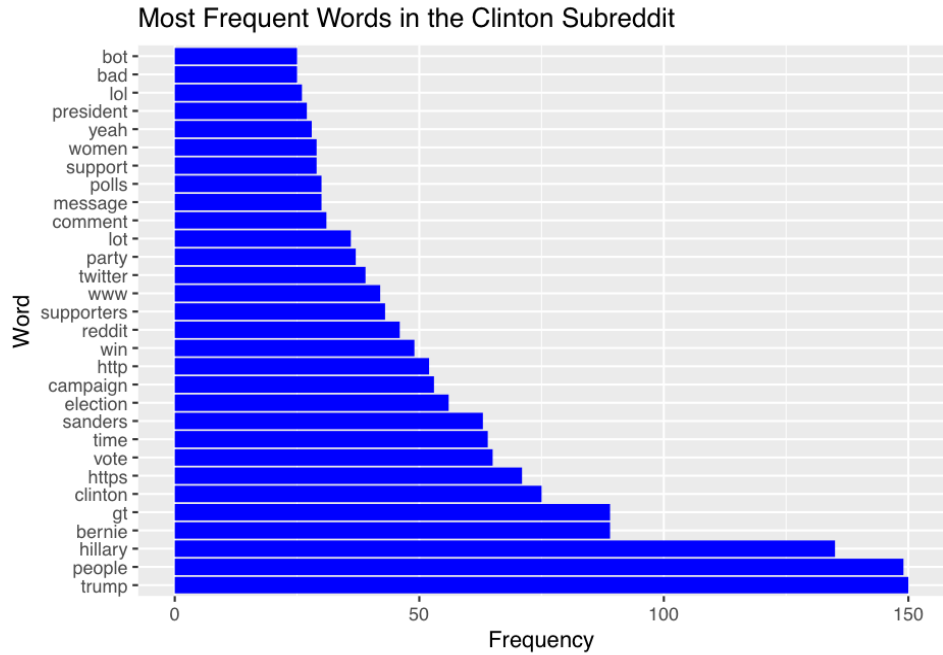
I know that I am not the only 3rd party voter that feels this way. So my question to you is, what do you say to people like me who are on the fence about voting 3rd party(Johnson/Stein) or for you?

Thank you very much for doing this AMA.

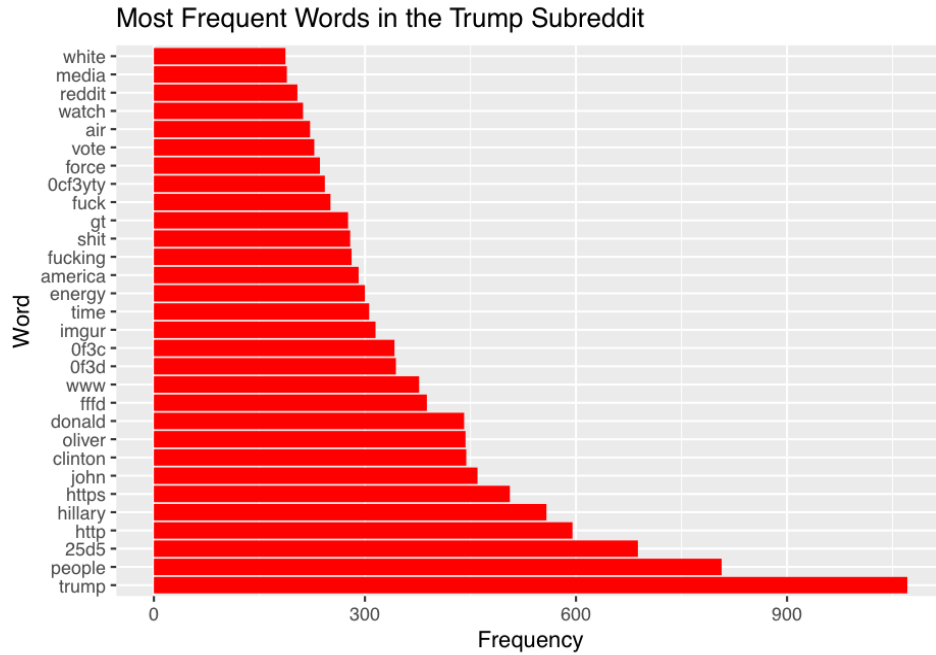
Figure 2: A top post on r/the_Donald, November 2016

Word Use Comparison

The first question to ask is how use of words compare in the subreddits. Following are a set of bargraphs to show the 30 most frequent words in the respective subreddits.



Trump is one of the most frequent word seen in the Clinton subreddit, which implies that the topic of Trump was the most spoken about. Most of the other words are not unexpected, words such as Bernie and Sanders (Clinton's primary challenger) and a few words relating to the campaign process, like supporters, people and campaign. Interestingly, there are a few examples of words left over from people providing web links to other pages, https (which begins a url link), www (the same), and mention of twitter. This implies one of most common forms of commenting on the Clinton subreddit is linking to an external source, including twitter.



For the Trump subreddit the most frequent words include: Islam, swear words, and derogatory words. Many of the words may be associated with anger against the Clinton campaign, mirroring the words of candidate Trump on the election trail. The Trump subreddit also had a very surprising amount of the collection of characters '0cf3tyY'. This turns out to be the remanant of a link, <https://i.imgur.com/0cf3tyY.jpg>, of a fake picture that Trump tweeted that implies Hillary Clinton is allied with a KKK supporter.

What about words that increase in frequency with respect to the other subreddit?

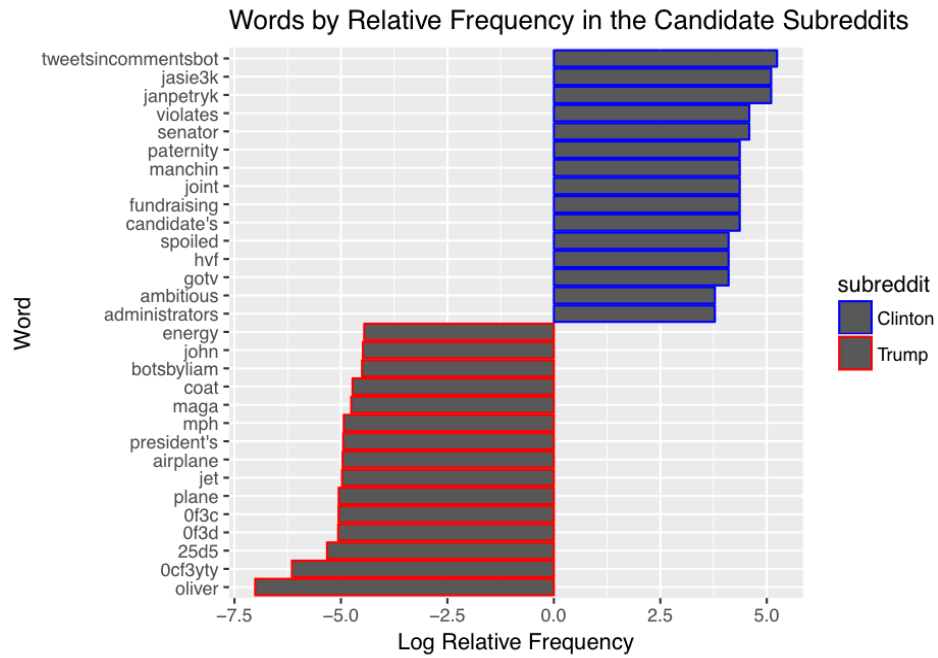
Here we look at log odds ratio comparison of most common words in the subreddits, where the log ration is given by:

$$R = \log_2 \left(\frac{\frac{\text{Frequency for Specific Word in Clinton subreddit}}{\text{Total Frequency for Words in Clinton subreddit}}}{\frac{\text{Frequency for Specific Word in Trump subreddit}}{\text{Total Frequency for Words in Trump subreddit}}} \right)$$

This gives us a way of looking at which words are overrepresented in each subreddit.



Figure 3: The picture so commonly linked to on the Trump subreddit.



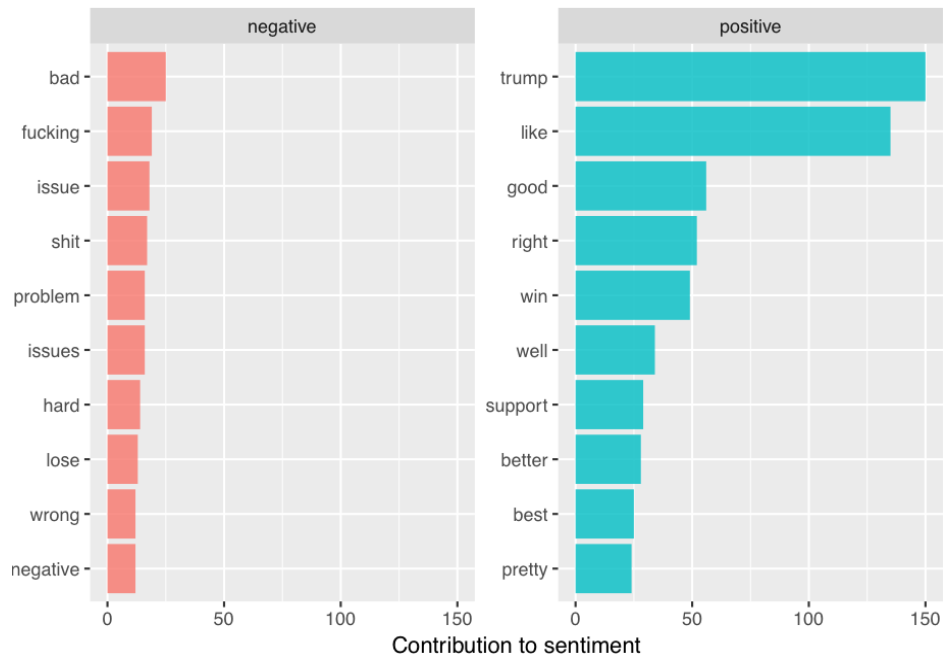
The Clinton subreddit had many, many more instances of the 'tweetsincommentsbot,' which is a reply to comments that link to a tweet. This implies the comments in the Clinton subreddit link to tweets much more

often than Trump commenters. Links to the statistical analysis site 538 are also much more frequent in the Clinton subreddit, as are words that may be commonly used to describe candidate Trump by Clinton supporters (for example emotional). On the other hand, Trump commenters are much more likely to mention ISIS, and the ingur link mentioned before, and (seemingly) words describing animals.

Sentiment Analysis

In order to gain a view of the overall attitudes (sentiments) being expressed in the subreddits, we conducted a sentiment analysis. We used the Bing lexicon for our sentiment analysis. The Bing lexicon categorizes words into two sentiment categories, positive and negative. The following is a sentiment comparison between the Clinton subreddit and Trump subreddit.

Sentiment Analysis Clinton

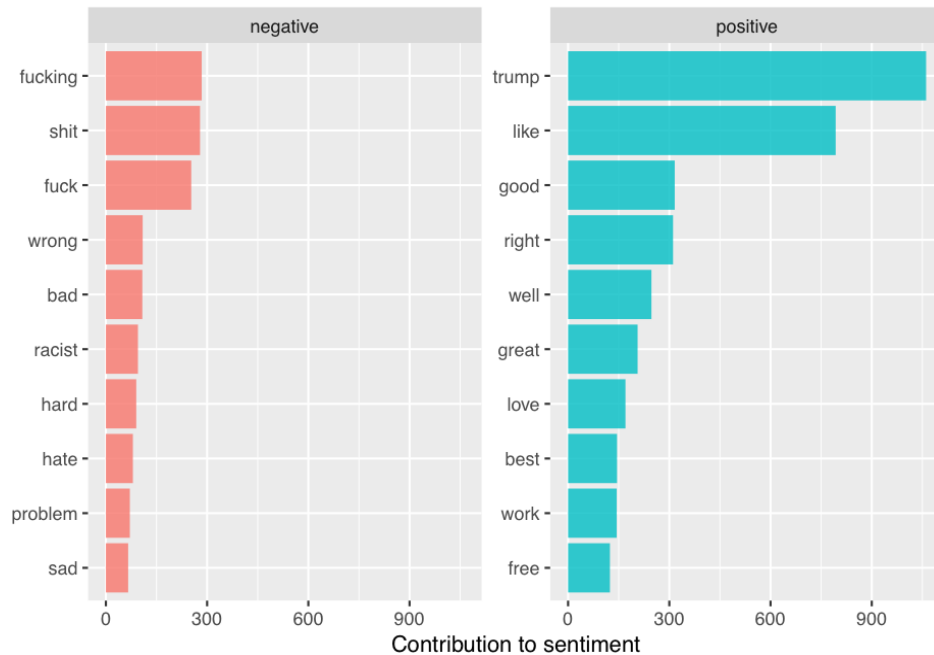


```
## # A tibble: 10 x 2
##   sentiment    n
##   <chr> <int>
## 1  negative 3986
## 2  positive 3455
## 3    trust 2011
## 4    fear 1837
## 5    anger 1654
## 6  sadness 1546
## 7 anticipation 1423
## 8    disgust 1258
```

```
## 9      joy 1111
## 10     surprise 978
```

Looking at the Clinton sentiment analysis, there is no clear word leader in negative sentiments or positive sentiments. The word Trump appears the most frequently for positive sentiments. However, it cannot be determined if the word Trump is being used as a verb or the candidate's name. One would assume it is most frequently being used as a proper noun, thus one cannot really make any conclusions about the positive sentiment categorization of the word Trump. The overall sentiment analysis implies there is no general divide in negative or positive comments, instead the comments are in general pretty varied in their sentiments. Interestingly, the word pretty is a sentiment expressed in the Clinton comments. Thus, there is evidence of female words used in the Clinton comments.

Sentiment Analysis Trump



```
## # A tibble: 10 x 2
##   sentiment    n
##   <chr> <int>
## 1 positive 10700
## 2 negative 10464
## 3 trust    6856
## 4 fear     5513
## 5 anticipation 4992
## 6 anger     4923
## 7 sadness  4101
## 8 joy      4071
## 9 disgust  3836
```



```
## 10      surprise  3446
```

For Trump, the negative sentiment words include: fake, liar, expletive, racist, evil. It also includes many expletives, which we did not find in the Clinton sentiment analysis. The positive sentiment words include: Trump, love and right. Far and away the leader for negative sentiment is the word liar, which Trump himself used many times to describe candidate Clinton on the election trail. While there are no clear leaders in the positive Trump sentiment analysis either, the words appear to have stronger connotations (for positive and negative sentiments) than the words found in the Clinton sentiment analysis.

Naive Bayesian Classifier

The Naive Bayesian Classifier is a classification model that uses the Bayes theorem (with a strong independence assumption between variables) to classify observations. The classifier may be applied to text data by counting the presence of words in a given feature. Here we train a Naive Bayesian Classifier to the Trump/Clinton comments dataset, to form a model that can classify comments as belonging in the Trump or Clinton subreddit. We build a classification model so that we may be able to bring in a new data set of other comments from another subreddit and consider how many of those comments belong in the Trump or Clinton subreddit.

R/politics is the general subreddit for political discussion and posts on reddit, and it underwent a fair amount of turmoil surrounding the November 2016 election results. The subreddit was counted as biased against Trump supporters happy with the election results, and driving them out by banning them and deleting their comments. The Naive Bayesian Classifier is applied to a document matrix of all comments from r/politics in November 2016, showing the amount of comments which are likely to be aligned with either subreddit. In doing so, we can see whether users and their comments who align with Clinton vastly outnumber those who support Trump in r/politics subreddit following the election.

```
## [1] 197
```

```
##      usekernel fL adjust Accuracy  Kappa AccuracySD KappaSD
## 1      FALSE  0      1  0.8567 0.1347    0.00915 0.04524
## 2       TRUE  0      1  0.8567 0.1347    0.00915 0.04524
```

The model has accuracy of around 86%, according to error reported by cross validation. The r/politics data set for November was obtained by the same methods as before, and as applied to the Bayes classifier below. The output is the amount of comments predicted to be aligned with either the Trump or Clinton subreddit.

```
## predictions
## Clinton  Trump
##      112    888
```

The vast majority of comments made (and not deleted) in the r/politics subreddit in November 2016 are predicted to be aligned with the Trump subreddit. This is counter to the narrative that r/politics didn't allow Trump users in the subreddit following Trump's election, and instead it would imply that the majority of users in r/politics in November are aligned with Trump regardless.

Conclusions

The steps taken here have conducted statistical analysis to text data of all comments in 2016 in two datasets combined. We've shown the most frequent words used in each respective, as well as the relatively more common words. Lastly, we've trained a model for text data classification on the data set, before bringing in a supposedly neutral data set to apply the model to for a real application of the statistical analysis.

In the future, the analysis would be more complete if conducted on the entire dataset. As of now, the analysis were conducted on very large subsets of the relevant datasets so that the computing power we have access to can work with the data. By sheer sample size, the analysis will accurately predict a analysis for the entire

dataset, yet would be more complete if run with every observation. The code included in this post may be edited only slightly, by removing the subsample commands, to be applied to the very very large .csv files for comments if the user has access to the computing power needed in R (RAM of around 10GB to host the needed objects).