Jenny Ackerman

DSCI415 Final Project

12/12/19

**Background on the datasets**

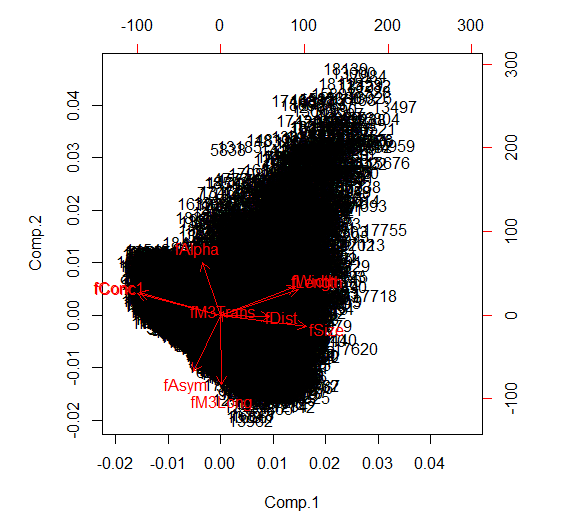
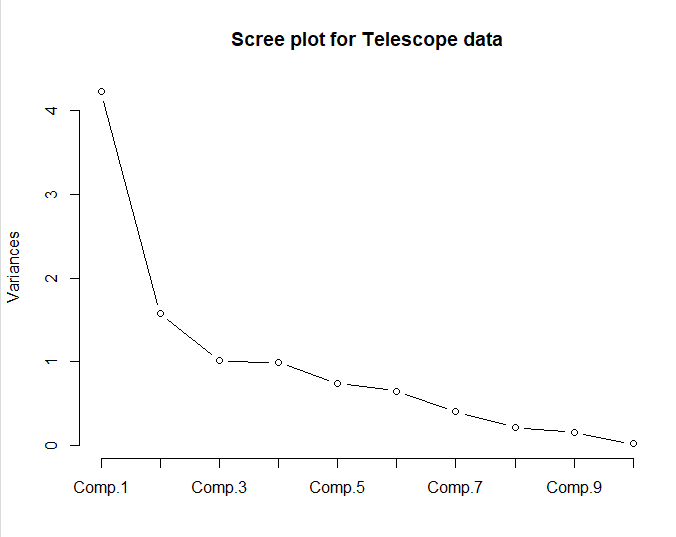
I used three different datasets I used a dataset from the Kaggle website and used two datasets pulled from twitter. The Kaggle data that I used was a Magic Gamma telescope data from 2004. This is a was generated by a Monte Carlo program, to simulate extensive air showers of high energy gamma particles in a ground-based atmospheric. There are different patterns a primary gamma signal and the background signals.

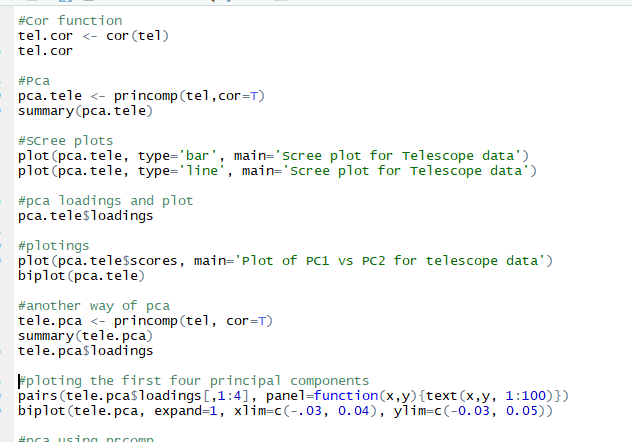
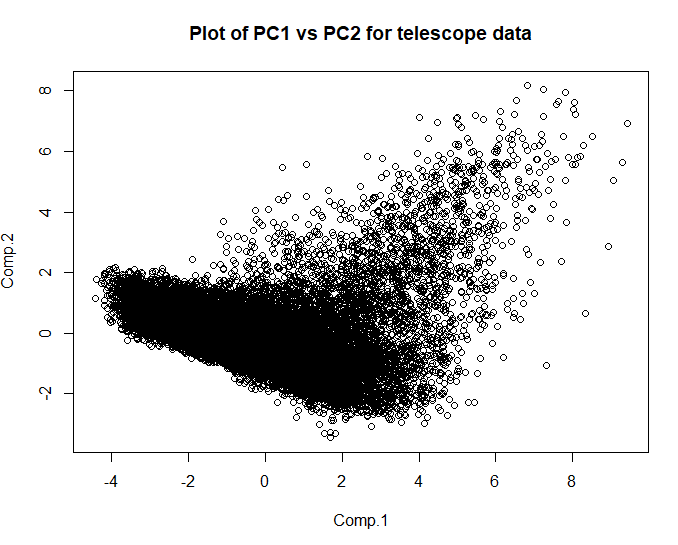
The twitter data was pulled by Deppa, I decided to use the trending hashtags of #LaughingStockoftheWorld and #Trumpisalaughingstock for my analysis. These two dataset were really similar so I combined both datasets into one.

**Magic Gamma Telescope data 2004 analysis**

I started with the telescope data which had around 19020 observations with 12 variables that included a class variable of either gamma samples or background samples. With this dataset I performed a PCA, started ICA and MSD, and did a particle cluster analysis.

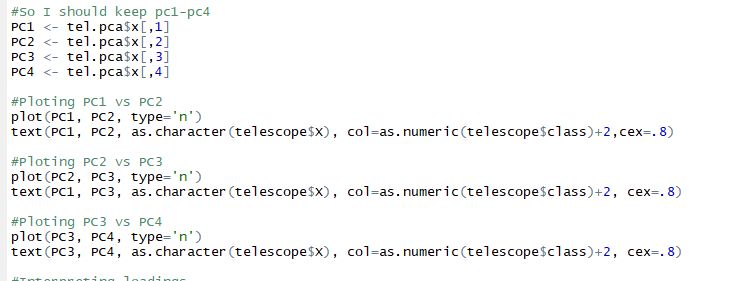
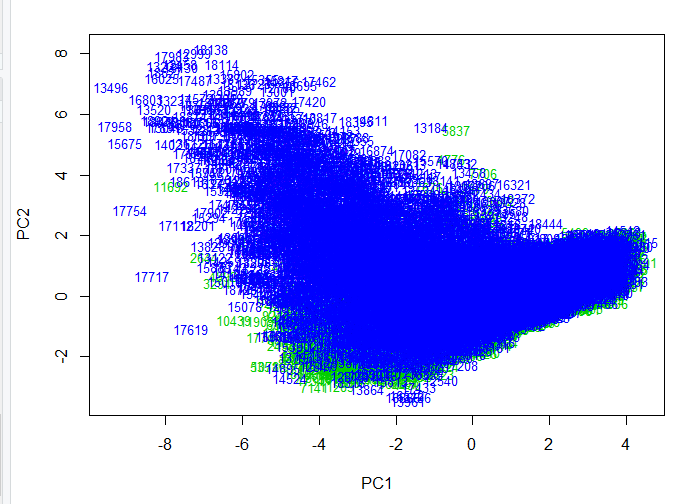
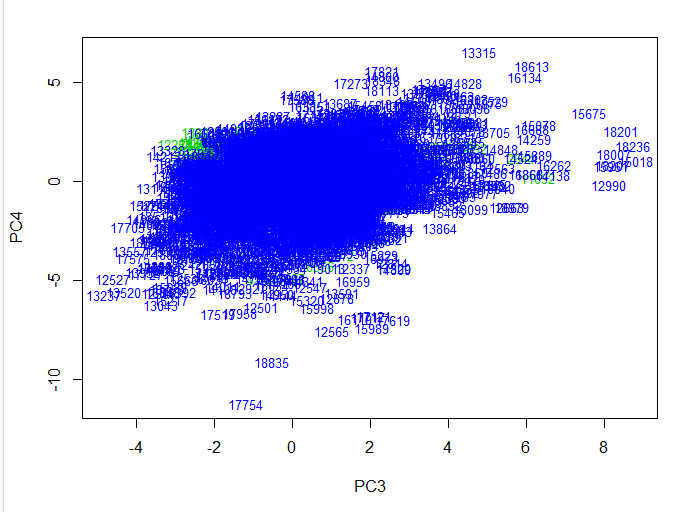
With the PCA I found that the best component to use was up to four components based on the scree plots and loadings. The graphs of the initial pc1 vs pc2 plot using princomp and it was really messy and didn’t really show anything useful, same with the initial biplot of the data.



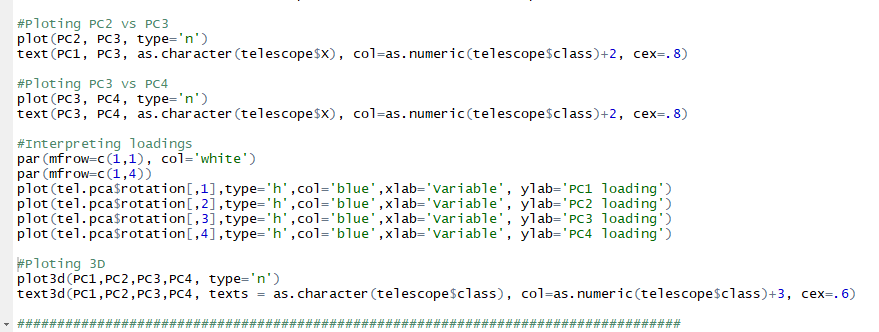


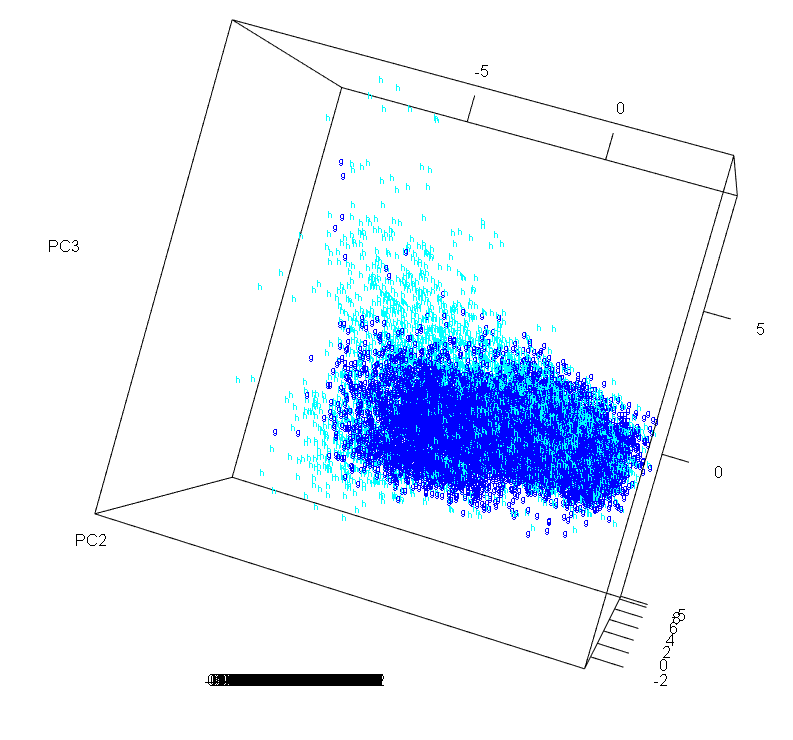
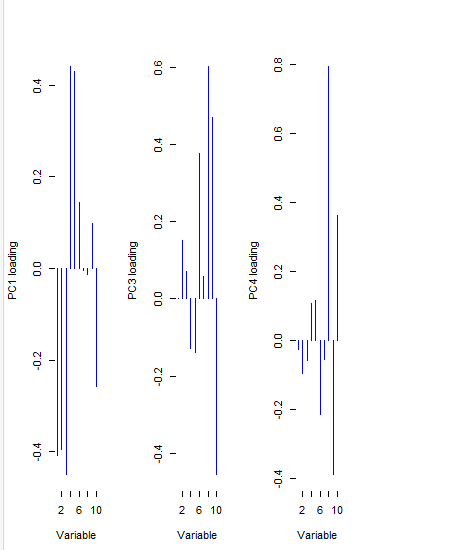
Then I plotted the first PCA to each others PC1 vs PC2 and PC3 vs PC4

These plots are always very unclear in that they don’t show a lot



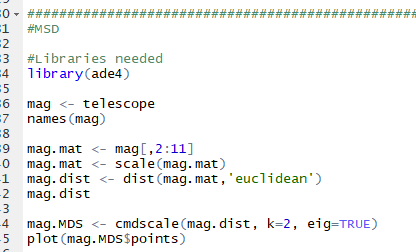
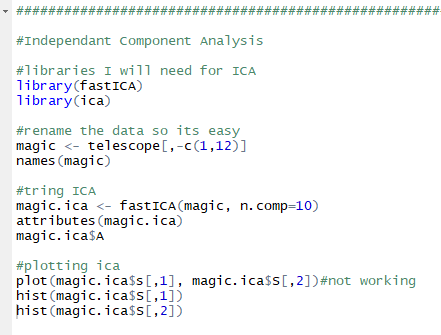
I also plotted the first four loadings and looked at them and they didn’t really show anything interesting



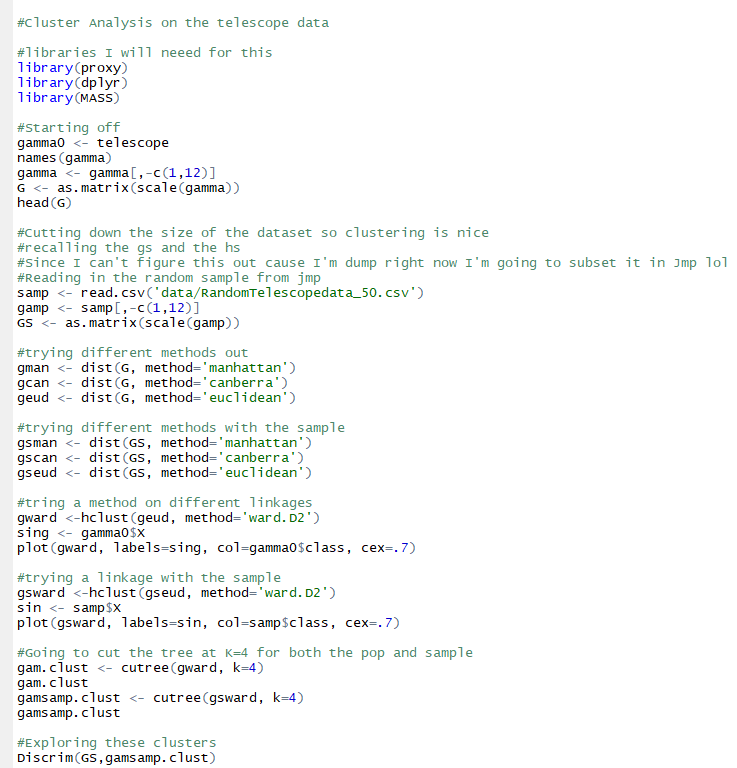


The 3d plot shows a little more distinction between the two classes given in the dataset but not really.

Then I started a ICA and MDS and I didn’t have enough time to finish, so here is my code for it

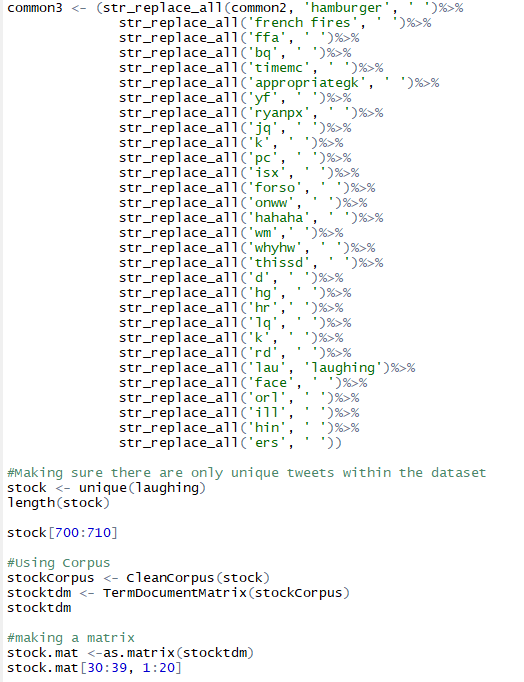
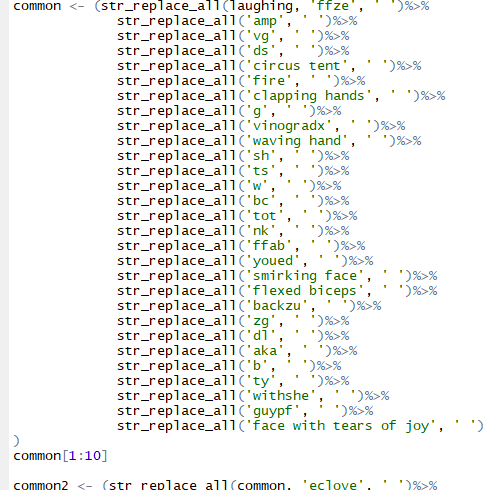
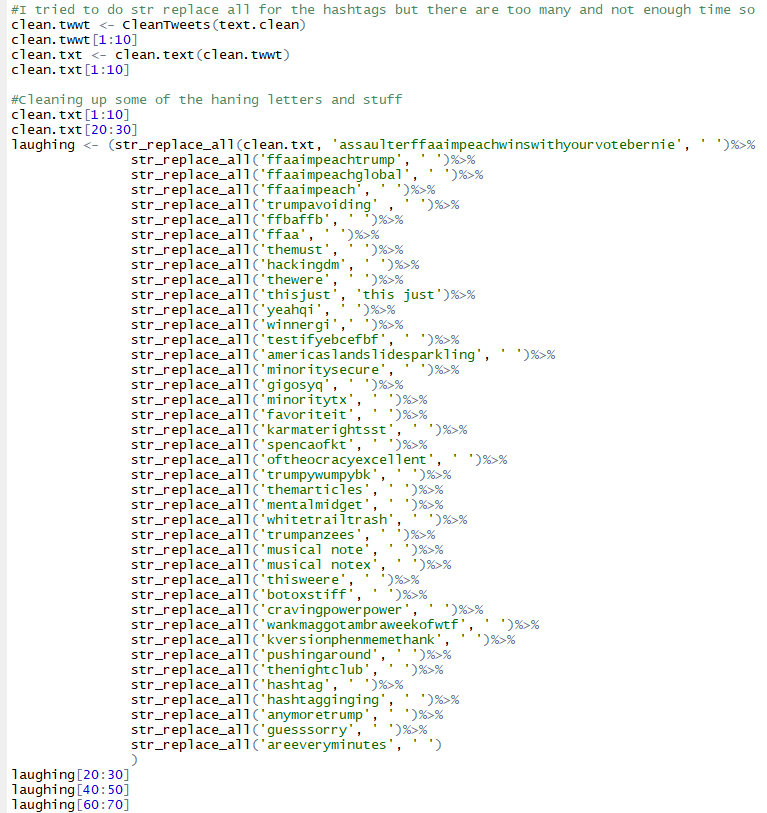
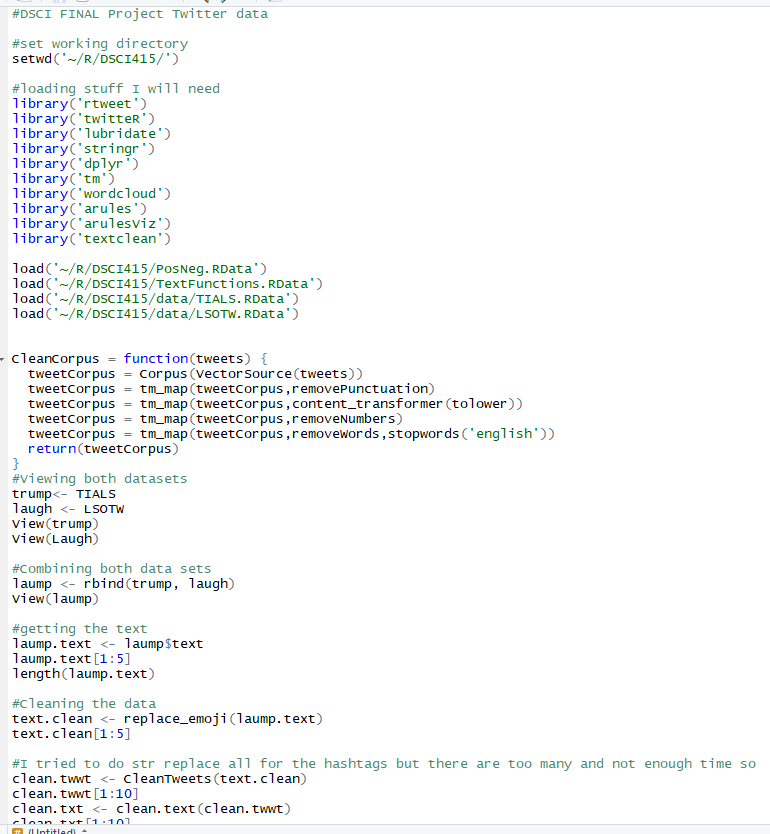


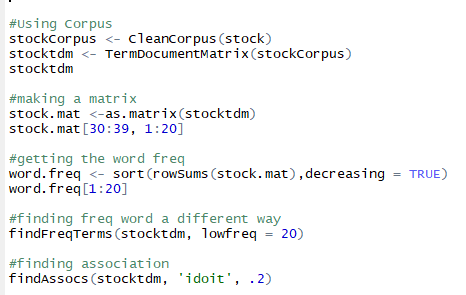
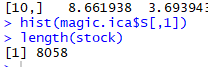
Finally I started a cluster analysis one the data and it worked only one time, and it won’t run since, that’s because the data set is so large but here is my code, I did end up trying to use a random sample but I didn’t get it completed in time for the presentation



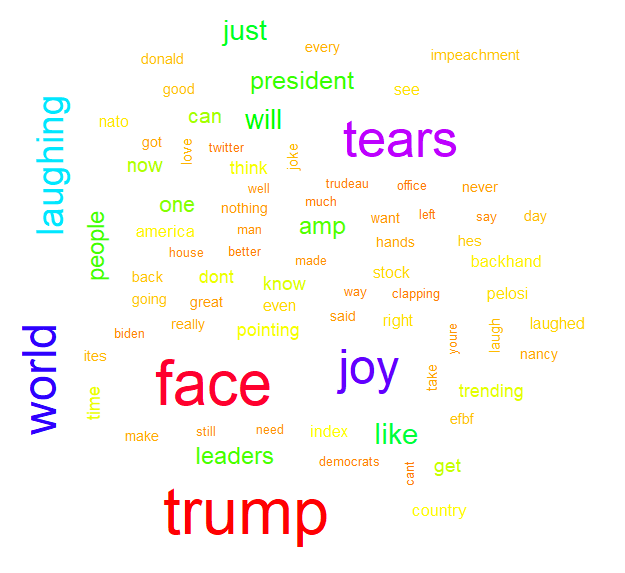
**Twitter data sets**

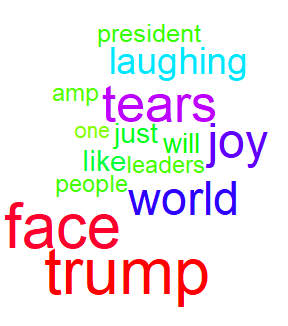
So I combined both datasets and clean them the best I could but clean corpus and the stop words wasn’t really working for me and there I had to create many str replace all functions that took a lot of time. I cleaned and cleaned the data

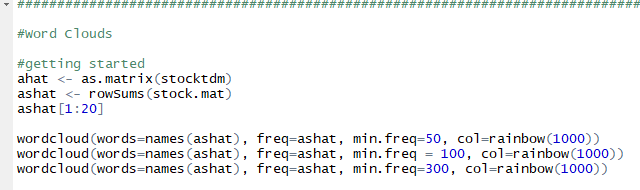




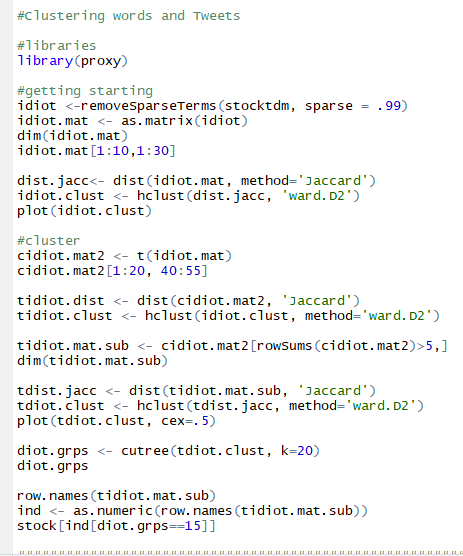
Next I started out using word clouds and seeing what those told, I created three wordclouds

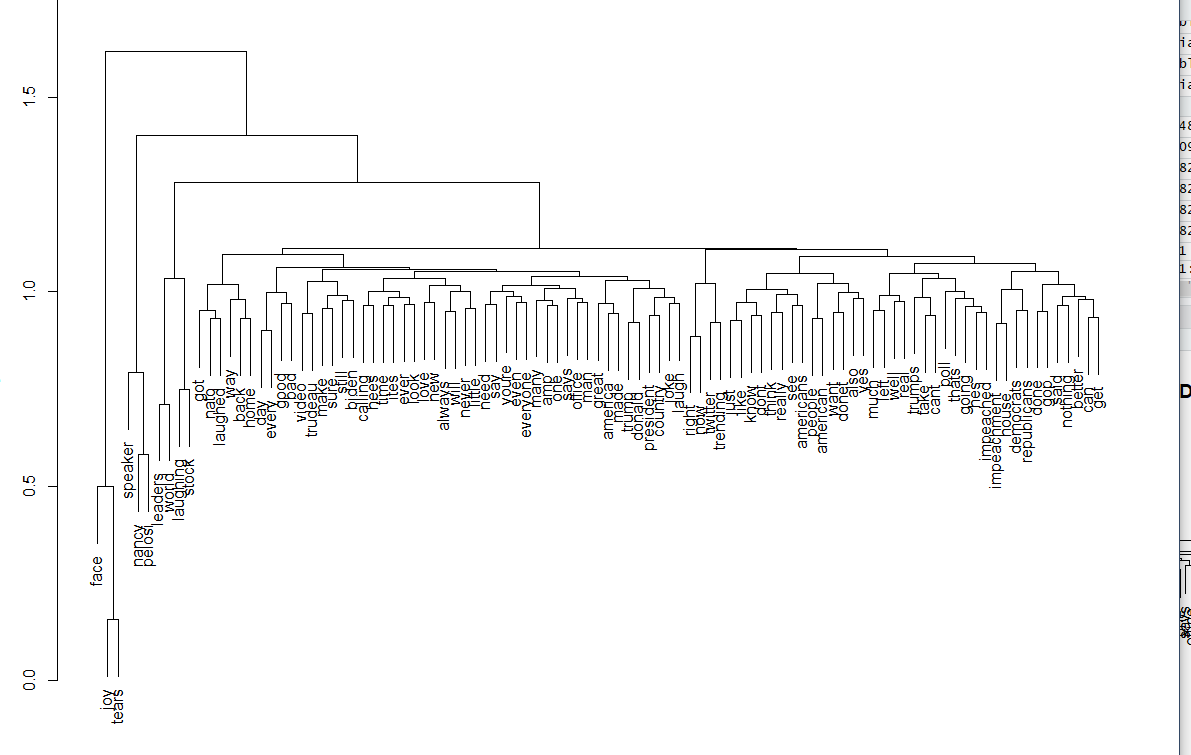
This world cloud is set to a min.freq of 50 and you can see the most common words are trump, laughing, tears, world, joy and face. This word cloud is set to a min.freq of 100 and again the most common words are face, trump, joy, tears, laughing, world.

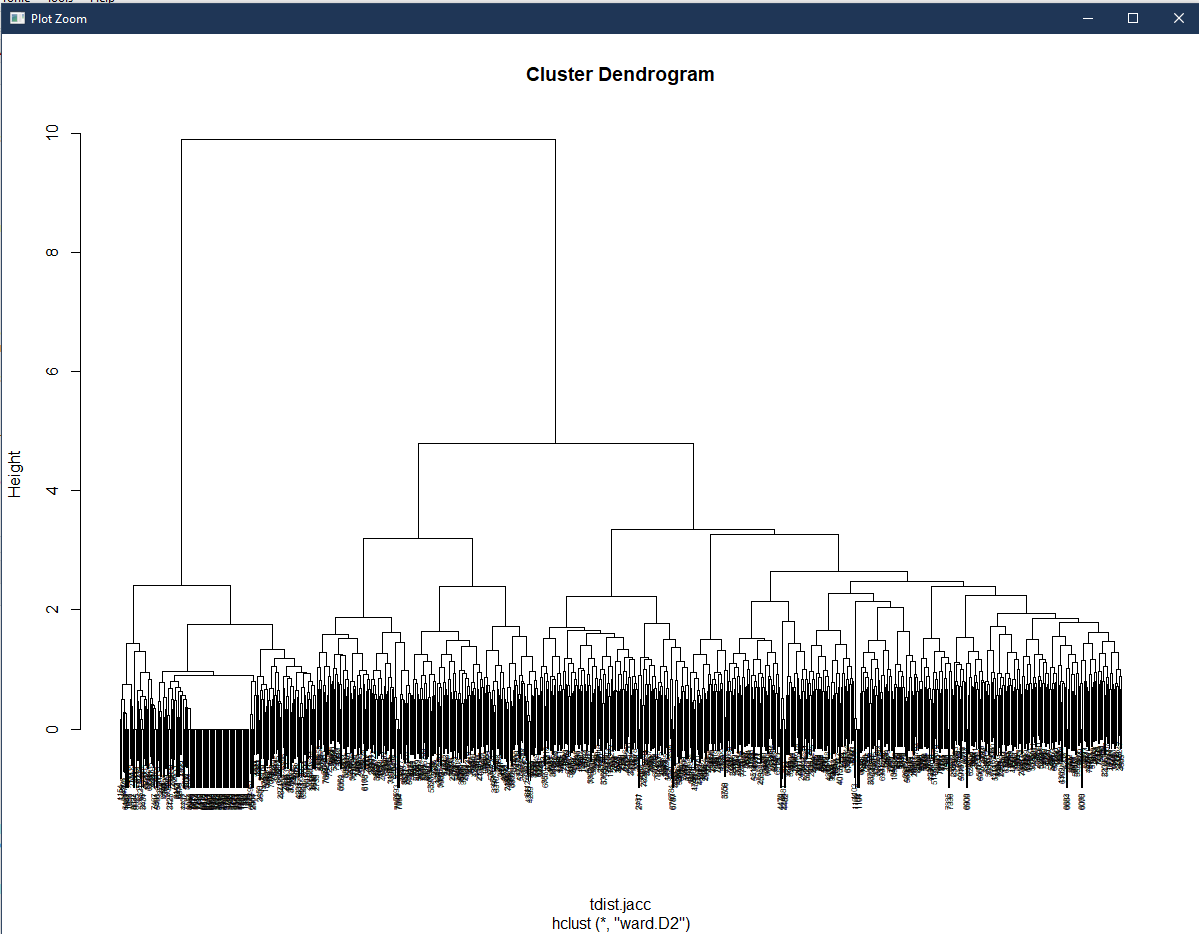
This word cloud is set to a min.freq of 300 narrows downs the words again and other words that haven’t been that relative are words like, people, like, will, just, one, amp and president.

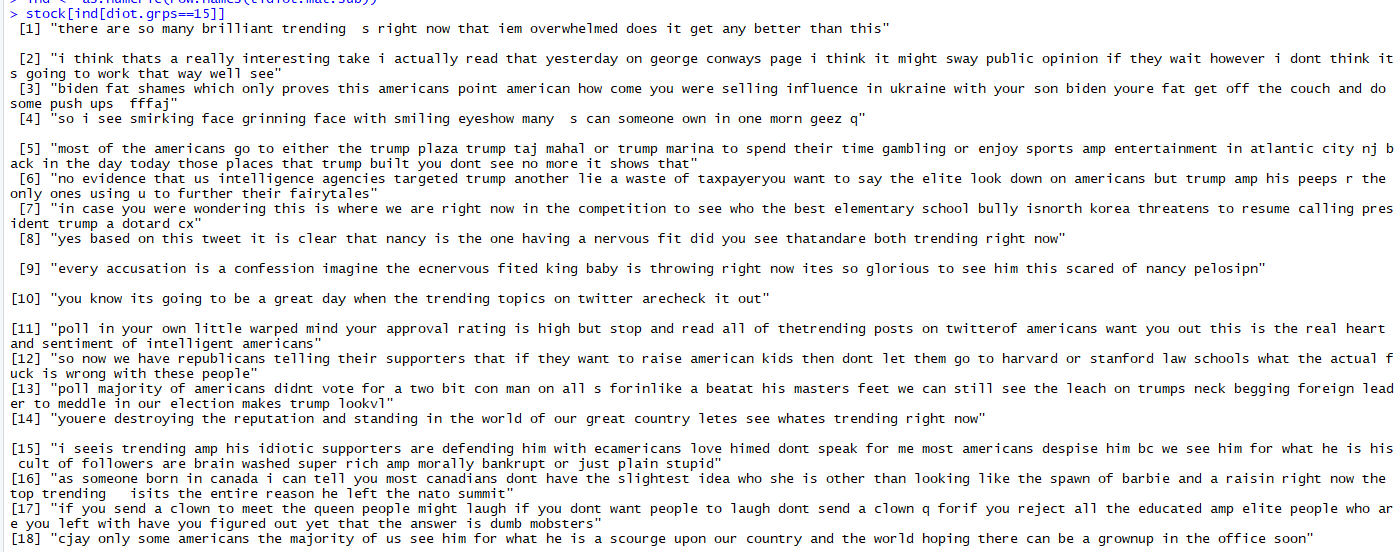


I then performed a cluster analysis on this data set

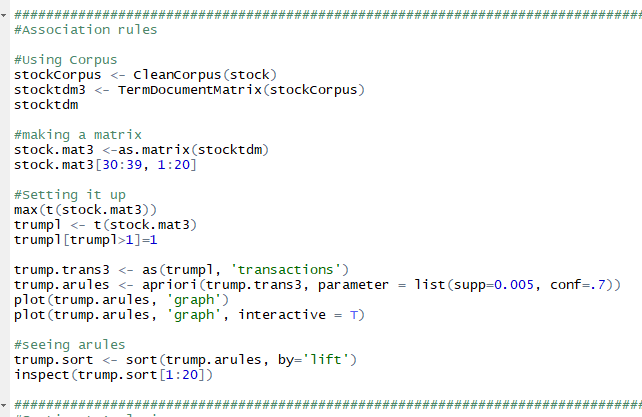


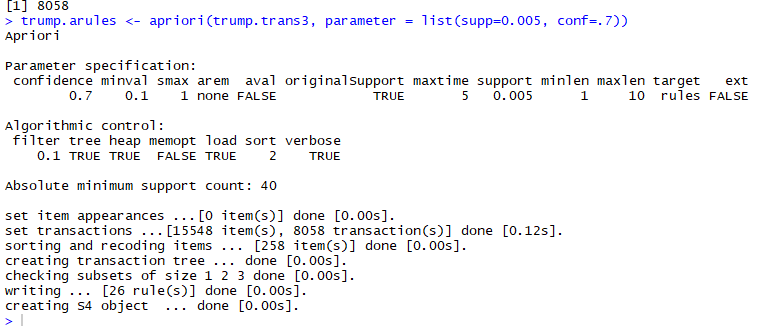


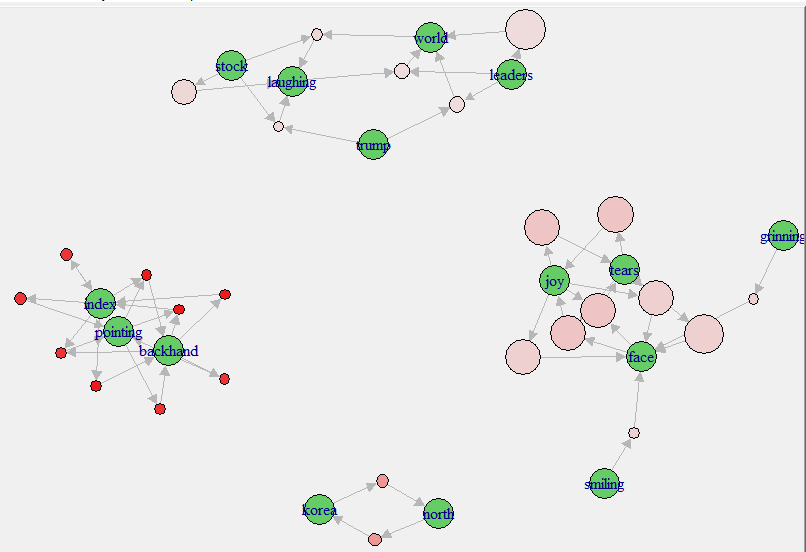




Then I did association rules and found around 24ish rules







And this is the plot for the most popular association rules