Twitter Data Mining and Predictive Modeling in R

Eliana Espinosa[1], Reinaldo Sanchez-Arias, Ph.D[2]

[1] St. Thomas University, School of Science, Miami Gardens, FL

Introduction

R, an open source statistical programming language [1], can be used to gather information from the social media platform Twitter, from which tweets are gathered from various news sources, celebrities, political figures, and some official colleges accounts. Other information such as screen names, number of tweets, number of followers, list of friends, and locations can be collected using the `twitteR` package [2] in combination with the Twitter application programming interface (Twitter API). After collecting this data, one can perform *text mining* by counting the word frequency in news sources' tweets, creating data visualizations to represent frequency of words, and conduct a *sentiment analysis* to understand and measure the impact of certain topics and opinions expressed in this social media venue. Spatial visualizations are also created in the form of interactive maps using the location data collected from different Twitter accounts. This project explores the various ways that Twitter can be used to gather information on certain topics and how this data could be used to help predict some of the behaviors and qualities on how people communicate through this social media source, as well as how different topics are perceived by society.

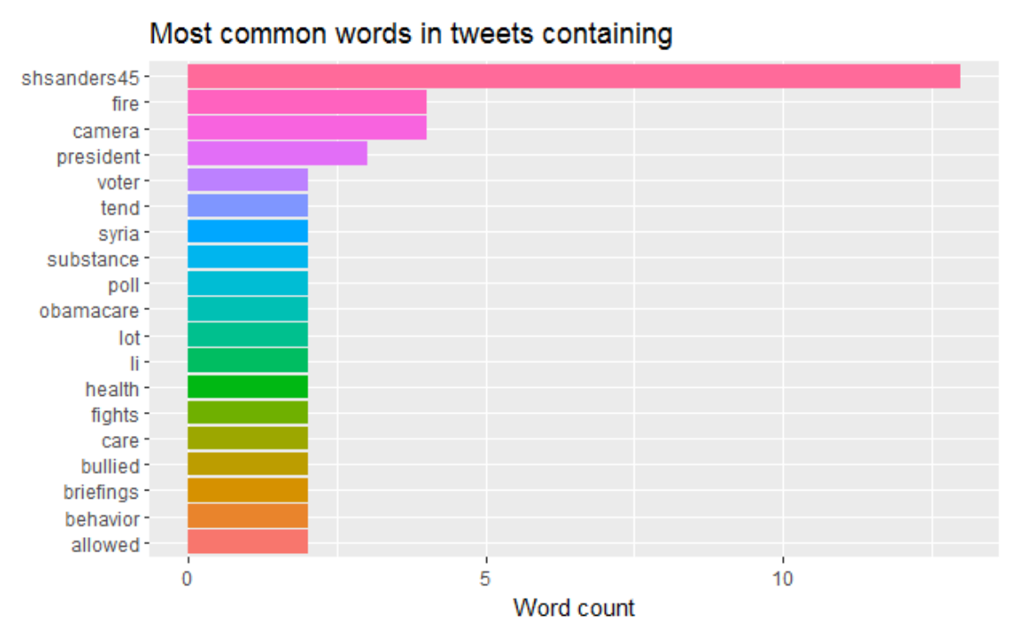
Methods

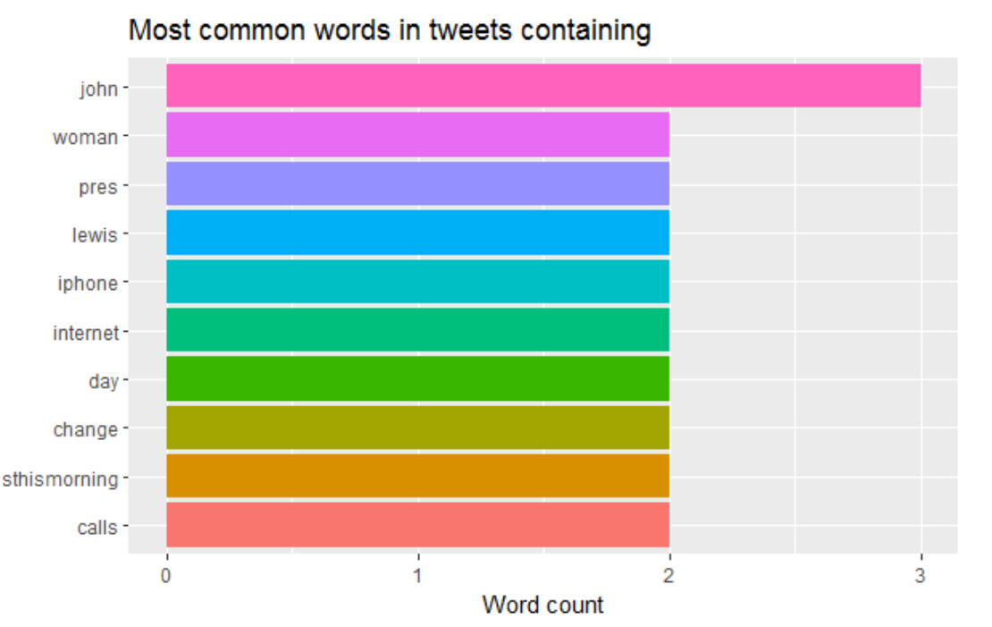
We used methods from the paper *"Mining Big Data to Extract Patterns and Predict Real-Life Outcomes"* by Michael Kosinski [3]. Some of the mathematical techniques used in this work include Singular Value Decomposition (SVD), and logistic regression models. A training matrix was generated gathering different data from celebrities including the number of people they follow, their age, gender, political party, and number of tweets. Using that data we created the matrix below.

```
        Name                   Gender Age  Pol. View Tweet Num. Sentiment
 [1,] "Adam Sandler"         "M"    "50" "R"       "180"      "0.2"
 [2,] "Ben Affleck"          "M"    "44" "D"       "395"      "0.53"
 [3,] "Beyonce"              "F"    "35" "D"       "10"       "0.9"
 [4,] "Blake Lively"         "F"    "29" "D"       "25"       "0.23"
 [5,] "Chris Hemsworth"      "M"    "33" "R"       "170"      "0.65"
 [6,] "Christina Aguilera"   "F"    "36" "D"       "979"      "0.9"
 [7,] "Cristiano Ronaldo"    "M"    "32" "D"       "2916"     "1.05"
 [8,] "Hugh Jackman"         "M"    "48" "D"       "2942"     "-0.4"
 [9,] "Jason Mraz"           "M"    "40" "D"       "3460"     "0.62"
[10,] "Karim Benzema"        "M"    "29" "D"       "1156"     "0.7"
[11,] "Kourtney Kardashian"  "F"    "38" "D"       "12200"    "0.25"
[12,] "Liam Hemsworth"       "M"    "27" "D"       "149"      "1.05"
[13,] "Luis Suarez"          "M"    "30" "R"       "838"      "1.39"
[14,] "Madonna"              "F"    "58" "D"       "2452"     "0.4"
[15,] "Mariah Carey"         "F"    "47" "D"       "7365"     "0.44"
[16,] "Mark Wahlberg"        "M"    "46" "D"       "1030"     "0.77"
[17,] "Michelle Obama"       "F"    "53" "D"       "798"      "1.27"
[18,] "Sandra Oh"            "F"    "45" "D"       "535"      "0.3"
[19,] "Sofia Vergara"        "F"    "44" "D"       "6819"     "0.11"
[20,] "Sylvester Stallone"   "M"    "70" "R"       "1251"     "0.62"
[21,] "Tom Hanks"            "M"    "60" "D"       "799"      "0.3"
[22,] "Vicoria Beckham"      "F"    "43" "R"       "3156"     "0.65"
[23,] "Zac Efron"            "M"    "29" "D"       "1585"     "0.86"
```

Experimental Design

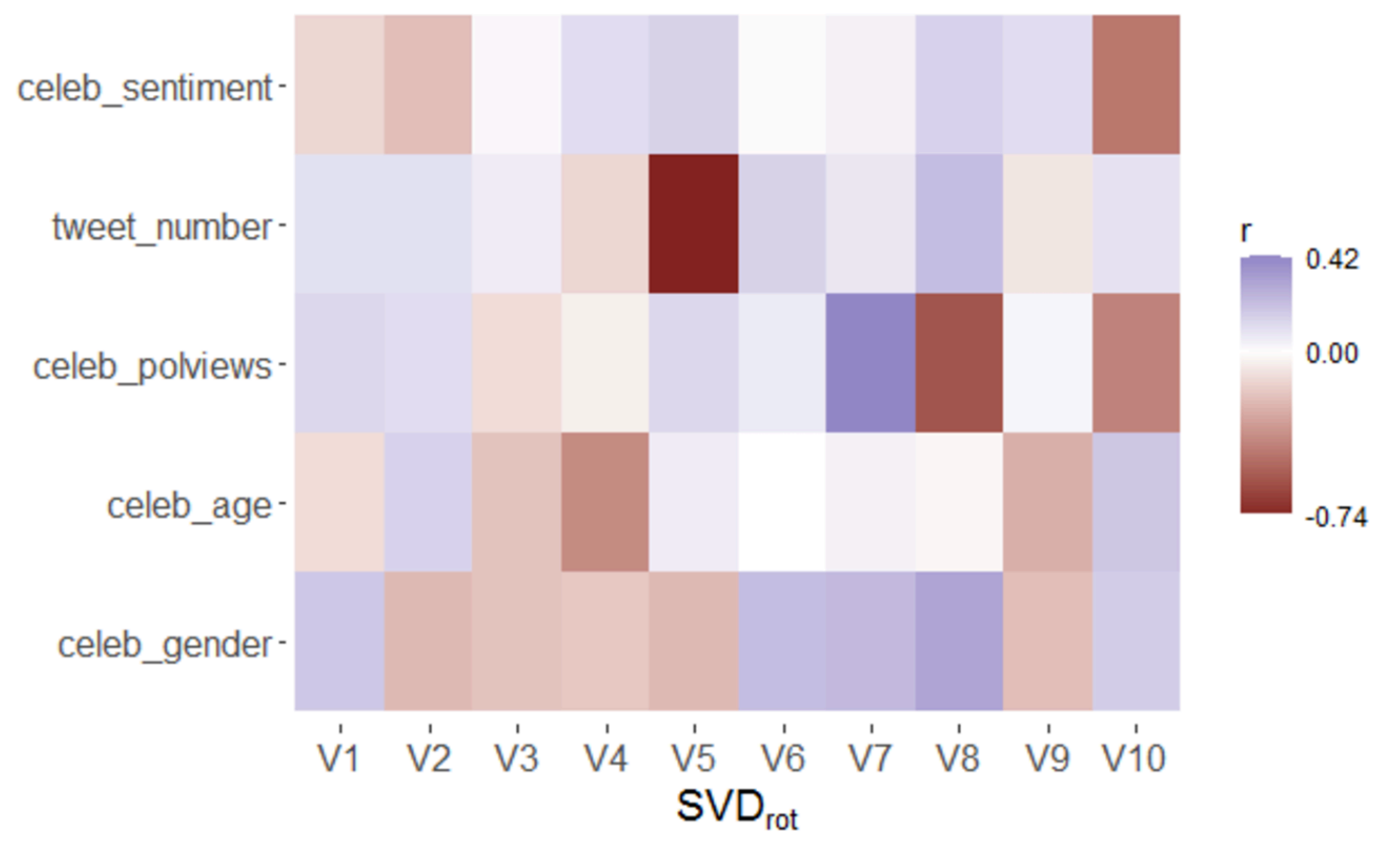
The `twitteR` package allowed us to collect different types of information from Twitter. Using the twitters of news sources, like ABC, CBS, CNN, FOX, and NBC, we were able to extract their tweets and perform text mining analysis. First, we collected 20 tweets from each of their Twitter accounts, then we used the `twListToDF()` function to convert the list of tweets into a data frame. From there we created a column of words that were used in each tweet and removed any unnecessary words (such as "the", "a", and "and") as well as any links, emojis, and symbols. After this column was created we used the `count()` function to count the number of times each word was repeated within the 20 tweets that were collected. Bar plots were created to show the words that were used the most in their tweets.



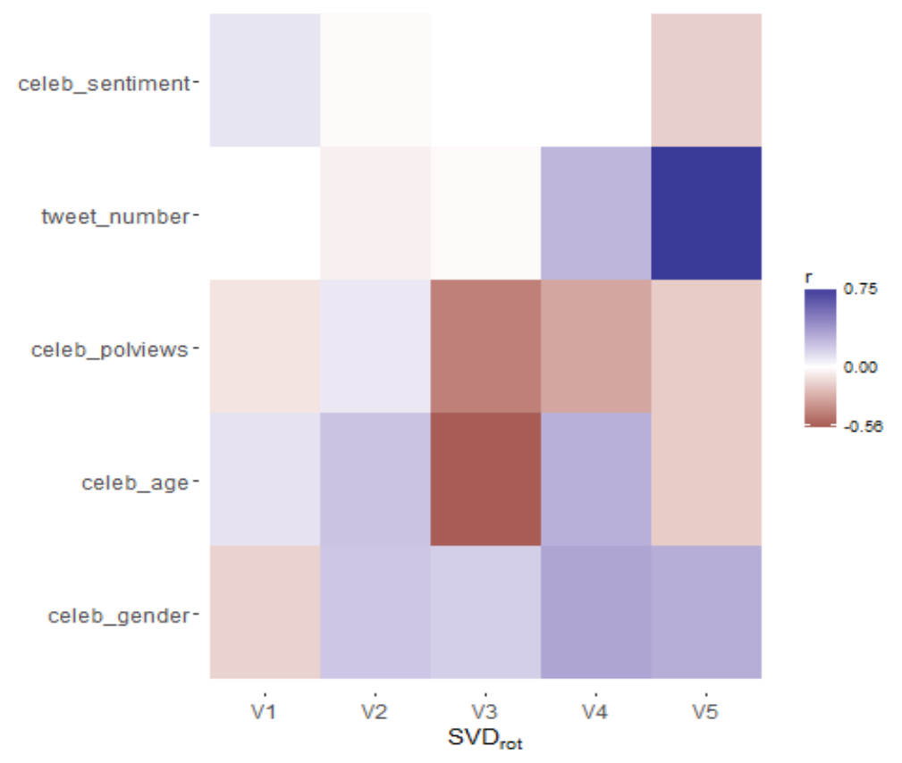


Exploratory Data Analysis and Mathematical Methods

Following the methodology Kosinski's paper [3], we were able to create a *heatmap* showing the correlation between the SVD dimensions and the psych demographic traits of each celebrity in the dataset we created. While the paper uses k = 5 SVD dimensions, we are also using k= 10 SVD dimensions.



Notice that the SVD dimension V5 correlates positively with the number of tweets and gender while it has a negative correlation with the sentiment, political views, and age.



Exploring Twitter data with the `twitteR` package

Using the `twitteR` package in R, we were able to gather the locations of people who follow the Twitter accounts of different universities. Once their locations were gathered, they were converted into longitude and latitude coordinates which then made it possible to plot their locations on a map. This helped us get a sense of the amount of diversity in each of the universities.

Here we present two types of maps showing the location of people followed by St Thomas University Twitter's account, both at the national and global levels.

Exploring Twitter data with the `wordcloud` package

We also present word clouds created using the `wordcloud()` function in R, as a way to better visualize the variety of words most commonly used by a given Twitter user.

Singular Value Decomposition

The factorization of a real or complex matrix.

$$A = U\Sigma V^T$$

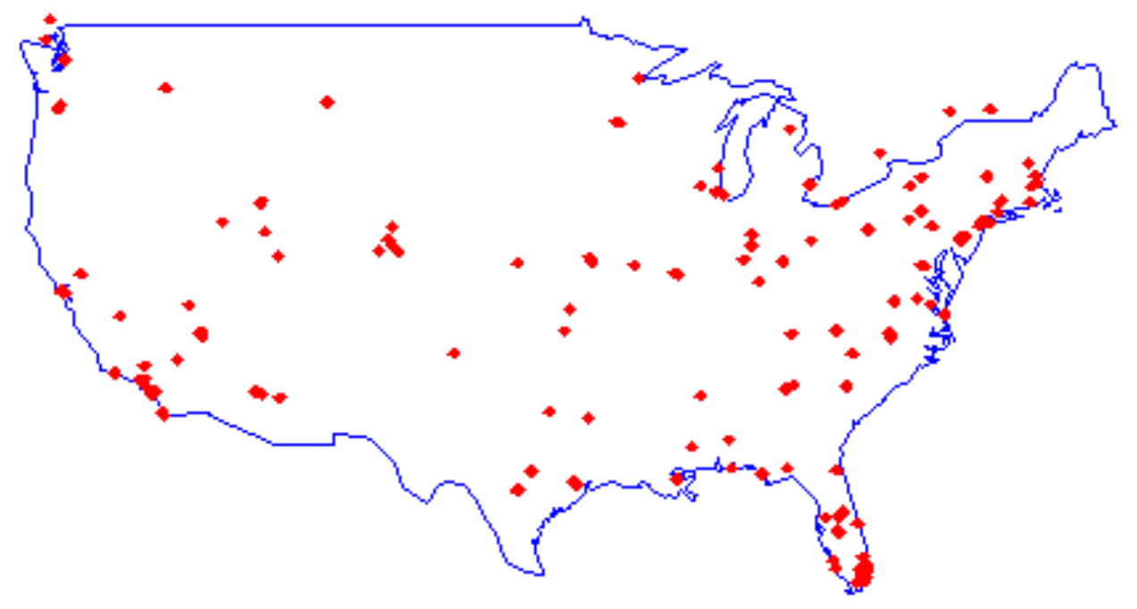
Where U is an m × m, Σ is an m × n, and V^T is an n × n

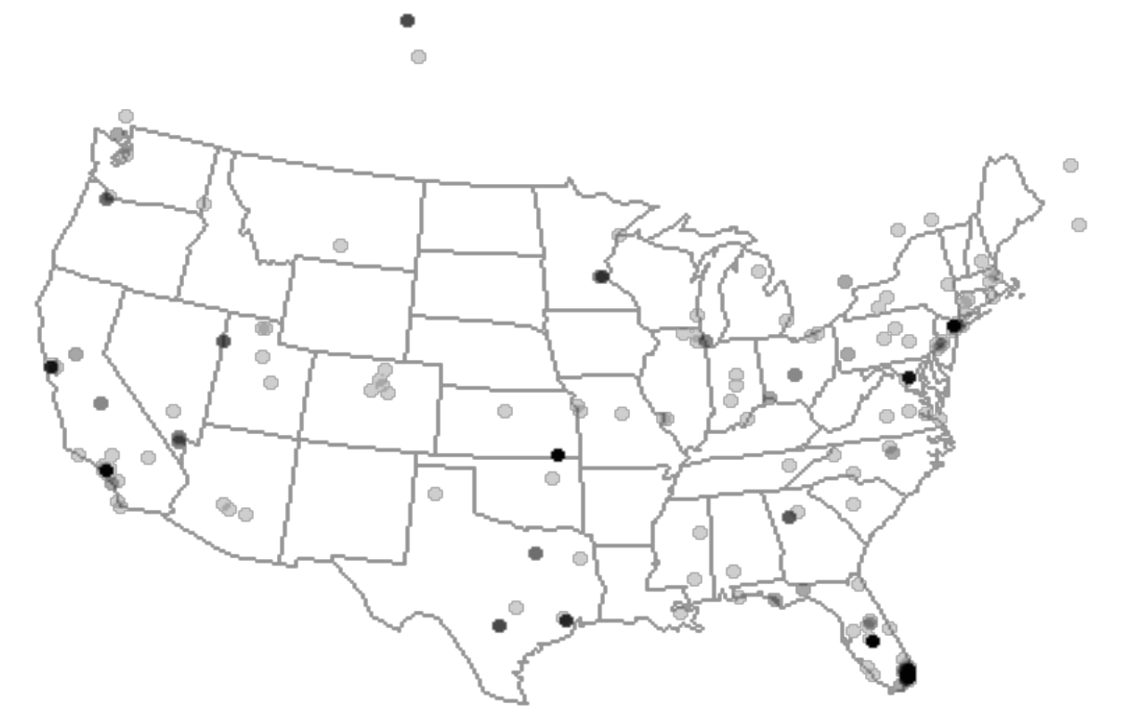
The singular value decomposition of A is:

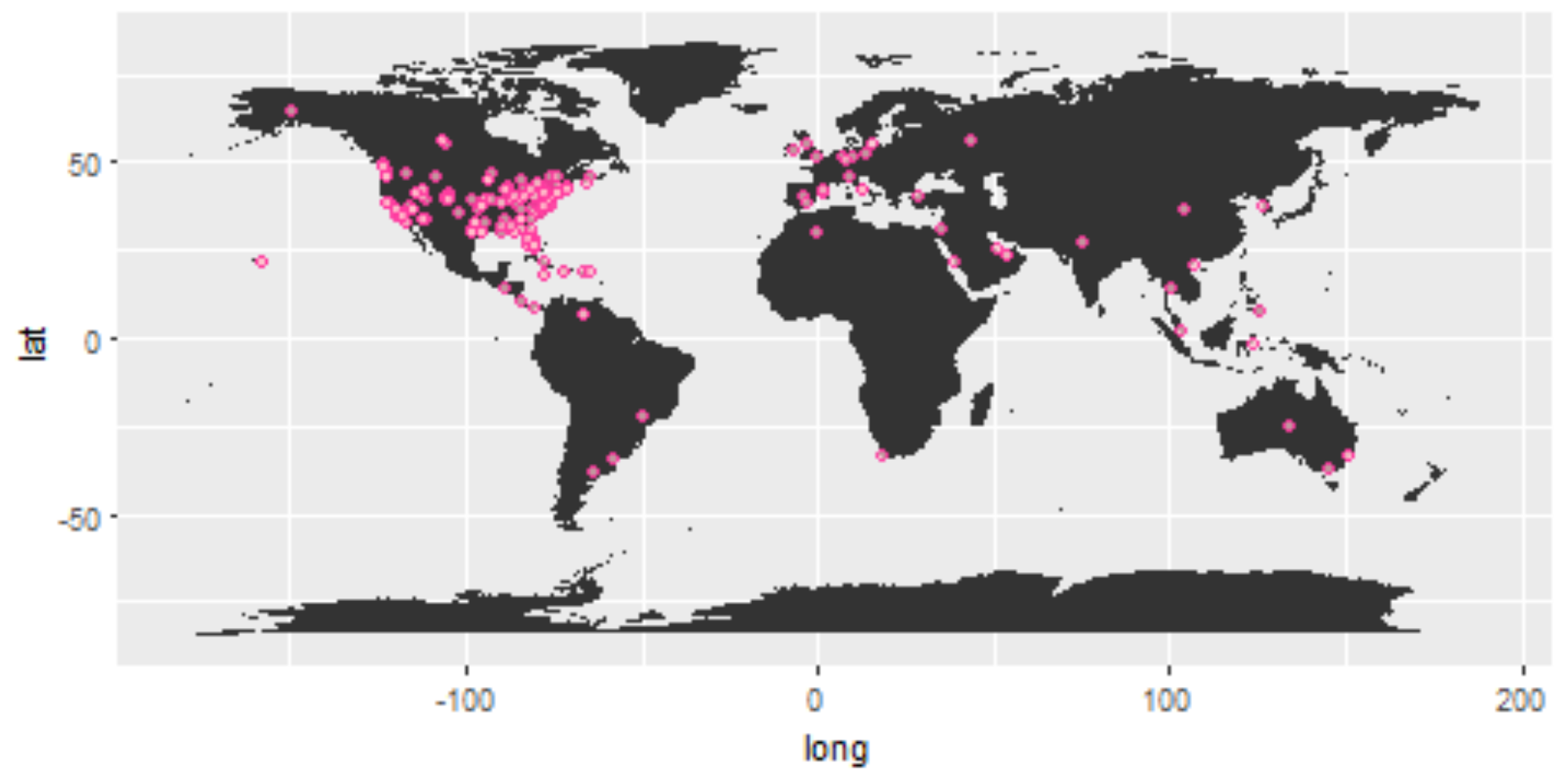
$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{6}}{3} & 0 & \frac{-1}{\sqrt{3}} \\ \frac{\sqrt{6}}{6} & \frac{-\sqrt{2}}{2} & \frac{1}{\sqrt{3}} \\ \frac{\sqrt{6}}{6} & \frac{\sqrt{2}}{2} & \frac{1}{\sqrt{3}} \end{bmatrix} \begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{-\sqrt{2}}{2} \end{bmatrix}$$

$$\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \ldots \geq \sigma_n$$

```
# load libraries
library(ggplot2)
library(reshape2)

# get correlations
x<-round(cor(u_rot, df_celeb[,-1], use="p"),2)

# reshape it in a ggplot2 friendly way
y<-melt(x)
colnames(y)<-c("SVD", "Trait", "r")

# produce the plot
qplot(x=SVD, y=Trait, data=y, fill=r, geom="tile") +
  scale_fill_gradient2(limits=range(x), breaks=c(min(x), 0, max(x)))+
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=14,face="bold"),
        panel.background = element_rect(fill='white', colour='white'))+
  labs(x=expression('SVD'[rot]), y=NULL)
```





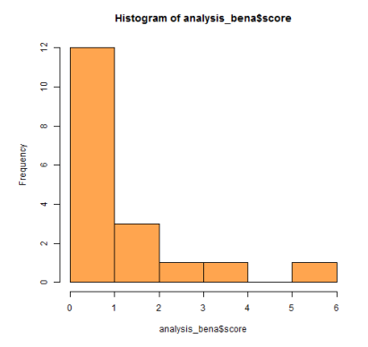


Sentiment Analysis

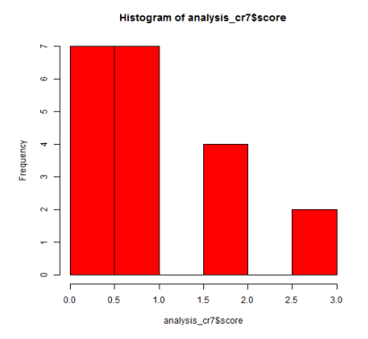
Sentiment analysis can be thought of as the exercise of taking a sentence, paragraph, document, or any piece of natural language, and determining whether that text's emotional tone is positive, negative or neutral.

We collected tweets from 23 celebrities that follow less than 100 people. We were able to calculate the sentiment of their tweets by using a list of negative and positive words. After we calculated the sentiment, we created a histogram of the sentiments for each celebrity.

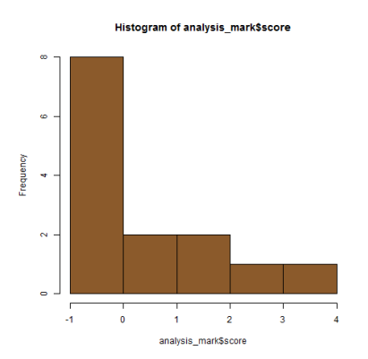
```
score.sentiment = function(tweets, pos.words, neg.words){
require(plyr)
require(stringr)
scores = laply(tweets, function(tweet, pos.words, neg.words) {
tweet = gsub('https://','',tweet)
tweet = gsub('http://','',tweet)
tweet=gsub('[^[:graph:]]', ' ',tweet)
tweet = gsub('[[:punct:]]', '', tweet)
tweet = gsub('[[:cntrl:]]', '', tweet)
tweet = gsub('\\d+', '', tweet) # removes numbers
tweet=str_replace_all(tweet,"[^[:graph:]]", " ")
tweet = tolower(tweet)
word.list = str_split(tweet, '\\s+')
words = unlist(word.list)
pos.matches = match(words, pos.words)
neg.matches = match(words, neg.words)
pos.matches = !is.na(pos.matches)
neg.matches = !is.na(neg.matches)
score = sum(pos.matches) - sum(neg.matches)
return(score)
}, pos.words, neg.words )
scores.df = data.frame(score=scores, text=tweets)
return(scores.df)
}
```



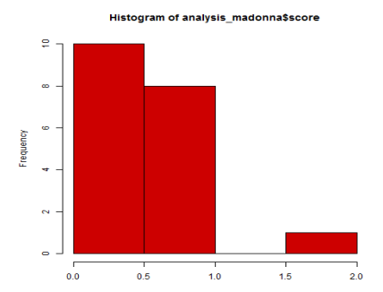
Sentiment of tweets by Ben Affleck



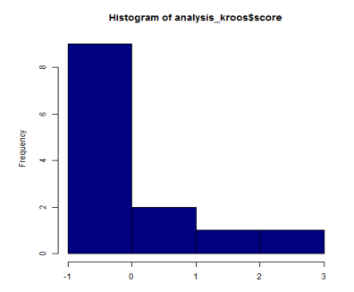
Sentiment of tweets by Cristiano Ronaldo



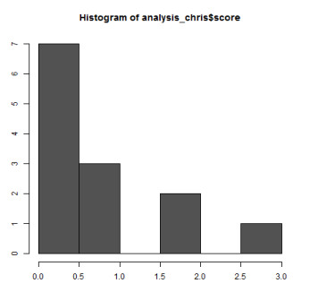
Sentiment of tweets by Mark Wahlberg



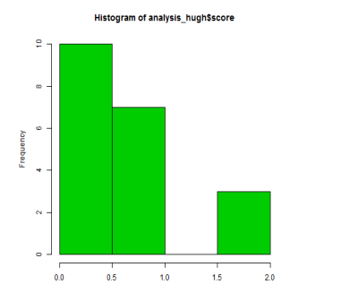
Sentiment of tweets by Madonna



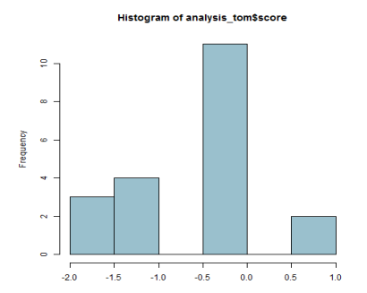
Sentiment of tweets by Toni Kroos



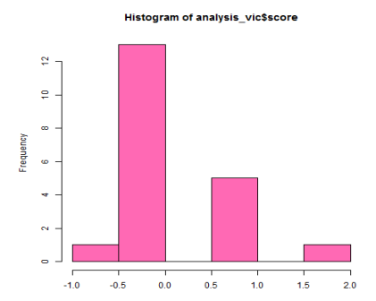
Sentiment of tweets by Chris Hemsworth



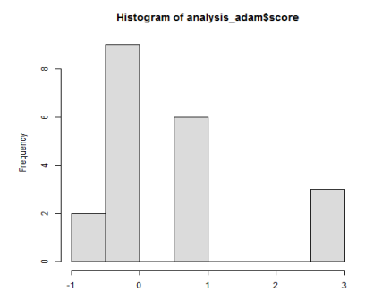
Sentiment of tweets by Hugh Jackman



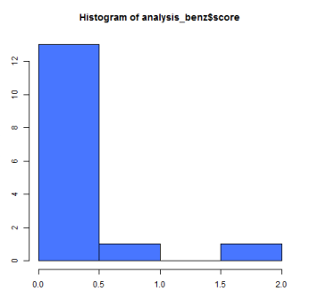
Sentiment of tweets by Tom Hanks



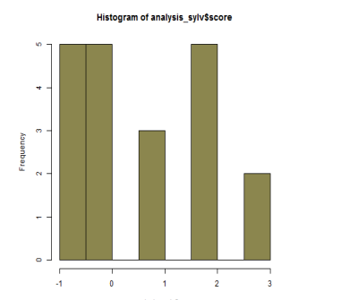
Sentiment of tweets by Victoria Beckham



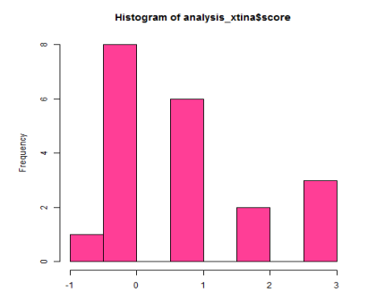
Sentiment of tweets by Adam Sandler



Sentiment of tweets by Benzema



Sentiment of tweets by Sylvester Stallone



Sentiment of tweets by Christina Aguilera

RStudio

Programming was essential in the development of this project. R is an open source language widely used in the data science community, with focus on statistical data analysis, data visualization and machine learning methods. During this project, tools from the `tidyverse` package [4] were used for data wrangling and data visualization with the help of RStudio [5] , an open source integrated development environment (IDE) for R.

References

[1] R: free software environment for statistical computing and graphics. https://www.r-project.org/

[2] twitteR https://cran.r-project.org/web/packages/twitteR/README.html

[3] *"Mining Big Data to Extract Patterns and Predict Real-Life Outcomes"* by M. Kosinski, Y. Wang, H. Lakkaraju, and J. Leskovec, Psychological Methods, 2016.

[4] tidyverse: an opinionated collection of R packages designed for data science. https://www.tidyverse.org/

[5] Rstudio: open-source integrated development environment (IDE) for R https://www.rstudio.com/

Acknowledgements

This work was made possible by the support of the School of Science at St. Thomas University and the collaboration of Professor Sanchez-Arias.

This project was supported, in part, by U.S. Department of Education grant award P03C1160161 (STEM SPACE), P031C160143 (STEM EngInE), P120A160036 (STEM ISLE), 1161177 (STEP Up), P120A140012 (SPARC). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the respective funding agency.