Introduction

Twitter is a powerful social platform that, in the modern day, can affect just about anyone or anything. Measuring twitter sentiment for stocks has been a hot ticket over the past five to ten years. Typically when a stock is green for the day (positive), then you will see a positive sentiment for that stock. Vice-versa, if the stock is in the red for the day (negative) then you will most likely see a negative sentiment on Twitter for that stock. For the day of April 20, 2018 the top gaining stocks for the day were TransUnion ($TRU), Telefonaktiebolaget LM Ericsson ($ERIC) and Pinnacle Foods, Inc. ($PF). For the same day, the three losing stocks for the day were Sketchers U.S.A., INC ($SKX), Sage Therapeutics ($SAGE) and Gentex Corporation ($GNTX). Using the Twitter API, extracting 100 tweets for each stock on that day and then storing them into a corpus. Once in a corpus, the preprocessing will take place to clean the text to make sure that it is easier to measure the sentiment of the top gaining and losing stocks on Twitter for April 20, 2018.

A)

The gaining stocks (TRU, ERIC, PF) happened to have more popularity on Twitter when searching for the 100 tweets. There was nothing that had to be done to extract the tweets for the gaining stocks of the day. The losing stocks (SAGE, SKX, GNTX) of the day required searching a hashtag of the company's name along with the cash symbol and ticker to get the full 100 tweets for each. Once the tweets were scraped off of Twitter, the tweets were then combined into their respective categories, gainers and losers (for the day) with three hundred tweets in each list.

B)

The corpora was formed by taking the two lists of tweets and passing them to a created function that will extract the text using *lapply* to return just the text from each tweet. Within the *lapply* function the *iconv* converted all the text to ASCII format. This allows all the text to be character-encoded, which is in the most common format for text in computers. Once the text is formatted, it is then passed to *VectorSource()*, which takes each tweet as a document and then the *Corpus()* function to store the documents in a corpora.

Corpus for Gaining Stocks:

```
inspect(data.corpus1)

<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 300
```

Corpus for Losing Stocks:

```
inspect(data.corpus2)

<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 300
```

C)

The first pre-processing step was completed with passing the tweets into a function that created a list and corpora in ASCII format for each of the two lists of tweets. The next pre-processing steps will be to make sure that the text is tidy and clean to be analyzed for sentiment. The pre-processing function created includes the following pre-processing techniques in order:

1. Convert all the text to lower case.
2. Remove the URL so that there is no confusion with text left behind from removing punctuation.
3. Remove punctuation
4. Remove all stop-words in the English language
5. Words that begin with numbers are removed
6. Stem the document to reduce all words to the root word
7. Remove any extra white spaces that may remain once the text is processed, the corpora will be returned and ready to be analyzed for word frequencies and sentiment.

D)

Now that the corpora have been processed, passing each corpora to the *TermDocumentMatrix()* function which takes all the documents in the corpora and returns a matrix that describes the frequency of terms that occur in the documents.

Term Document Matrix for Gaining Stocks:

```
inspect(tdm1)

<<TermDocumentMatrix (terms: 733, documents: 300)>>
Non-/sparse entries: 3234/216666
Sparsity           : 99%
Maximal term length: 19
Weighting          : term frequency (tf)
Sample             :
        Docs
Terms    149 204 232 238 27 29 46 5 65 96
  clf      1   0   0   0  1  1  0 0  1  1
  drna     1   0   0   0  1  1  0 0  0  0
  eric     1   0   0   0  1  1  0 0  1  1
  gainer   0   0   0   0  1  0  0 0  1  0
  llnw     1   1   1   1  1  1  1 1  1  0
  nap      1   0   0   0  1  1  0 0  1  0
  ssw      1   0   0   0  1  1  0 0  1  0
  top      0   1   1   0  1  0  1 1  1  0
  tru      1   1   1   0  1  1  1 1  1  1
  vlrx     1   0   0   0  1  1  0 0  1  0
```

Term Document Matrix for Losing Stocks:

```
inspect(tdm2)

<<TermDocumentMatrix (terms: 1014, documents: 300)>>
Non-/sparse entries: 3213/300987
Sparsity           : 99%
Maximal term length: 19
Weighting          : term frequency (tf)
Sample             :
       Docs
Terms   107 259 266 281 282 283 284 285 287 288
  amp     0   0   0   0   0   0   0   0   0   0
  bhge    0   1   1   1   1   1   1   1   1   1
  clf     0   1   1   1   1   1   1   1   1   1
  earn    0   1   1   1   1   1   1   1   1   1
  gntx    0   1   1   1   1   1   1   1   1   1
  man     0   1   1   1   1   1   1   1   1   1
  sage    0   0   0   0   0   0   0   0   0   0
  skx     1   0   0   0   0   0   0   0   0   0
  stock   1   0   0   0   0   0   0   0   0   0
  swk     0   1   1   1   1   1   1   1   1   1
```

The terms above can be seen and appear to be other stocks that were negative or positive on the day. This is due to that when someone tweets out about a stock, typically they include other stock tickers in the tweet as well.

E)

The Term Document Matrix allows for an easy creation of a matrix of terms and their frequencies in each corpora. The terms will be looked at as their row sums in appearing each corpora. In decreasing order (most frequent to least frequent), the top five most frequent terms in each corpora are:

Most frequent terms for Gaining Stocks:

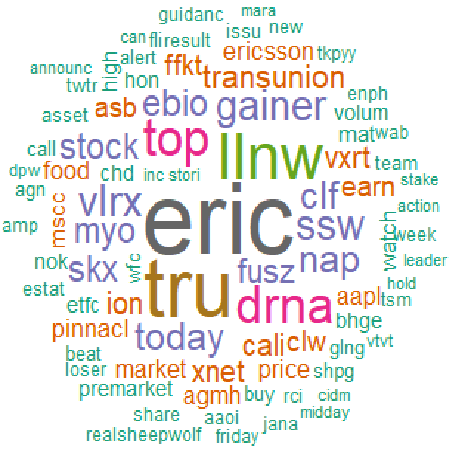
```
head(pos_sortedfreq)

eric  tru llnw  top drna vlrx
 130  104   80   61   60   48
```

Most frequent terms for Losing Stocks:

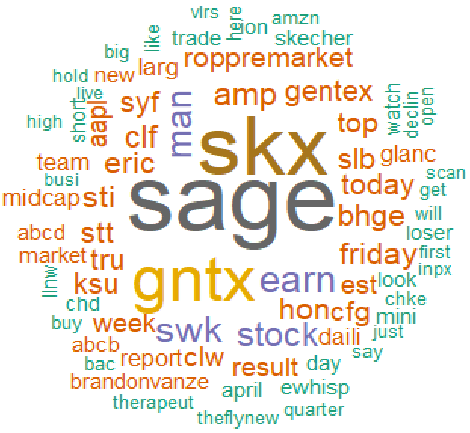
```
head(neg_sortedfreq)

sage  skx gntx earn  man  swk
 106   92   67   34   33   33
```

Visualizing the most frequent terms is often done using a wordcloud. The word-clouds for each, the gaining stocks and losing stocks are shown below:

The wordcloud for the Gaining Stocks



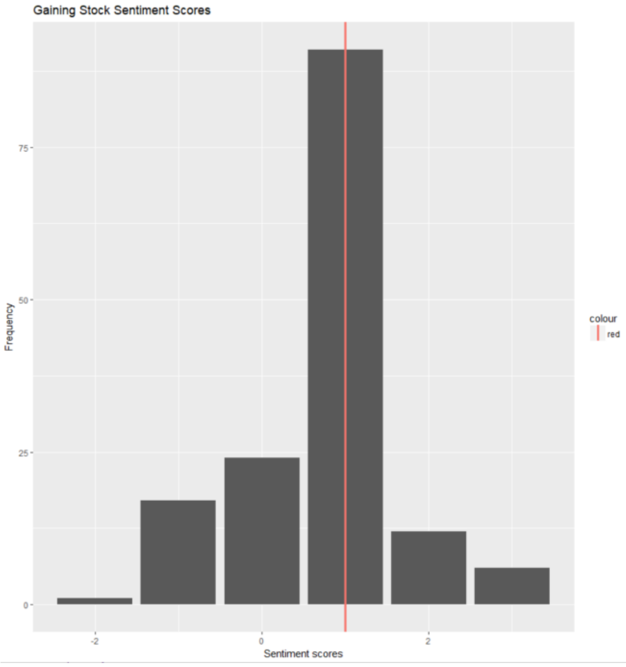
The wordcloud for the Losing Stocks

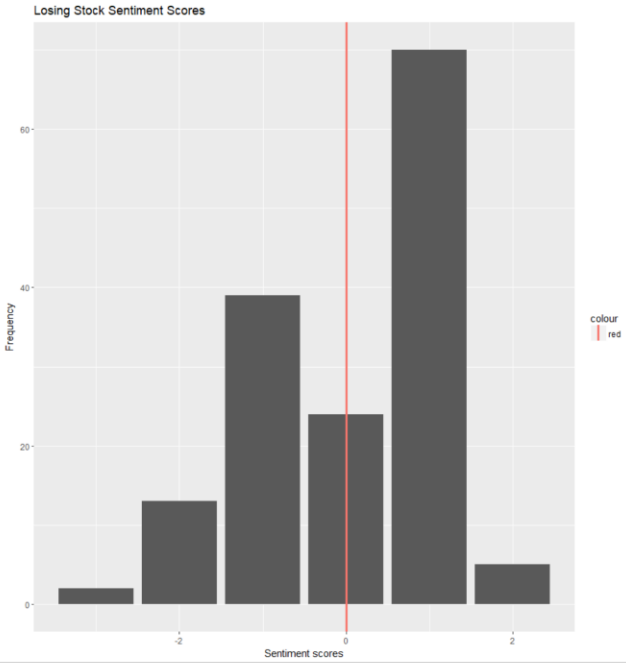


F)

The Sentiment of the tweets for both the Gaining and Losing stocks will be calculated by taking the text from each group and passing them through a created function. The created function will create a vector of words and then find which words are positive and which are negative. By choice, the function will exclude the tweets that are neither negative nor positive form the final tweets that will allow for a better visual of the sentiment for each group.

The bar plots for both the Gaining stocks and Losing stocks are side by side for comparison.



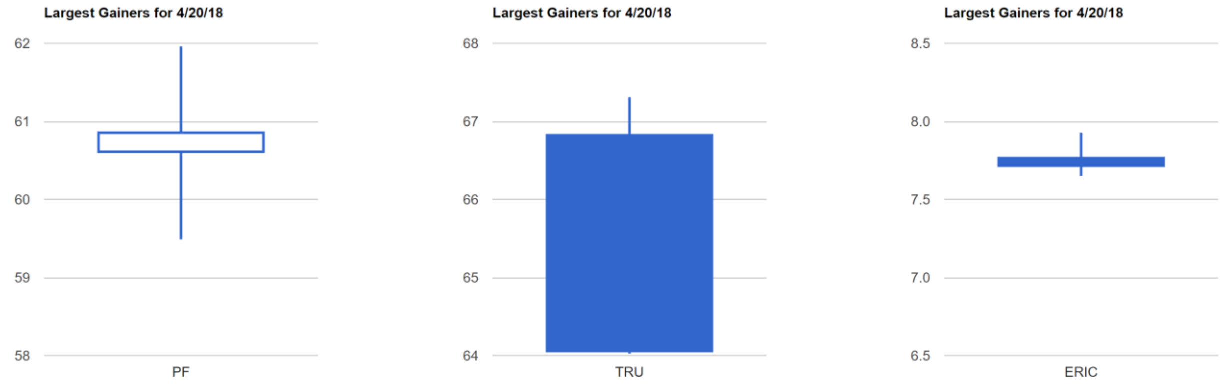


The median is chosen when plotting the bar plots so, that the middle of the data is shown. When looking at the median line in each of the bar-plots the Gaining stocks clearly have a positive sentiment in the tweets extracted. The Losing stocks have a neutral sentiment based on the average of the sentiment scores from the tweets. This is due to, Gentex Corporation, one of the losing stocks having a positive "outlook" on its earnings before the market open and then missed earnings when they were reported as the market opened. There are also mixed tweets regarding Sketchers U.S.A., Inc and the stock price vs options, the options still have a positive outlook. This caused a mixture in sentiment among the tweets that were gathered.

G)

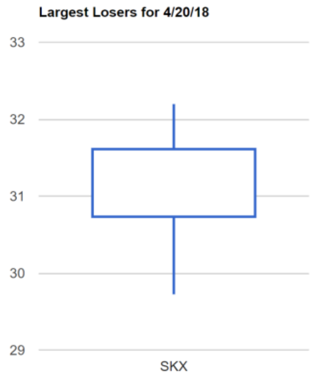
GoogleVis charts are created using the *quantmod* library to gather the data for the stocks. The candlestick charts are broken up by gaining and losing stocks. The three candlestick charts for each stock in each category are "merged" together using *gvisMerge.* The charts can be seen below:

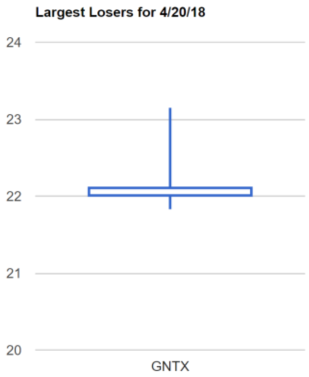
Three Gaining stocks of the day:

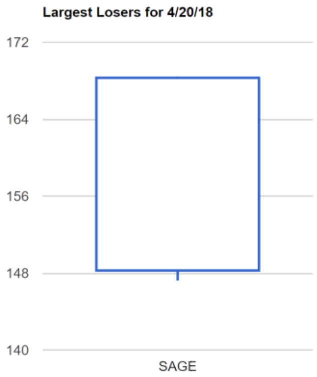


As you can see, Pinnacle Foods (PF) actually declined for the day on April 20th. Its pre-market hours on that day are where the price jumped from $55.20 to $60.88 on the open. Although the price of the stock declined during market hours, the pre-market hours are still factored into the "gain" for the overall stock price.

Three Losing stocks of the day:







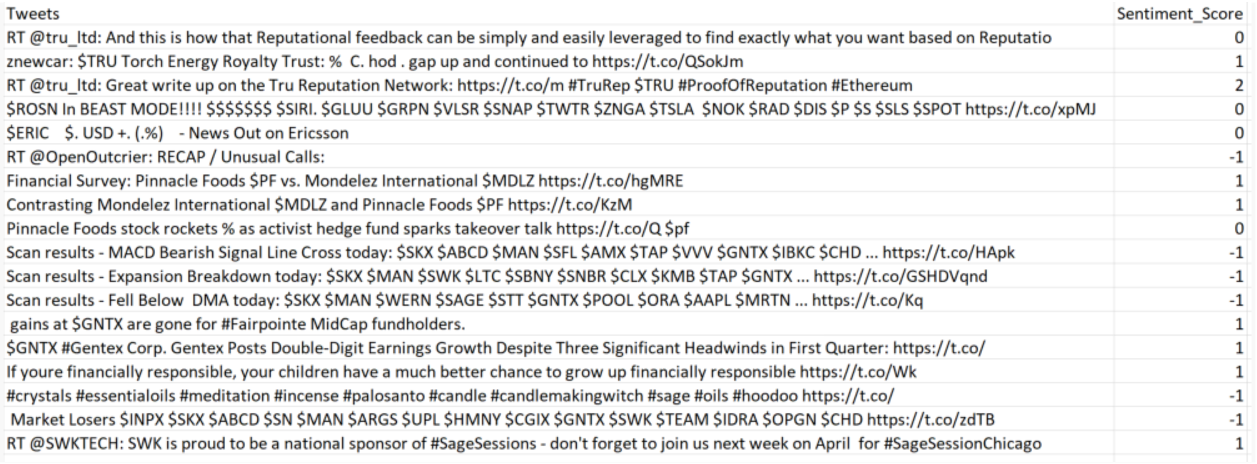
The three losing stocks of the day are all seen declining on April 20th, 2018.

Conclusion

In conclusion, as you can see, the three Gaining stocks have an overall positive sentiment because they were in the green for the day and investors were happy to be making money off of those stocks. The positive sentiment for the Gaining stocks was to be expected. The Losing stock sentiment, one would think would be a negative sentiment. Due to somewhat unique circumstances on this day the three Losing Stocks also had an overall positive sentiment. Looking into why that may be, Gentex Corporation (GNTX) had positive outlook going into earnings on April 20th, then once reported that the company missed on its earnings, the stock declined into negative territory. Although there was a miss on earnings for Gentex Corporation, the “Buy” opinion on the company’s stock remained. Sketchers (SKX) has remained an “Overweight” at top firms such as Morgan Stanley with a price target of $225 (currently the stock price is $148), which means that the sentiment could show that with the drop in price of SKX, that it could be a good buying opportunity. These are two factors that have played into the positive sentiment for the Losing stocks of the day for April 20, 2018.

Tweets Data Frame

Three tweets from each one of the three stocks in each category (Gaining & Losing) is shown below with its sentiment score. For the sake of showing the table below, the “NA” results in the sentiment score training were labeled as 0. The tweet text is shown before any preprocessing was done. This allows to see the link to the news article or tweet.



Tweets	Sentiment_Score
RT @tru_ltd: And this is how that Reputational feedback can be simply and easily leveraged to find exactly what you want based on Reputatio	0
znewcar: $TRU Torch Energy Royalty Trust: % C. hod . gap up and continued to https://t.co/QSokJm	1
RT @tru_ltd: Great write up on the Tru Reputation Network: https://t.co/m #TruRep $TRU #ProofOfReputation #Ethereum	2
$ROSN In BEAST MODE!!!! $$$$$$$ $SIRI. $GLUU $GRPN $VLSR $SNAP $TWTR $ZNGA $TSLA $NOK $RAD $DIS $P $S $SLS $SPOT https://t.co/xpMJ	0
$ERIC $. USD +. (.%) - News Out on Ericsson	0
RT @OpenOutcrier: RECAP / Unusual Calls:	-1
Financial Survey: Pinnacle Foods $PF vs. Mondelez International $MDLZ https://t.co/hgMRE	1
Contrasting Mondelez International $MDLZ and Pinnacle Foods $PF https://t.co/KzM	1
Pinnacle Foods stock rockets % as activist hedge fund sparks takeover talk https://t.co/Q $pf	0
Scan results - MACD Bearish Signal Line Cross today: $SKX $ABCD $MAN $SFL $AMX $TAP $VVV $GNTX $IBKC $CHD ... https://t.co/HApk	-1
Scan results - Expansion Breakdown today: $SKX $MAN $SWK $LTC $SBNY $SNBR $CLX $KMB $TAP $GNTX ... https://t.co/GSHDVqnd	-1
Scan results - Fell Below DMA today: $SKX $MAN $WERN $SAGE $STT $GNTX $POOL $ORA $AAPL $MRTN ... https://t.co/Kq	-1
gains at $GNTX are gone for #Fairpointe MidCap fundholders.	1
$GNTX #Gentex Corp. Gentex Posts Double-Digit Earnings Growth Despite Three Significant Headwinds in First Quarter: https://t.co/	1
If youre financially responsible, your children have a much better chance to grow up financially responsible https://t.co/Wk	1
#crystals #essentialoils #meditation #incense #palosanto #candle #candlemakingwitch #sage #oils #hoodoo https://t.co/	-1
Market Losers $INPX $SKX $ABCD $SN $MAN $ARGS $UPL $HMNY $CGIX $GNTX $SWK $TEAM $IDRA $OPGN $CHD https://t.co/zdTB	-1
RT @SWKTECH: SWK is proud to be a national sponsor of #SageSessions - don't forget to join us next week on April for #SageSessionChicago	1