TWITTER GEOLOCATION PROJECT

Victoria's Secret Fashion Show

LU LI

U67923571

https://github.com/elenaluuu/Twitter_project

Table of Contents

Introduction

This project is about getting twitter information about Victoria's Secret Fashion Show in 2016. The main purpose is comparing tweets between location in USA. "VSFashionShow" is used as the keyword when filtering tweets. Text mining is also found in the process of project. Word cloud are used to represent hot concerns of people who write tweets about this fashion show. Maps are generated concerning about different aspect of tweet between different locations. what's more, sentiment analysis is also presented based on the tweet texts.

Analysis Procedures

1. Set up twitter developers account
2. Connect to twitter API in R and save "*my_oauth.Rdata"*
3. Search twitter about the keyword "**VSFashionShow**" and save it into "tweets.*json"* file.
4. Transform the *json* file into data frame in R and save into RDS file for further reproduction.
5. There are 57695 tweets downloaded, with 42 variables, including tweets texts, locations, geolocations, number of favorites etc.
6. First, I present what people concerns about Victoria's Secret Fashion Show in 2016 by mining tweet texts, finding keyword frequencies and representing into Word cloud.
7. Text description statistics: word length, number of characters, number of tags, number of @mentions and number of unique words; Analysis possible relationships between.
8. Second, I reduce the dataset into tweets only in USA (with geolocations). Draw USA map with tweets distributions. Counting number of tweets in each states.
9. Look at California specific. Compare tweets length and number of unique words between geolocations in California.
10. Sentimental analysis in USA, compare people's attitude between locations
11. Sentimental analysis about Lady Gaga and Adriana Lima, compare people's attitude between locations.

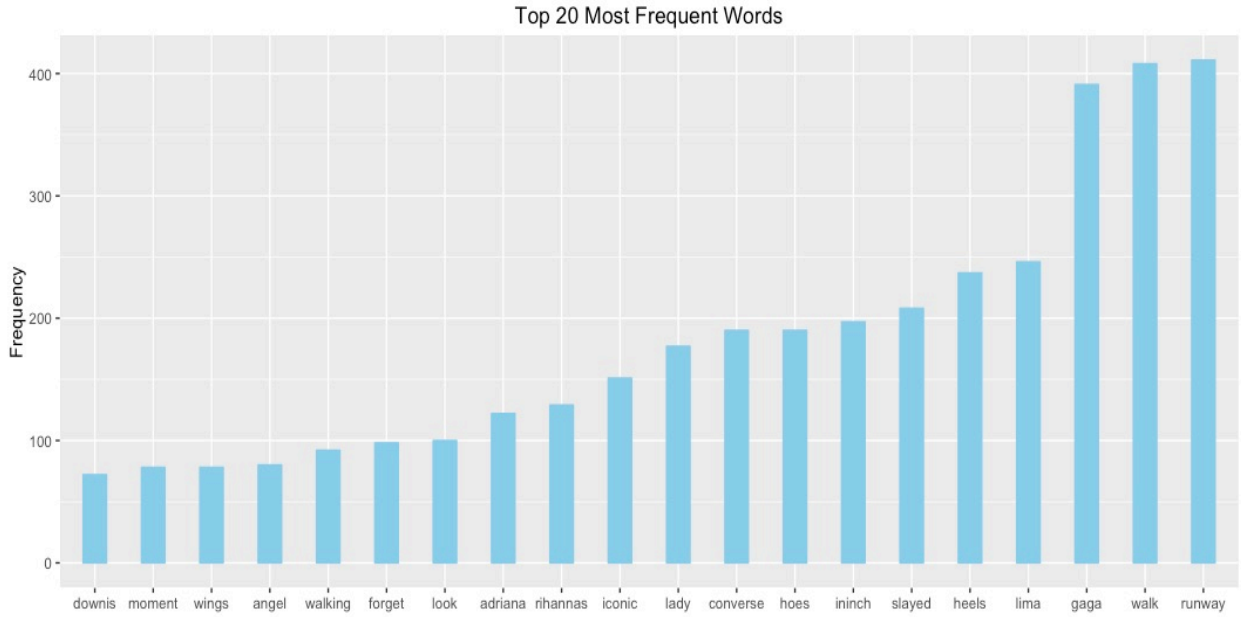
General Text Description

Word cloud

The tweet data was captured on the second day of 2016 Victoria's Secret Fashion Show was broadcasted. The topic "***VSFashionShow***" was of great concerns at that time. First, I explore how people on Twitter think of this hot fashion show. By extracting the tweet text messages from the original data frame, I cleaned the text by getting rid of URLs, hashtags, @mentions, unnecessary spaces, retweet headers and meaningless daily words (such as we, can, she, able, etc.). Package "wordcloud2" was used to generate the word cloud here. Two word clouds were generated, one is at above, the letter cloud is on the title page.

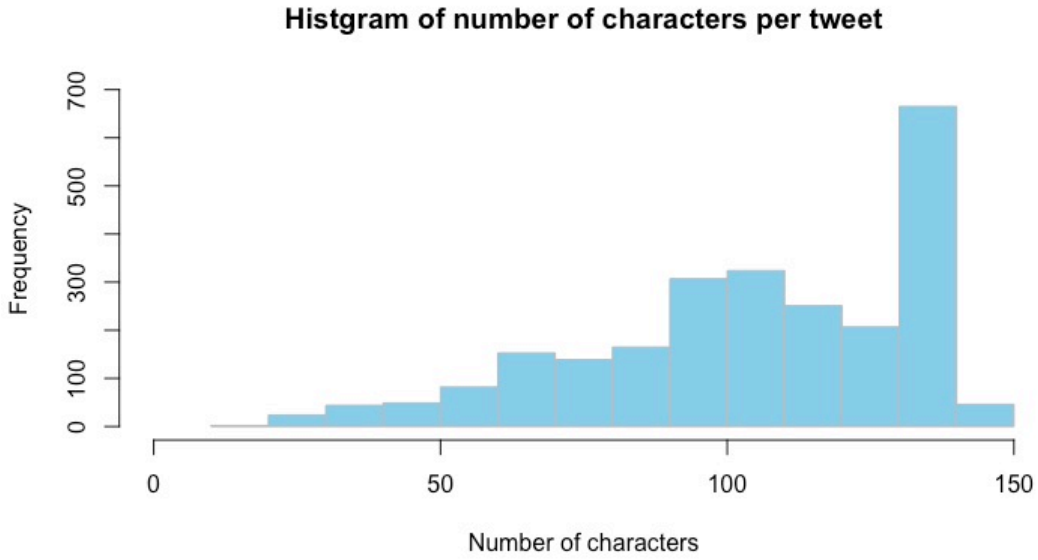
Top 20 Most Frequency Words

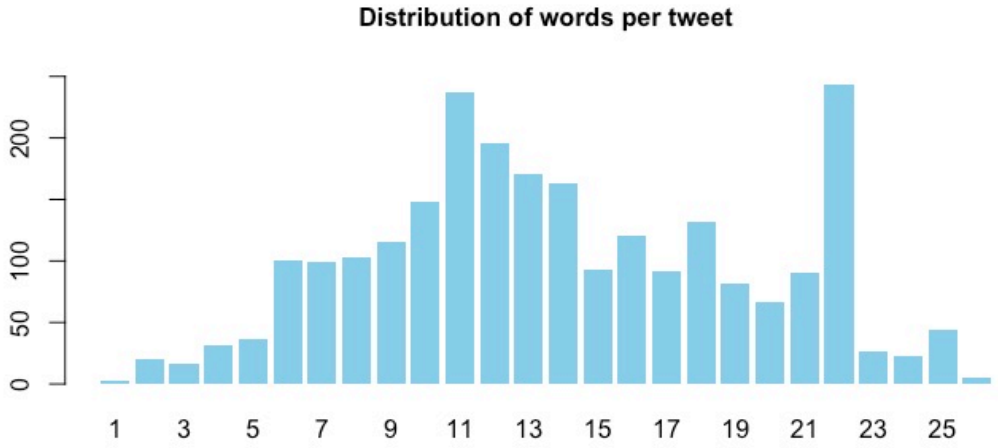
The most frequency words about *VSFashionShow* are *runway, walk, gaga, lima, heels, slayed, ininch, hoes, converse, lady, iconic, Rihanna, Adriana, look, forget, walking, angles, wings, moments, downis*. Lady Gaga and Adriana Lima are the people of greatest concerns of this fashion show this year. Rihanna were mentioned more frequent than expectation as people recalled her performance on *VSFashionShow* in previous years. I will further present sentimental analysis of Lady Gaga and Adriana Lima in the later report.

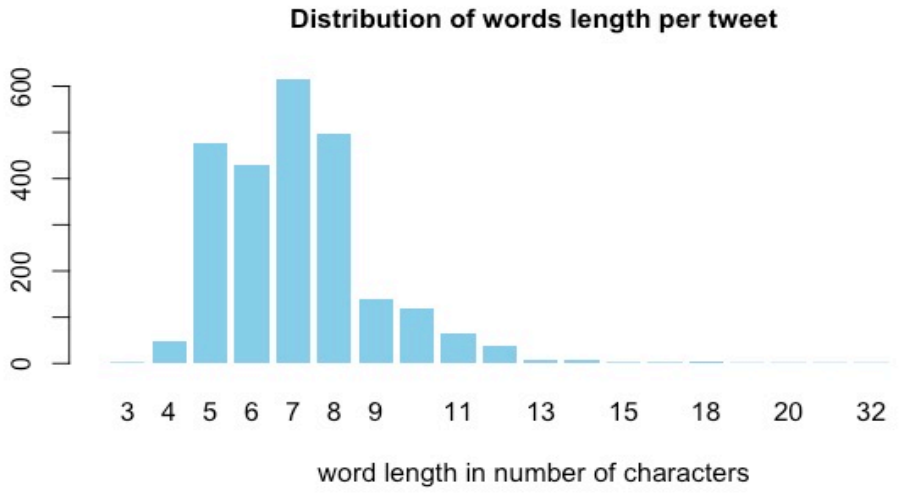


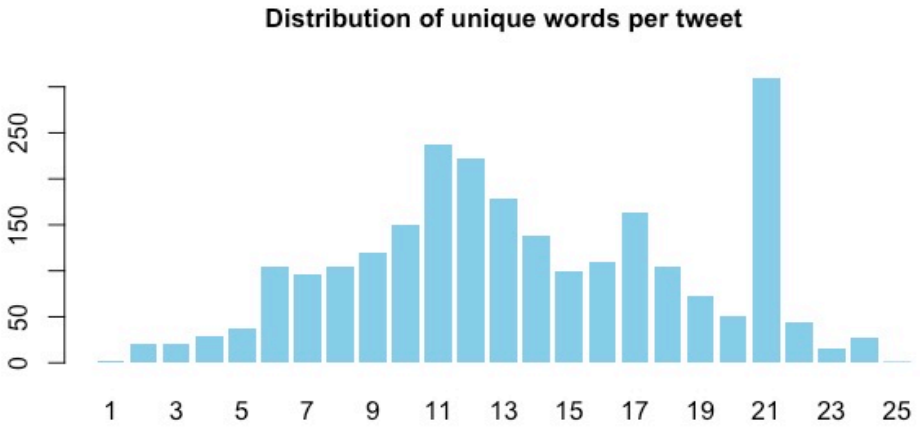
Frequency Analysis of texts

In this section, I did further frequency analysis of tweet texts in terms of number of characters per tweet, word length per tweet, number of unique words per tweet, number of hashtag per tweet, number of @mentions per tweet.




Distribution of words per tweet
0
50
100
200
1
3
5
7
9
11
13
15
17
19
21
23
25

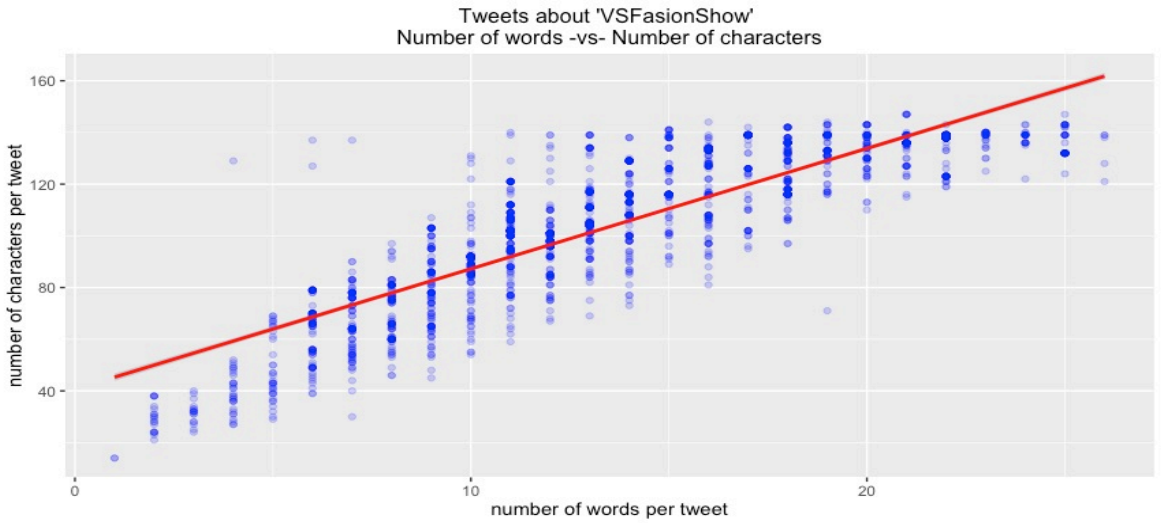

Distribution of words length per tweet
0
200
400
600
3
4
5
6
7
8
9
11
13
15
18
20
32
word length in number of characters

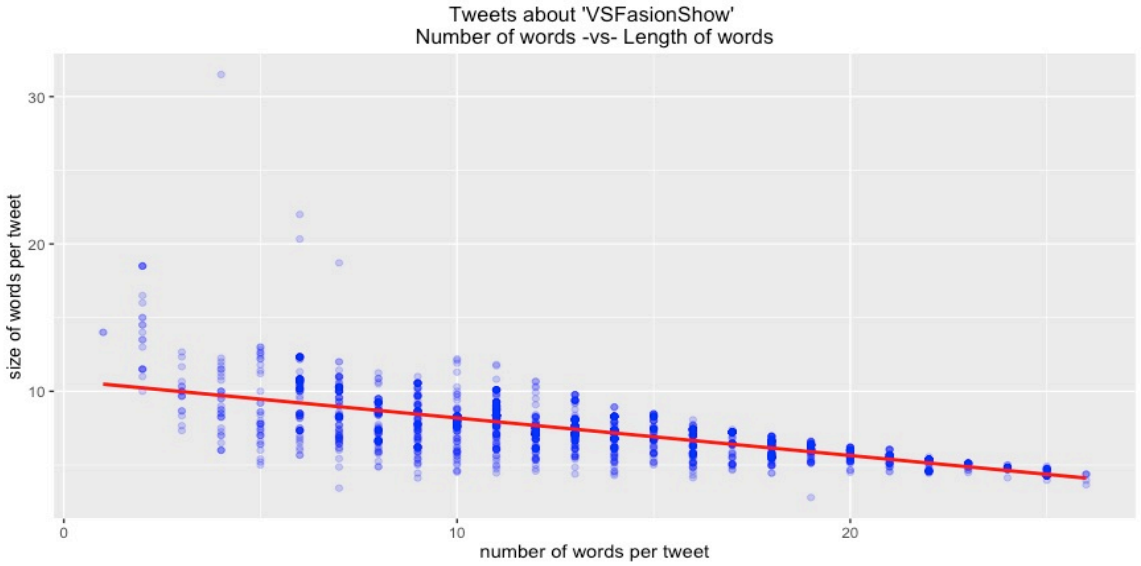

Distribution of unique words per tweet
0
50
150
250
1
3
5
7
9
11
13
15
17
19
21
23
25

Frequency	0	1	2	3	4	>4
#hashtags	323	1917	172	35	6	5
#@mention	205	1750	378	110	12	3

Table1: Frequency of number of hashtags and @mention per tweet

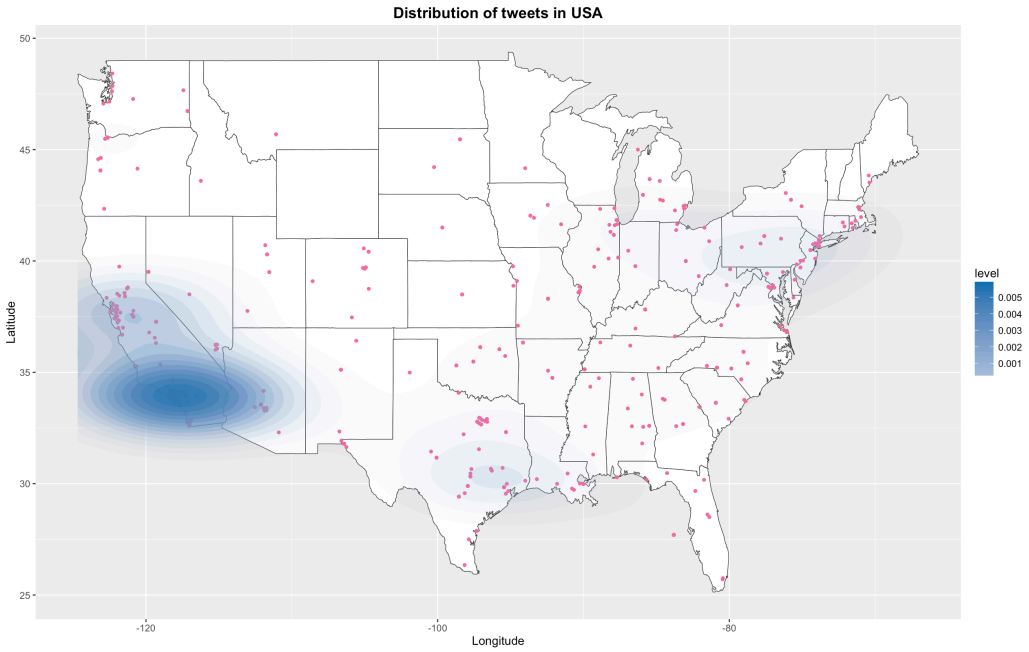
Secondly, I explored people's habits about writing tweet texts. It can be seen that as the number of words increases in one tweet, the number of character increases but the size of words per tweet decreases. These two patterns might be similar as the overall writing patterns in Twitter, not only appears in "*VSFashionShow*" topic.



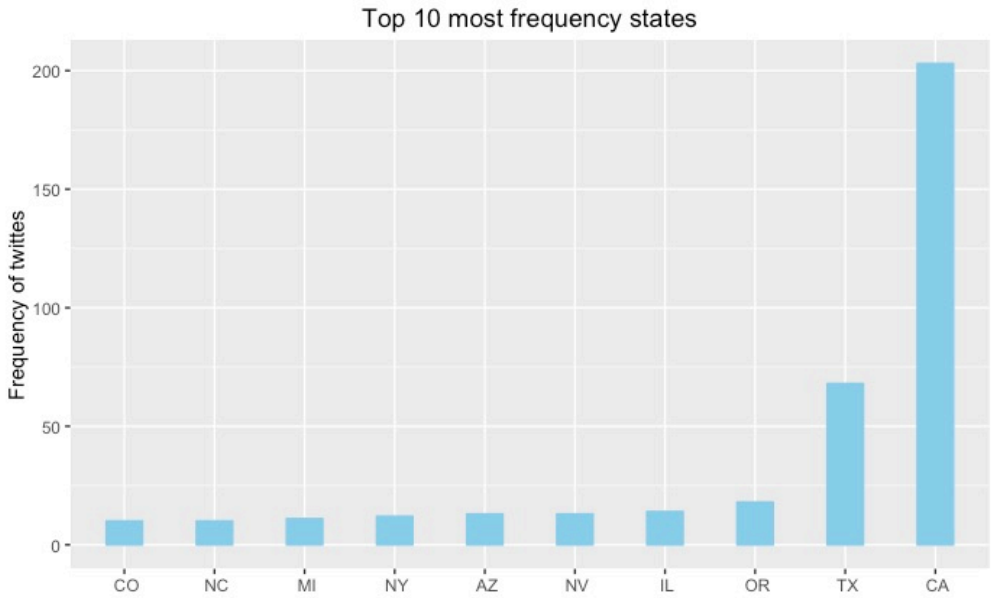


Mapping

Distribution of Tweets in USA



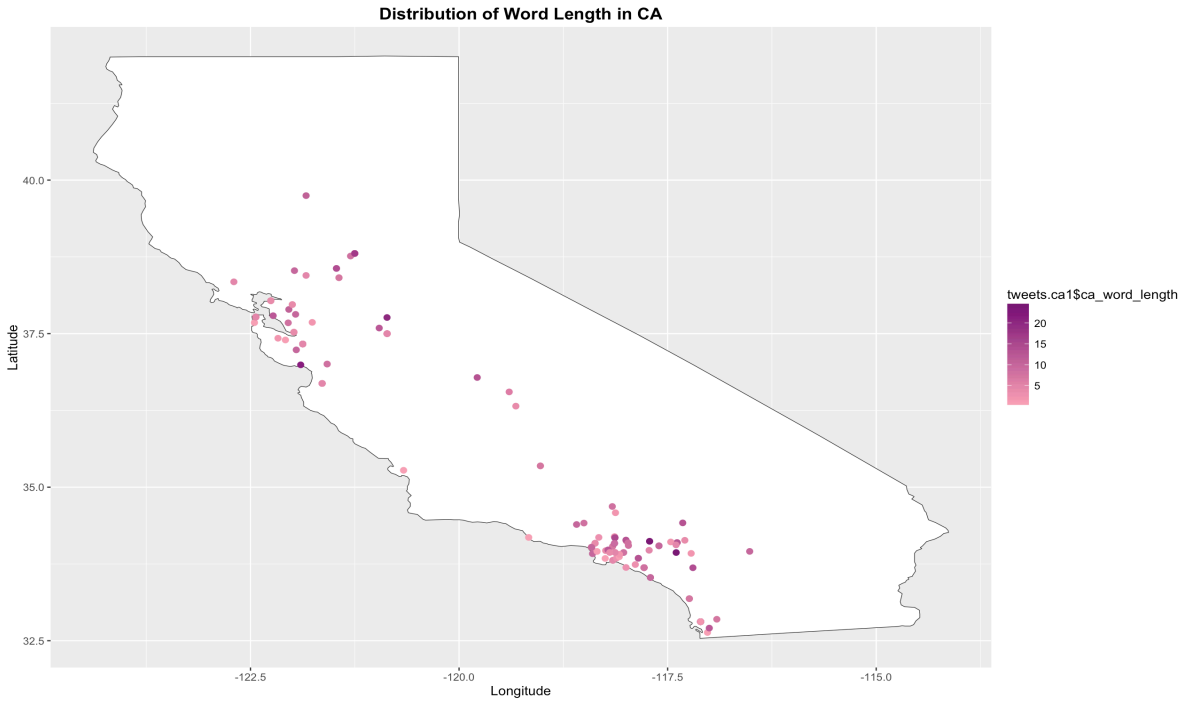
First, I generated a map of distribution of Tweets geolocations in USA, with associated heat levels. It can be seen that among all the states in USA, California is the “hottest” states about the topic “VSFashionShow”, especially Los Angeles area. Texas is the second. People in the north are less likely to concern about this fashion show.



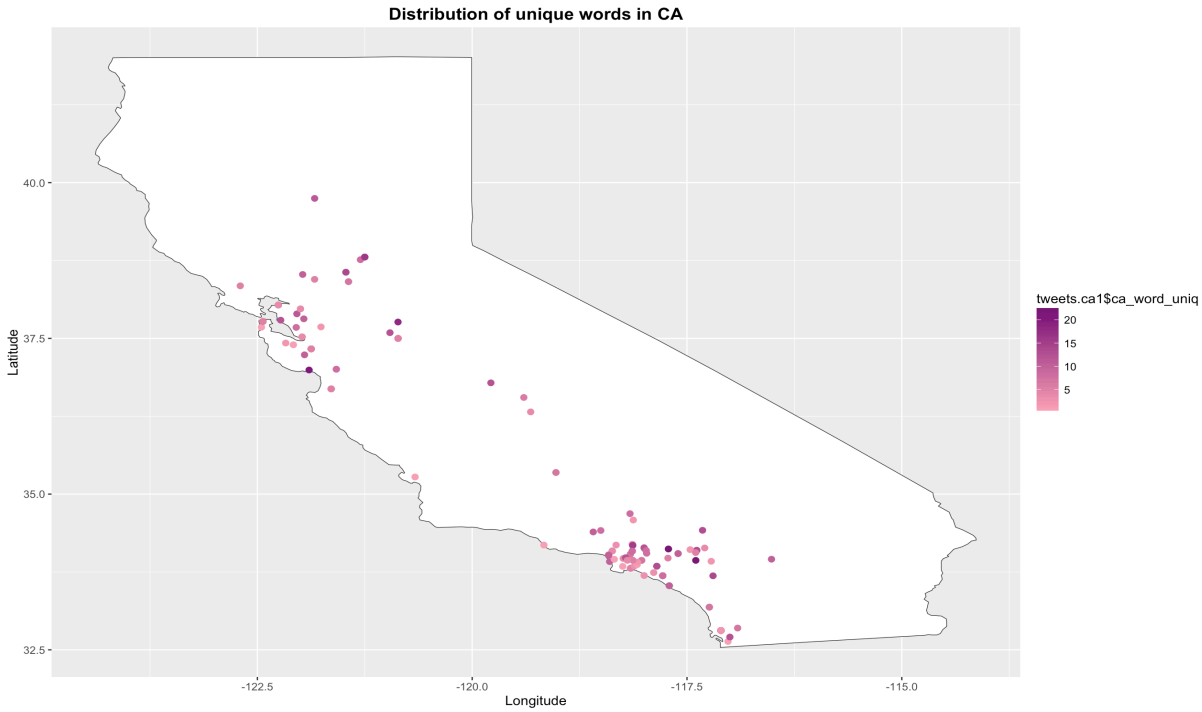
Top 10 states

As I counted the number of tweets in each state, same result appears as the heat map above. California has the largest number of tweets (approx. 200), and Texas has the second largest number of tweets. Oregon is the third, but the number of tweets captured with geolocations is under 25.

California: text frequency analysis



Most tweets in California are found around two big cities: San Francisco and Los Angeles. These two maps are generated to compare word length per tweet between locations (map1) and compare number of unique words per tweet between location (map2). It can be seen that people around Los Angeles are slightly more likely to write more words and more unique words than people around San Francisco.

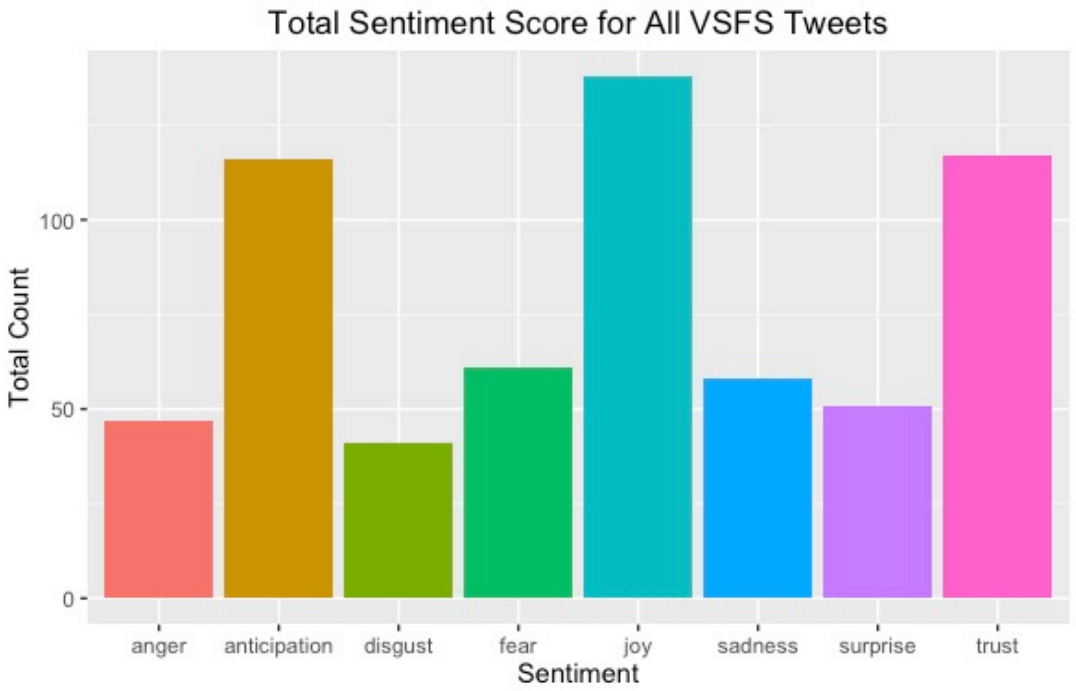


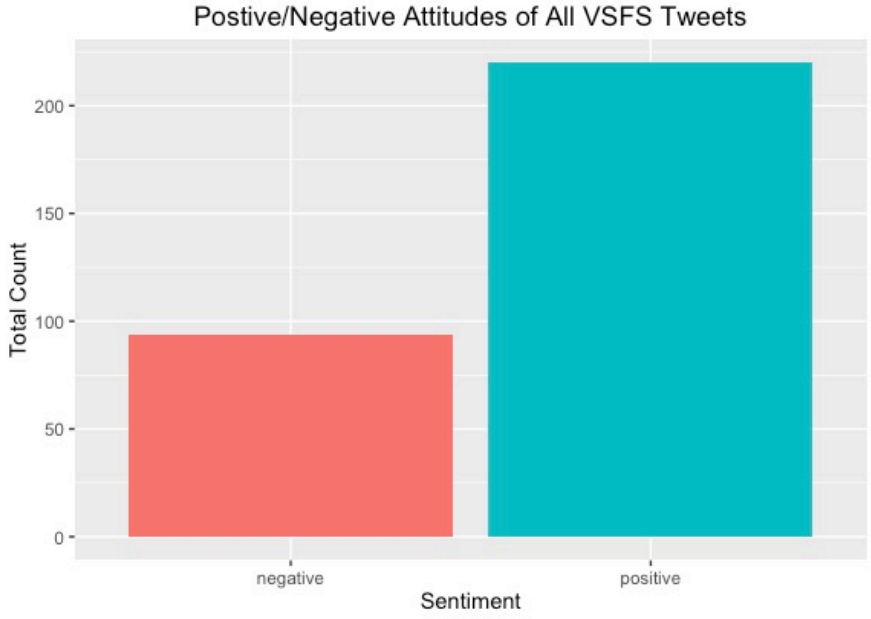
Sentiment Analysis

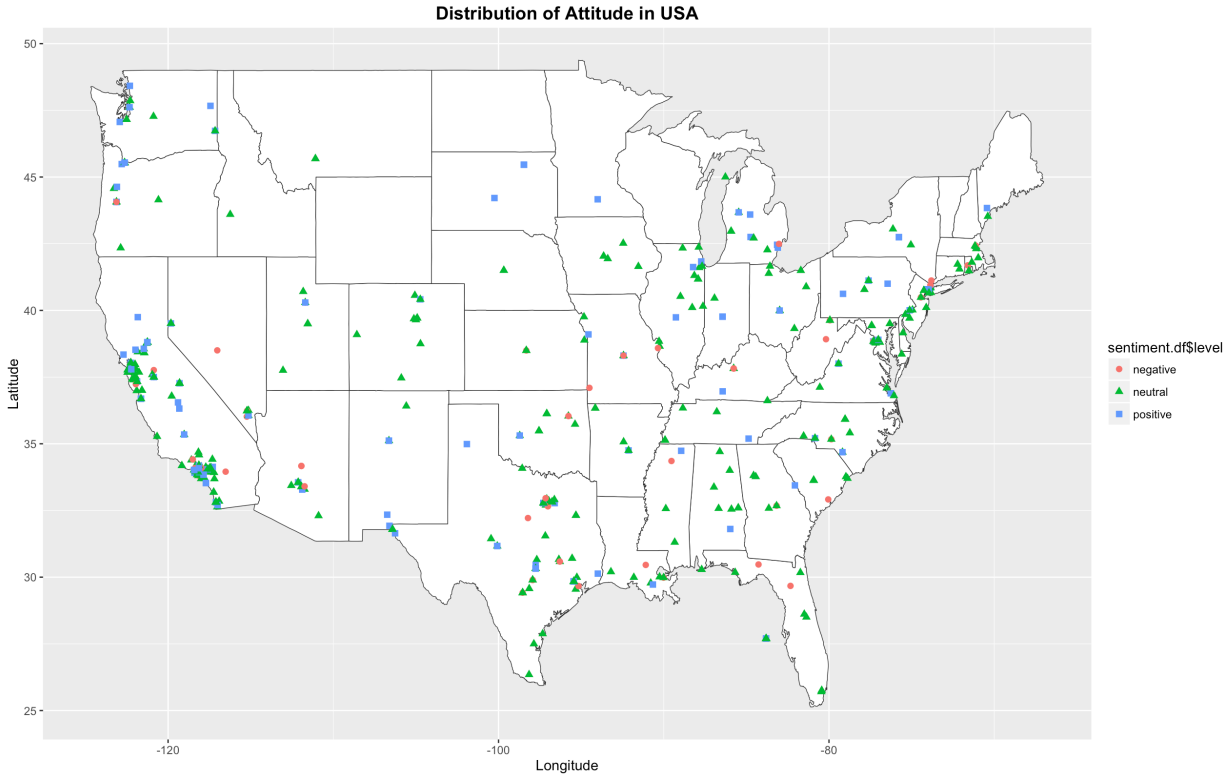
Sentiment analysis, same as opinion mining, is the process of deriving the opinion or attitude of a speaker or writer. In this project, I used sentiment analysis of *Saif Mohammad's NRC Emotion* lexicon in *Syuzhet* package. "The NRC emotion lexicon is a list of words and their associations with eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive)"(See http://www.purl.org/net/NRCemotionlexicon).

Analysis in USA

Trust, joy and anticipation are the most frequent emotions in captured tweets and also approximately 70% of people represent positive attitude toward the fashion show. Thus we can say that people who watching the show and then write tweets on Twitter are more likely to have positive opinion about the Victoria's Secret Fashion Show in 2016.


Total Sentiment Score for All VSFS Tweets
Total Count
100
50
0
anger
anticipation
disgust
fear
joy
sadness
surprise
trust
Sentiment

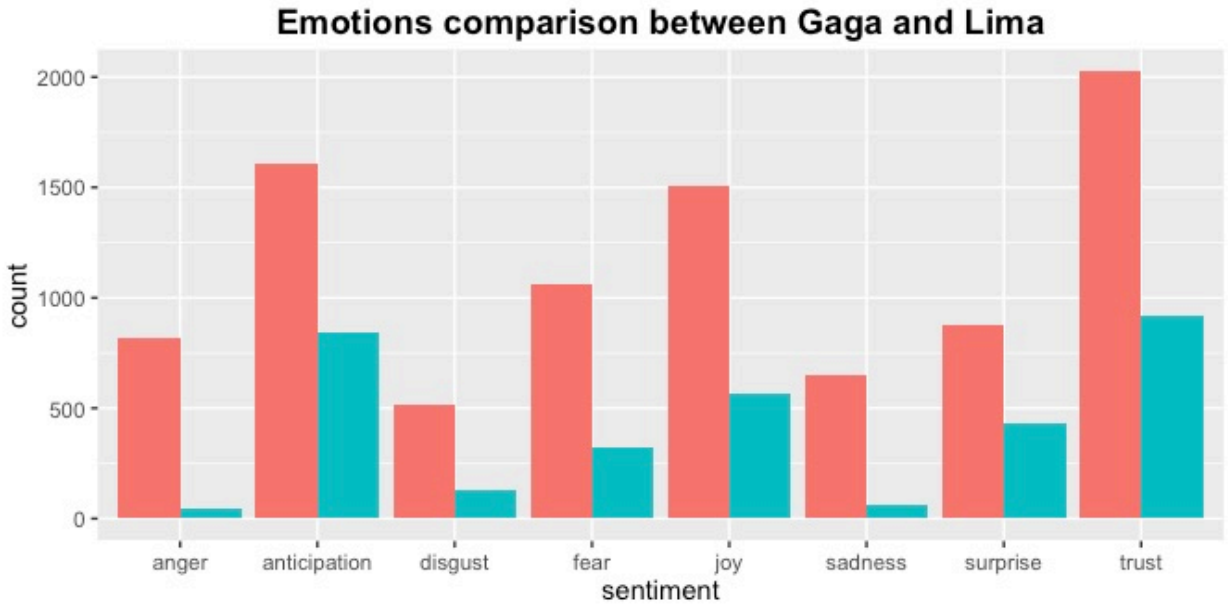

Postive/Negative Attitudes of All VSFS Tweets
Total Count
200
150
100
50
0
negative
positive
Sentiment



The level of attitude is set to be positive when number of positive words is greater than the number of negative words per tweet; negative when number of positive words is less than the number of negative words per tweet; neutral when they are equal. Based on this setting, more people have a positive attitude about the show than people with negative attitude among the country. More specifically, the different between the number of people with positive attitude and with negative attitude gets larger. People on the northeast are more likely to have a neutral attitude about the show.
This map can be also viewed in shiny app.

Analysis between Lady Gaga and Adriana Lima

From word cloud in previous section, we note that Lady Gaga is of greatest interest from people on Twitter and Adriana Lima is the top 1 popular model in the show. I did sentiment analysis on these two people to see how people on Twitter think about them. Similar emotions are found for both two people, also same as emotions about the overall fashion show. People put more positive attitude on both two persons than negative attitude.


Emotions comparison between Gaga and Lima
2000
1500
1000
500
0
count
anger
anticipation
disgust
fear
joy
sadness
surprise
trust
sentiment