Social Media Analytics

Group 9 - Thomas George Thomas, Yang Liu, Pratyush Pothuneedi

12/8/2021

1.Introduction

We take a look at 1.6 million tweets and find interesting patterns s solve our business queries. The techniques used include Text mining, sentimental analysis, probability, building a time series data from the existing data set and clustering related data on various parameters.

Data Description

The data set contains 1.6 million tweets with the following 6 fields:

- target: the polarity of the tweet (0 = negative, 2 = neutral, 4 = positive)
- ids: The id of the tweet (2087)
- date: the date of the tweet (Sat May 16 23:58:44 UTC 2009)
- flag: The query (lyx). If there is no query, then this value is NO_QUERY.
- user: the user that tweeted (robotickilldozr)
- text: the text of the tweet (Lyx is cool)

Data Acquisition

We acquire the data from Kaggle: https://www.kaggle.com/kazanova/sentiment140

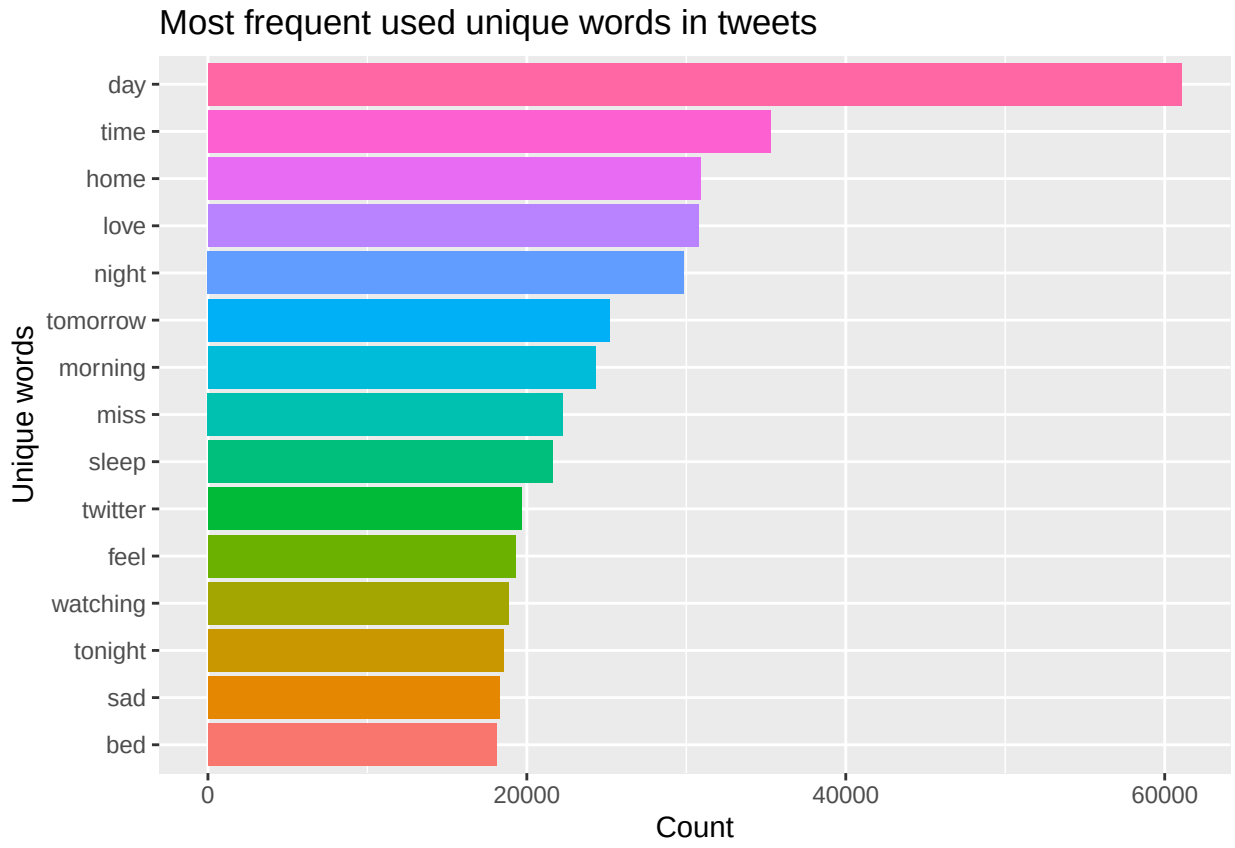
Taking 5 rows and a few columns

ids	date	text
1467810369	Mon Apr 06 22:19:45 PDT 2009	@switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer. You shoulda got David Carr of Third Day to do it. ;D
1467810672	Mon Apr 06 22:19:49 PDT 2009	is upset that he can't update his Facebook by texting it... and might cry as a result School today also. Blah!
1467810917	Mon Apr 06 22:19:53 PDT 2009	@Kenichan I dived many times for the ball. Managed to save 50% The rest go out of bounds
1467811184	Mon Apr 06 22:19:57 PDT 2009	my whole body feels itchy and like its on fire
1467811193	Mon Apr 06 22:19:57 PDT 2009	@nationwideclass no, it's not behaving at all. i'm mad. why am i here? because I can't see you all over there.

2.Analytical Questions

1. Finding the frequently used unique words

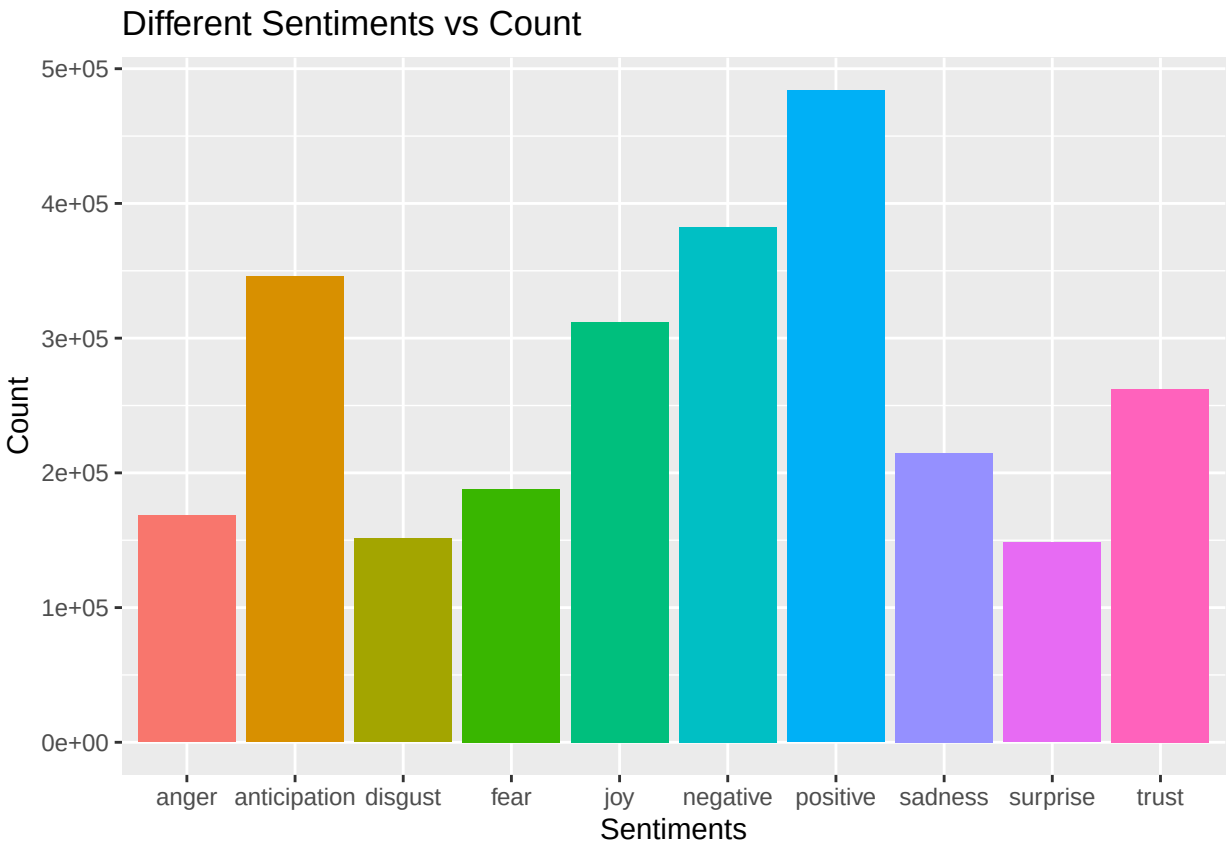
For this insight, we consider only the "original" thought of the user/author. We Remove stop words, username mentions, replies, and Re-tweets so that we only have the "original" tweets and visualize our findings.



Observation: *Day* is the most frequently used word which has been used around 63,000 times out of the total of 1.6 million tweets. Following that, the words *Time*, *Home*, *love* and *night* have been used around 30,000 times each.

2. Sentimental Trends of Tweets

By utilizing the nrc library, we find different sentiments in each of the tweets.



Observation: Positive, negative, anticipation are the top three most tweeted sentiments. Another trend is that there are equal number of Anger, disgust and surprise sentiment tweets. A lot of Users have have tweeted about issues that they fear and trust.

@Yang Please do this as your 2nd question in Time series # 3. Extract different months from the date column and determine the sentiments related to the month

Adding the month column to the dataset

```
tidy_tweets <- tidy_tweets %>%
  mutate(elements = str_split(date, fixed(" "), n=6)) %>%
    mutate(Month = map_chr(elements, 2),
           Day = map_chr(elements, 1),
           Time = map_chr(elements, 4), .keep="unused")
```

```
tidy_tweets %>%
  group_by(Day,sentiment) %>%
  summarize(Count=n()) %>%
  arrange(desc(Count)) %>%
  arrange(Day) %>%
  top_n(5)
```

```
## `summarise()` has grouped output by 'Day'. You can override using the `.groups` argument.

## Selecting by Count
```

```
## # A tibble: 35 x 3
## # Groups:   Day [7]
##    Day   sentiment    Count
##    <chr> <chr>        <int>
##  1 Fri   positive     65573
##  2 Fri   negative     52425
##  3 Fri   anticipation 46507
##  4 Fri   joy          42295
##  5 Fri   trust        35180
##  6 Mon   positive     91825
##  7 Mon   negative     70264
##  8 Mon   anticipation 65332
##  9 Mon   joy          58388
## 10 Mon   trust        51136
## # ... with 25 more rows
```