Project Report on Twitter Analysis: #OhCanada

Neha Deshmukh

12 July 2019

Project Introduction

For the purpose of Data Visualisation project, I have attempted to perform a twitter analysis of tweets associated with the hashtag #OhCanada because of Canada day which was celebrated on 1st July. I have used Twitter API and rTweet package in R to collect the tweets and then performed visual analysis on it using both R and Tableau. Questions that I have attempted to answer are:

1. Apart from Canada and US, which all countries have tweeted about Canada day?
2. Top 25 frequent words used?
3. Positive and Negative sentiments of people regarding Canada Day.
4. Words associated with each sentiment.
5. Word Network to find out pairs of words occuring together frequently.

On the occasion of Canada Day, it would be interesting to see the reactions of people on twitter regarding it and it could be useful for performing sentiment analysis to understand what people feel like doing on this day.

Design Analysis

The 4 levels of visualisation design are:

1. **Domain situation**:

It speaks about who are the target users. For my twitter analysis on Canada Day, the target users would be businesses and retailers who are looking to expand their merchandise sale related to Canada Day in other countries. They can use the sentiments analysed here to check if they will get profit on their sale on Canada Day.

2. **Abstraction**:

It deals with translating the specifics of domain to the vocabulary of visualisation. It explains about the what and why part of data analysis. To answer the what part, my project uses the twitter data that I have collected using the R language and rTweet package. The visualisation then goes on to show relation between the countries and their say about Canada day. For doing the sentiment analysis, I have converted the text into tokens and then analysed its emotions in R. I have made 4 visualisations. Two of my visualisations are in R and the rest two visualisations are in Tableau.To answer the why part, the users are looking at my visualisations to get an idea about what people generally feel about Canada Day and how small retailers can get profit on this day.

3. **Idiom**:

It deals with the how part of data analysis. I have made 2 of my visualisations in R. They are bar graph to analyse the sentiments and a wordcloud to understand the word distribution in the twitter text. The other two visualisations deals with country wise tweets and 25 most frequently used words. These are in the form of a packed bubbles and tree maps resepctively in Tableau. I have manipulated the data source such that I exported the clean text from R and then uploaded that in tableau to perform an inner join with the data containing the frequency of each word.

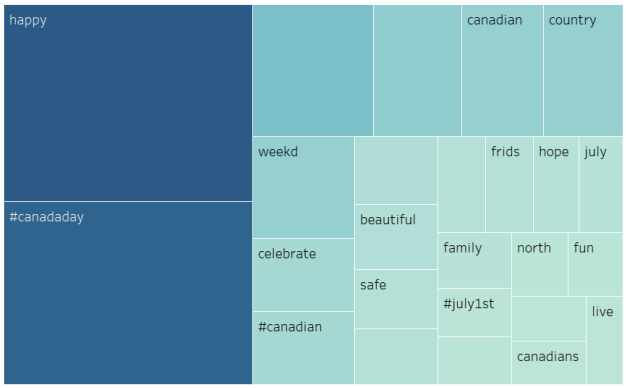
4. **Algorithm**:

The computational complexity of my visualisations is very low. Only the word network takes time to get generated because of many tweets that it needs to analyse.

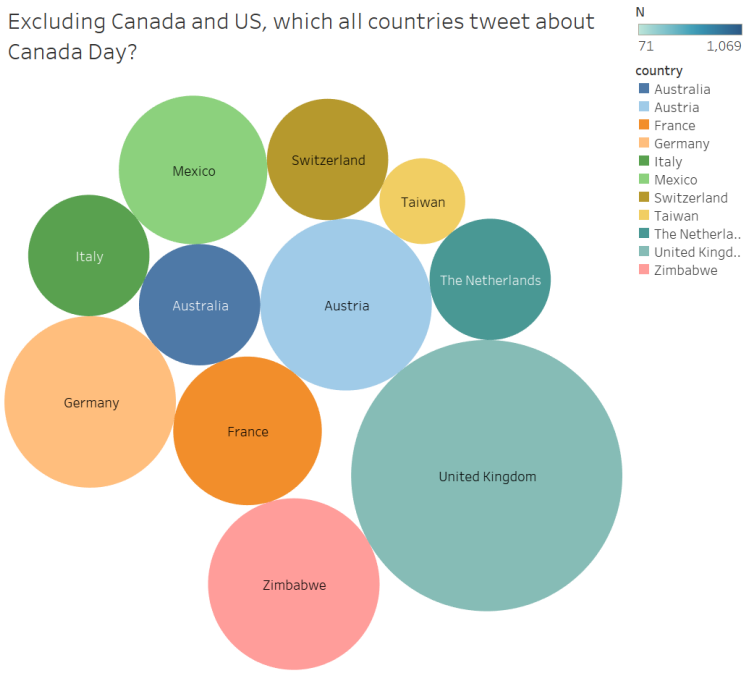
Twitter Data Analysis using Tableau

Canada Day twitter analysis

25 most used words in #ohcanada tweets



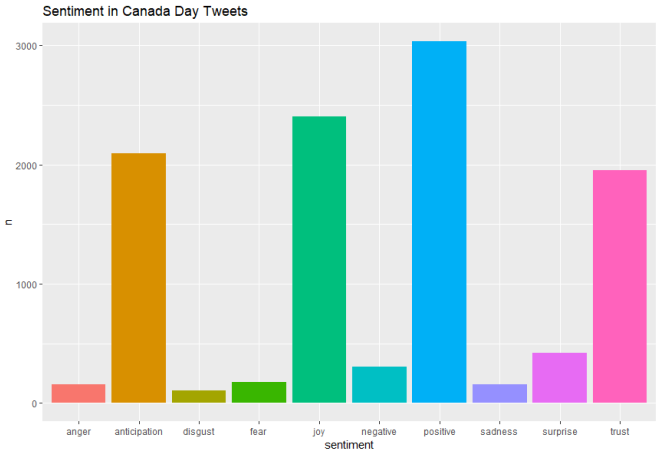
Excluding Canada and US, which all countries tweet about Canada Day?

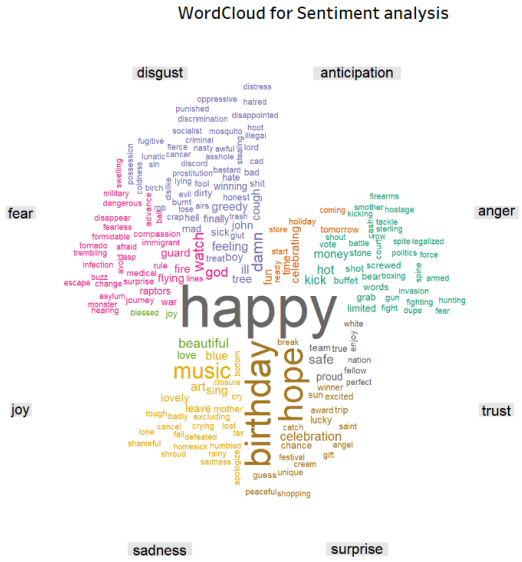


Referring to the graphs above, it is seen that 25 most frequently used words on Canada Day tweets are words like happy ,weekend, celebrate, canada and so on. Looking at the packed bubbles, it is interesting to note that apart from Canada and USA, there are several other countries tweeting about Canada Day like Mexico, taiwan and UK.

Sentiment Analysis using R

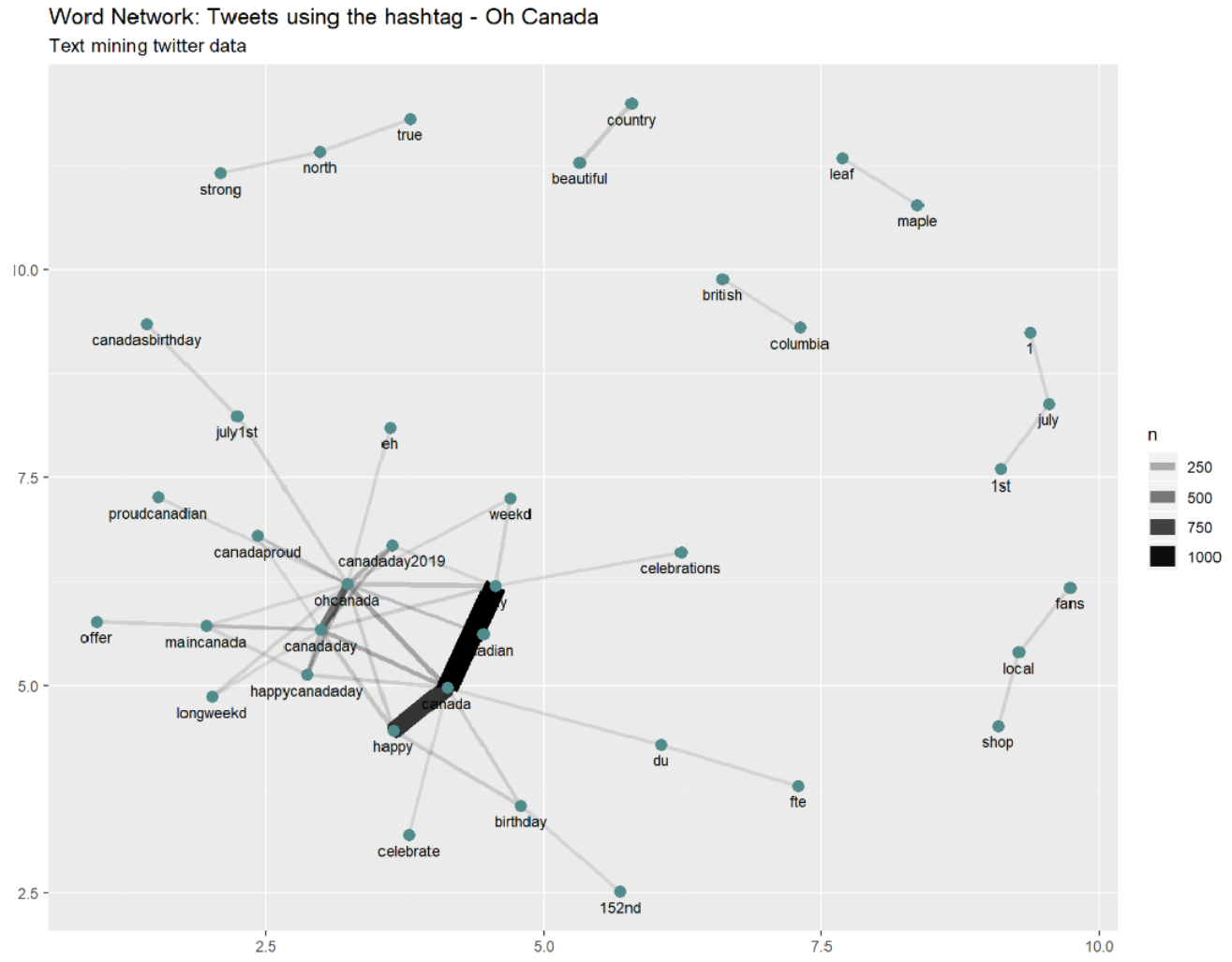
Sentiment analysis of Canada Day tweets





The two graphs shown above speak about the sentiments of the people regarding Canada Day. The bar graph shows that most of the tweets are positive and joyful. The other sentimenets are trust and anticipation. There are some negative and angry tweets too. The wordcloud shows that the most used word is happy along with birthday and hope. Some negative words used are disgust, oppressive and so on.

Word Network using R



This is the word network that I made using R. It uses packages like ggraph and igraph. It tells about the pairs of words occuring together frequently. As expected, the words happy, canada and day occur together more frequently as compared to other words. Apart from that, the words ohcanada also occurs more.

Conclusion

To conclude, through this analysis, I found out that there are many people outside Canada who are interested in Canada Day and are tweeting about it. Also, along with happy sentiments, there are some bad and negative sentiments about it among the twitter people too.

The Source Code for R visualisations and Twitter Data gathering

```
#Twitter gathering code

library(twitteR)
library(streamR)
library(ROAuth)
## install devtools package if it's not already
if (!requireNamespace("devtools", quietly = TRUE)) {
  install.packages("devtools")
}
## install dev version of rtweet from github
devtools::install_github("mkearney/rtweet")

## load rtweet package
library(rtweet)

#Authentication for rTweet
create_token(
  app = "rtweet_token",
  consumer_key = "TaEfmfVnuKODi9N5H00AmH3Gu",
  consumer_secret = "MiUzqheRhgP789bphr38tqdZTxCOXTuFWg4MlWM1c3JWJRSpgs",
```

```
  access_token = "1067108927907786753-MMymmV9akRi5BUqZX2P1Ua16XqfPLk",
  access_secret = "jiLlaL1jEqXwRPgfm5e9CqB3xGhX0G6o7Y5EdgMa0EvnB")

#Authentication using StreamR

consumerKey <- "TaEfmfVnuKODi9N5H00AmH3Gu"
consumerSecret <- "MiUzqheRhgP789bphr38tqdZTxCOXTuFWg4MlWM1c3JWJRSpgs"
accessToken = "1067108927907786753-MMymmV9akRi5BUqZX2P1Ua16XqfPLk"
accessTokenSecret = "jiLlaL1jEqXwRPgfm5e9CqB3xGhX0G6o7Y5EdgMa0EvnB"

oAuthToken <- createOAuthToken(consumerKey, consumerSecret, accessToken, accessTokenSecret)

#Pulling historical data from twitter
rt <- search_tweets("#OhCanada", n = 10000, language = "en", include_rts = FALSE)
rt
View(rt)

#Pulling streaming data from twitter
stream_tweets("#OhCanada",timeout = 60 * 60 * 6,
              file_name = "canada.json",
              parse = FALSE
)

canada <- parse_stream("canada.json")

#Adding an extra column for the method in rt dataframe
library(dplyr)
rt %>% mutate(Method = "REST technique")
#Adding an extra column for method in Mars1 dataframe
canada %>% mutate(Method = "Streaming API")
View(canada)
#Merge both dataframes
canada <- rbind(rt, canada)
canada

#Dataframe to CSV
library(data.table)
fwrite(canada, file = "E:/neha/studies/trent study material/Data Analysis with R/canada.csv")

#Code for cleaning the dataset and sentiment analysis

library(ggplot2)
library(dplyr)
library(tidytext)
library(igraph)
library(ggraph)
library(stringr)
library(wordcloud)
library(reshape2)
library(widyr)
canada <- read.csv(file="E:/neha/studies/trent study material/Data Visualisation/canada.csv", header=TRUE)
head(canada)
# Tokenising and cleaning
token.pattern <- "([^A-Za-z_\\d#@']|'(?![A-Za-z_\\d#@]))"
clean.pattern = "https|la|mfclzsfjjr|de|tco|en|amp|[[:cntrl:]]|\\'|\\!\\,|\\?|\\.|\\:"

# Cleaning the dataset
clean.tweets <- canada %>%
  select(text, country, source)%>%
  mutate(text=iconv(text, "latin1", "ASCII", "")) %>%
  mutate(text=str_replace_all(text,clean.pattern, "")) %>%
  mutate(text=str_replace_all(text,"tco","")) %>%
  mutate(text=tolower(text))
View(clean.tweets)
# tokenizing the text column
tidy.all <- clean.tweets %>%
  unnest_tokens(word, text, token = "regex", pattern = token.pattern) %>%
  filter(!word %in% stop_words$word,
         str_detect(word, "[a-z]"))

View(tidy.all)

# Calculating the frequency of each word
```

```
frequency.all<- tidy.all %>%
  count(word, sort = TRUE)

write.csv(clean.tweets, "E:/neha/studies/trent study material/Data Visualisation/clean_tweets.csv", row.name
s=TRUE)

#Sentiment Analysis using NRC
tidy.all %>%
  filter(word!="canada")%>%
  inner_join(get_sentiments("nrc")) %>%
  count(sentiment, sort=TRUE)%>%
  ggplot(aes(sentiment, n, fill=sentiment)) +
  geom_bar(stat = "identity") +
  theme(legend.position="none")+
  labs(title = "Sentiment in Canada Day Tweets")

#Wordcloud
tidy.all %>%
  filter(word!="canada")%>%
  inner_join(get_sentiments("nrc")) %>%
  filter(!sentiment %in% c("positive",
                           "negative"))%>%
  count(word,sentiment, sort=TRUE)%>%
  acast(word ~ sentiment, value.var = "n", fill = 0) %>%
  comparison.cloud(colors = brewer.pal(8, "Dark2"),
                   title.size=1.0, max.words=200)

#Building the word network

library(ggplot2)
library(dplyr)
library(tidytext)
library(igraph)
library(ggraph)
library(stringr)
library(wordcloud)
library(reshape2)
library(widyr)
library(tidyr)
canada <- read.csv(file="E:/neha/studies/trent study material/Data Visualisation/canada.csv", header=TRUE)
head(canada)
# Tokenising and cleaning
token.pattern <- "([^A-Za-z_\\d#@']|'(?![A-Za-z_\\d#@]))"
clean.pattern = "https|la|mfclzsfjjr|de|tco|en|amp|[[:cntrl:]]|\\'|\\!\\,|\\?|\\.|\\:"

# Cleaning the dataset
clean.tweets <- canada %>%
  select(text, country, source)%>%
  mutate(text=iconv(text, "latin1", "ASCII", "")) %>%
  mutate(text=str_replace_all(text,clean.pattern, "")) %>%
  mutate(text=str_replace_all(text,"tco","")) %>%
  mutate(text=tolower(text))

# tokenizing the text column
tidy.all <- clean.tweets %>%
  unnest_tokens(word, text, token = "ngrams", n=2) %>%
  filter(!word %in% stop_words$word,
         str_detect(word, "[a-z]"))

tidy.all %>%
  count(word, sort = TRUE)

canada_tweets_separated_words <- tidy.all %>%
  separate(word, c("word1", "word2"), sep = " ")

canada_tweets_filtered <- canada_tweets_separated_words %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word)

# new bigram counts:
canada_words_counts <- canada_tweets_filtered %>%
  count(word1, word2, sort = TRUE)
```

```r
# plot canada day word network
canada_words_counts %>%
  filter(n >= 18) %>%
  graph_from_data_frame() %>%
  ggraph(layout = "fr") +
  geom_edge_link(aes(edge_alpha = n, edge_width = n)) +
  geom_node_point(color = "darkslategray4", size = 3) +
  geom_node_text(aes(label = name), vjust = 1.8, size = 3) +
  labs(title = "Word Network: Tweets using the hashtag - Oh Canada",
       subtitle = "Text mining twitter data ",
       x = "", y = "")
```