AIDM7390 Data Mining and Knowledge Discovery for Digital Media
Group Project
Twitter Data Analysis

Front page:
AIDM7390

Title
100 Days Of Code

Student IDs
20465106 CHEN Xiaoqi (leader)
20426550 CAI Runlin
20465769 GUO Yuju
20449496 HUANG Zefei

1. Introduction:

AI has become a hot topic in society. More and more people pay attention to learning programming, and even self-learning code, hoping to hold the opportunity of computer science trend to improve their thinking ability and find a high-paying career. We hope to explore what programming are those people who are interested in computer science learning or what programming they want to learn, and what specific content they are interested in. So we can judge the application scenarios of computer science and the current career direction. This is also closely related to our profession.

We choose “100DaysOfCode” as our hashtag. Actually, it includes two rules, like people need to code minimum an hour every day for the next 100 days, and tweet your progress every day with the #100DaysOfCode hashtag. This topic is like a large community to share code communication, programming discussion. Many people check in on this hashtag and share their mood of code learning every day. Therefore, there is a lot of data for us to observe and analyze which part of the code people like and what specific content they are discussing.

2. Procedures:

- After we got the API, we first set the app with the name “zooloretto” and then got the access token.
- In R, we installed a “rtweet” package so that R can send requests to Twitter, then automatically retrieved the data and organize the returned data into tidy structures.
- We set the hashtag of “#100DaysOfCode” to 10000 visits, excluding retweets, and users using English as their twitter content.
- We used the access token in R to get the data we want, and we can see in the table what content the returned tweets contain.
- We use the "ggplot2" package to visualize the number and time of tweets in the past few days, and count them every 3 hours, so we can observe the #100DaysOfCode activity. And we plot frequency of tweets for five related users and all tweet statuses under this hashtag.
- We installed several packages like “tm”, “SnowballC”, “wordcloud”, “wordcloud2” and “RColorBrewer” for text mining and word cloud.
- We used “RSentiment” packages and “syuzhet” packages for sentiments analysis. We can

get the emotion of the tweet, such as positivity, enjoyment, sadness, depression, etc.

- We used Latent Dirichlet Allocationanalyswas aiming at detecting sub-topics from the general debate.
- We also installed several packages like “ggmap” packages and “maps” packages for visualizing. We added some geographic information to the twitter data using the “rtweet::lat_lng()” function, and then we used the “ggplot2::map_data()”function to get the "world" data. Finally we layered the twitter data onto the map with “ggplot2::geom_point()” by specifying the *long* and *lat* to *x* and *y*. We can see where twitter users are on the map and which countries they belong to, so we can see which countries are more interested in the topic

```r
install.packages('rtweet')
library('rtweet')

token <- create_token(app = 'Zooloretto',
                      consumer_key = 'BpRSq1TOabc2RsBOlRV3HQraB',
                      consumer_secret = 'NdSZNKMgcSS23lvXV4zHPWaqmpO4DptBLh777bK1c7wbVTDPul',
                      access_token = '1327201654018490368-IuKsF3YAZmMCInO87m5iqVdu4MnjdM',
                      access_secret = '7hA7xB6N49fcBmv8ux2ui0aQQiLQZteikrjZoS1PWVfkt',
                      set_renv = TRUE)

Code<- search_tweets('#100DaysOfCode',n=10000,include_rts = FALSE,lang='en')
Code
```

```r
#####################################################
# Information of official account of #100DaysOfCode #
#####################################################

daysoc <- lookup_users('_100DaysOfCode')
daysoc$name
daysoc$description
daysoc$followers_count

################
# search tweet #
################

data.frame(Code)
colnames(Code)
Code[100,]$text
Code[100,]$screen_name
Code[100,]$created_at
Code[100,]$retweet_count

######################
# Twitter rate limits #
######################

LCode <- search_tweets('#100DaysOfCode', n = 1000000, retryonratelimit = TRUE)
LCode
```

```r
install.packages('ggplot2')
library('ggplot2')

ts_plot(Code,'3 hours')+
  ggplot2::theme_minimal()+
  ggplot2::theme(plot.title = ggplot2::element_text(face = 'bold'))+
  ggplot2::labs(
    x = NULL, y = NULL,
    title = "Frequency of #100DaysOfCode Twitter statuses from past 4 days",
    subtitle = "Twitter status (tweet) counts aggregated using three-hour intervals",
    caption = "\nSource: Data collected from Twitter's REST API via rtweet"
  )

########################
# Get 5 users timelines #
########################

Codeuser <- get_timelines(c("_100DaysOfCode","ka11away","amanhimself","ossia","freeCodeCamp"),n = 3200)

Codeuser %>%
  dplyr::filter(created_at > "2020-11-1") %>%
  dplyr::group_by(screen_name)%>%
  ts_plot("days", trim = 1L) +
  ggplot2::geom_point()+
  ggplot2::theme_minimal()+
  ggplot2::theme(
    legend.title = ggplot2::element_blank(),
    legend.position = "bottom",
    plot.title = ggplot2::element_text(face='bold'))+
  ggplot2::labs(
    x=NULL, y=NULL,
    title="Frequency of Twitter statuses posted by 100DaysOfCode and following",
    subtitle="Twitters status (tweet) count aggregated by day from Nov 2020",
    caption="\nSource: Data collected fron Twitter's REST API via rtweet"
  )
```

```
##############################
# Text mining and word cloud #
##############################

install.packages('tm')
install.packages('SnowballC')
install.packages('wordcloud')
install.packages('RColorBrewer')

library('tm')
library('SnowballC')
library('wordcloud')
library('RColorBrewer')

Codecode.v <- VectorSource(Code$text)
Codecode.c <- SimpleCorpus(Codecode.v)

inspect(Codecode.c)

Codecode.c.p <- tm_map(Codecode.c, content_transformer(tolower))
Codecode.c.p <- tm_map(Codecode.c.p, removeNumbers)
Codecode.c.p <- tm_map(Codecode.c.p, removeWords, stopwords('english'))
Codecode.c.p <- tm_map(Codecode.c.p,removeWords,c("day"))
Codecode.c.p <- tm_map(Codecode.c.p, removePunctuation)
Codecode.c.p <- tm_map(Codecode.c.p, stripWhitespace)

inspect(Codecode.c.p)

dtm <-TermDocumentMatrix(Codecode.c.p)
m <- as.matrix(dtm)
v <- sort(rowSums(m), decreasing = TRUE)
d <- data.frame(word = names(v), freq = v)
head(d,10)

set.seed(1234)
wordcloud(words = d$word, freq = d$freq, min.freq = 1,
          max.words = 200, random.order = FALSE,random.color = TRUE, rot.per = 0.25,
          colors = brewer.pal(13,"Paired"))

################
# word cloud 2 #
################

install.packages("wordcloud2")
library("wordcloud2")
wordcloud2(d,size = 1, color = "random-light",shape = "triangle-forward")

#########
# count #
#########

top10 <- head(d, 10)
top10
barplot(freq ~ word, data = top10, width =2,border = NA,las=2,
        main = "Top 10 most frequent words",cex.main=1,col = terrain.colors(10))

###########
# network #
###########

install.packages("topicmodels")
install.packages("lubridate")
install.packages("SentimentAnalysis")
install.packages("ggpubr")
install.packages("dplyr")
install.packages("tidytext")
install.packages("quanteda")
install.packages('textdata')
library("topicmodels")
library("lubridate")
library("SentimentAnalysis")
library("ggpubr")
library("dplyr")
library("tidytext")
library("quanteda")
library('textdata')

text_code <- Codecode.c.p
text_df <- data.frame(text_clean = get("content",text_code),stringsAsFactors = FALSE)
Code$text <- text_df$text_clean
toks <- tokens(Code$text)

set.seed(30)
fcmat <- fcm(toks, context = "document", tri = FALSE)
feat <- names(topfeatures(fcmat, 30))
fcm_select(fcmat, pattern = feat) %>%
textplot_network(min_freq = 0.5)
```

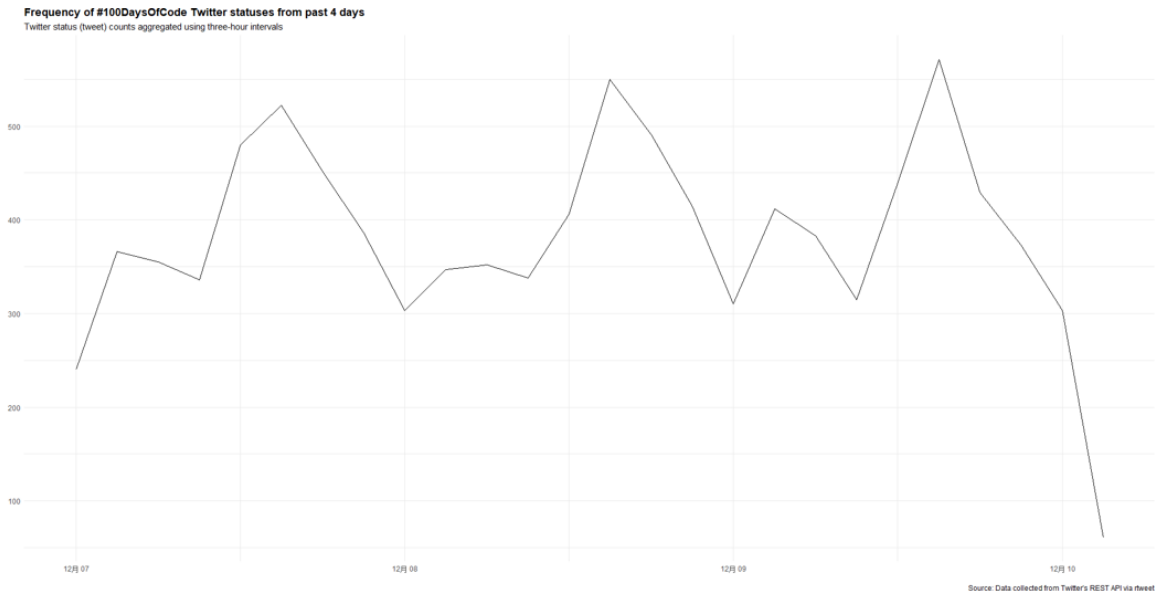
```
##########################
# Topic modelling with LDA #
##########################

set.seed(100)
samp <- sample(nrow(Code),3500)

corpus_sub <- Corpus(VectorSource(Code$text[samp]))
dtm_sub = DocumentTermMatrix(corpus_sub)
doc.length = apply(dtm_sub, 1, sum)
dtm_sub = dtm_sub[doc.length > 0,]

k <- 3
DO <- LDA(dtm_sub, k, method = "Gibbs", control = list(nstart = 5, seed = list(2003,5,63,100001,765),
                                                       best = TRUE, burnin = 4000, iter = 2000, thin = 400))

topics <- as.matrix(topics(DO))
terms <- as.matrix(terms(DO, 10))
topics_prob <- as.matrix(DO@gamma)

topics_beta <- tidy(DO, matrix = "beta")

top_terms_b <- topics_beta %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

theme_set(theme_classic())
top_terms_b %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) + labs(x = 'words', y = NULL) +
  geom_col(show.legend = TRUE) + facet_wrap(~ topic, scales = "free") +
  theme(axis.text = element_text(angle = 30, vjust = 0.5, size = 8)) + coord_flip()
```

```
######################
# Sentiment analysis #
######################

install.packages('ROAuth')
install.packages('syuzhet')
install.packages('RSentiment')
install.packages('kableExtra')
install.packages('knitr')
install.packages('RColorBrewer')
library('ROAuth')
library('syuzhet')
library('RSentiment')
library('kableExtra')
library('knitr')
library('RColorBrewer')

mysentiment_code <- get_nrc_sentiment((Code$text))
Sentimentscores_code <- data.frame(colSums(mysentiment_code[,]))
names(Sentimentscores_code) <- 'Score'
Sentimentscores_code <- cbind('sentiment'=rownames(Sentimentscores_code),Sentimentscores_code)
rownames(Sentimentscores_code) <- NULL

ggplot(data = Sentimentscores_code,aes(x=sentiment,y=Score)) +
  geom_bar(aes(fill=sentiment),stat = 'identity',width=0.7) + theme(legend.position = 'right')+
  xlab('Sentiments')+ylab('Socres')+ggtitle('Sentiments of people behind the tweets on #100DaysofCode')
```

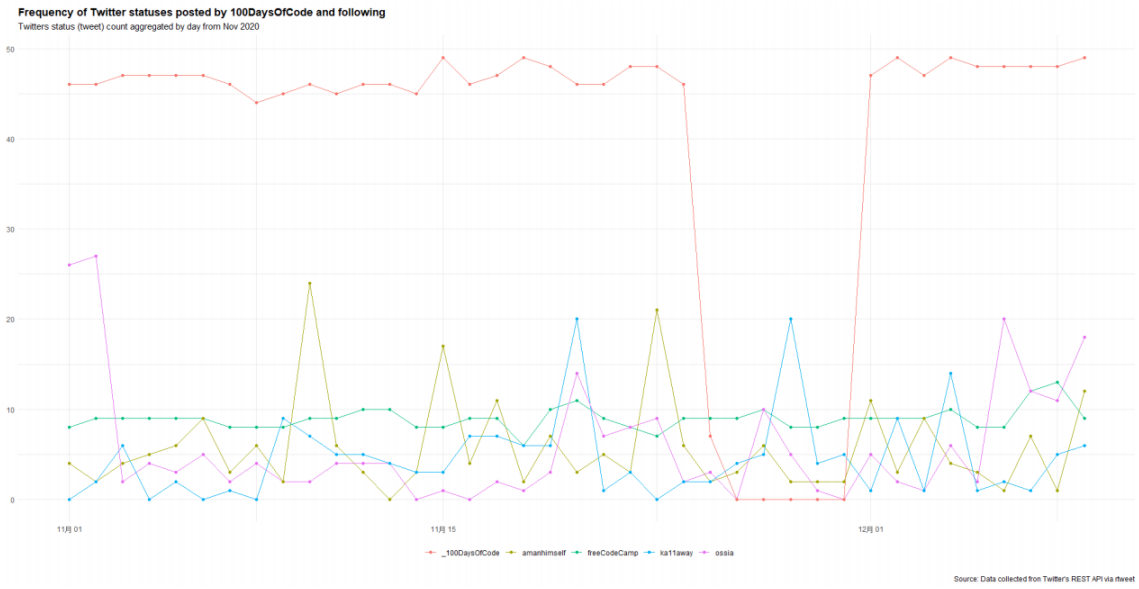
```
#######
# Map #
#######

install.packages('ggmap')
install.packages('maps')
install.packages('mapdata')
install.packages('igraph')
install.packages('gganimate')
install.packages('ggraph')
install.packages('ggalt')
install.packages('ggthemes')
library('ggmap')
library('maps')
library('mapdata')
library('igraph')
library('gganimate')
library('ggraph')
library('ggalt')
library('ggthemes')

CodeLoc <- rtweet::lat_lng(Code)
CodeLoc %>% names() %>% tail(2)
CodeLoc %>% dplyr::distinct(lng) %>% base::nrow()
CodeLoc %>% dplyr::distinct(lat) %>% base::nrow()
CodeLoc <- CodeLoc %>% dplyr::rename(long = lng)

World <- ggplot2::map_data('world')
World %>% glimpse(1000)

ggWorldMap <- ggplot2::ggplot() +
  ggplot2::geom_polygon(data = World,
                        aes(x = long,
                            y = lat,
                            group = group),
                        fill = "grey82",
                        color = "white",
                        alpha = 0.6)

gg_Code_title <- "#100DaysOfCode tweets worldwide"
gg_Code_cap <- "Tweets collected with tweet the hashtags #100DaysOfCode"

gg_Code_map <- ggWorldMap +
  ggplot2::coord_quickmap() +
  ggplot2::geom_point(data = CodeLoc,
                      aes(x = long, y = lat),
                      size = 0.7, # reduce size of points
                      color = "firebrick") + ggplot2::labs(title = gg_Code_title,caption = gg_Code_cap)

gg_Code_map
```

3. Results：

We use time plot to show the frequency of the hashtag 100DaysOfCode Twitter statuses from past 4 days. It can be seen from the figure that the number of daily releases in the past 4 days has exceeded 3000, so this hashtag has a wealth of data for us to analyze the content of the code.

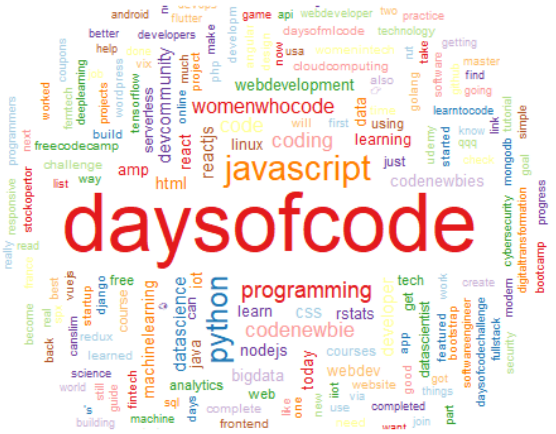


We do a tweet analysis on the official account "_100DaysOfCode" and the 4 people it followed. It can be found that "amahimself" and "_100DaysOfCode" are most closely related. When the number of "_100DaysOfCode" tweets increases, "amahimself" also increases most of the time.



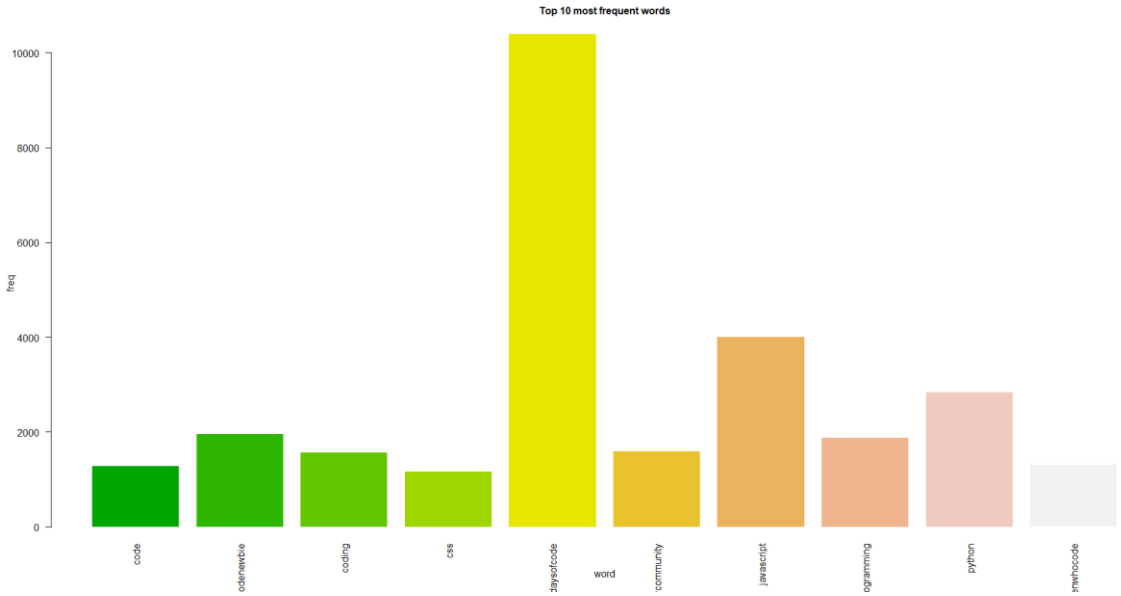
We use Twitter content to make word clouds. In addition to the core word *daysofcode*, *JavaScript* and *Python* appear the most frequently, indicating that many people are interested in these two languages. And we can find that the content of this hashtag is related to data

science, such as *html* and *webdevelopment*.

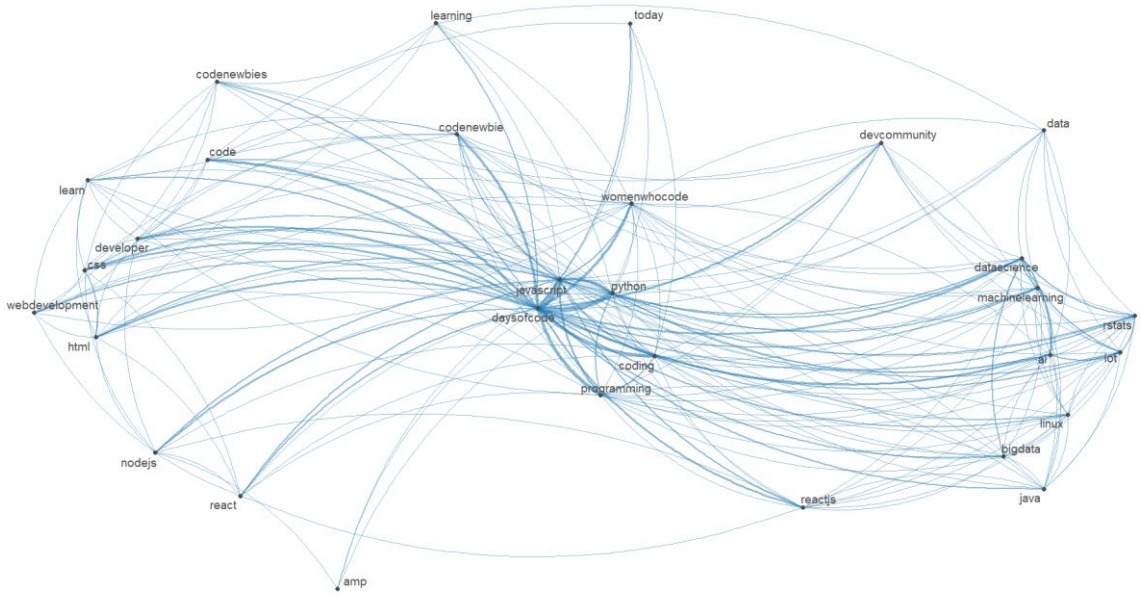


We can also see that *codenewbie* also accounts for a large proportion. It is consistent with the phenomenon we mentioned before that many people started to learn code. In the traditional impression, the code discussion area is full of males, and the appearance of *womenwhocode* indicates that more women are joining code.

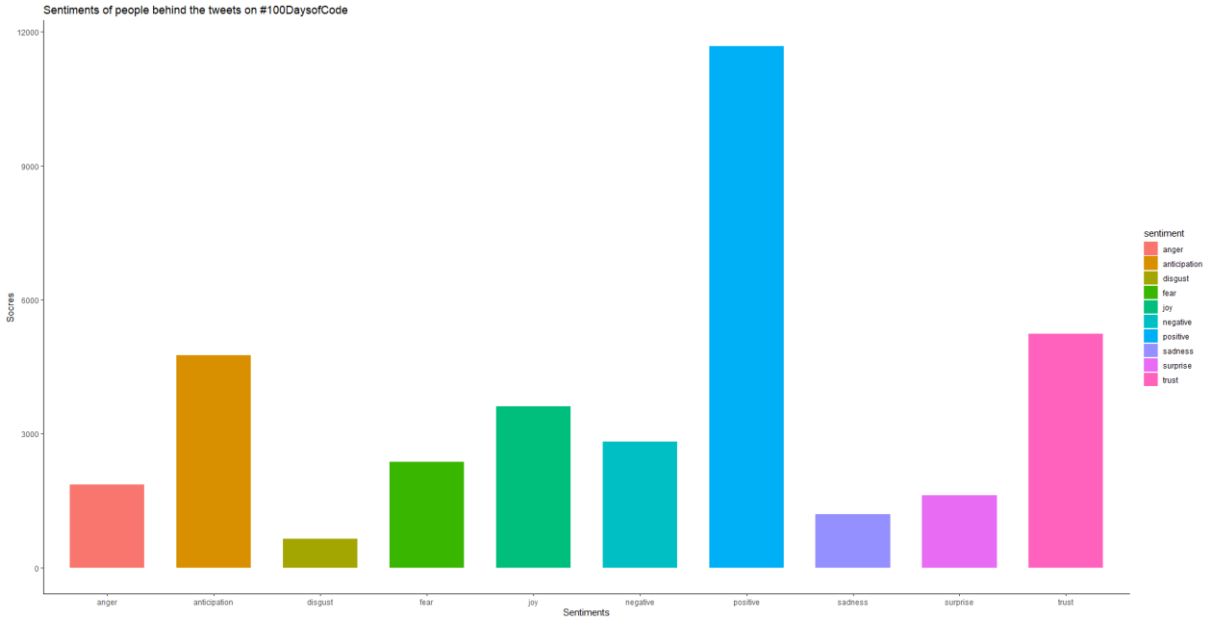
Through the bar chart below, we can more clearly see that more people learn java than those who learn python.



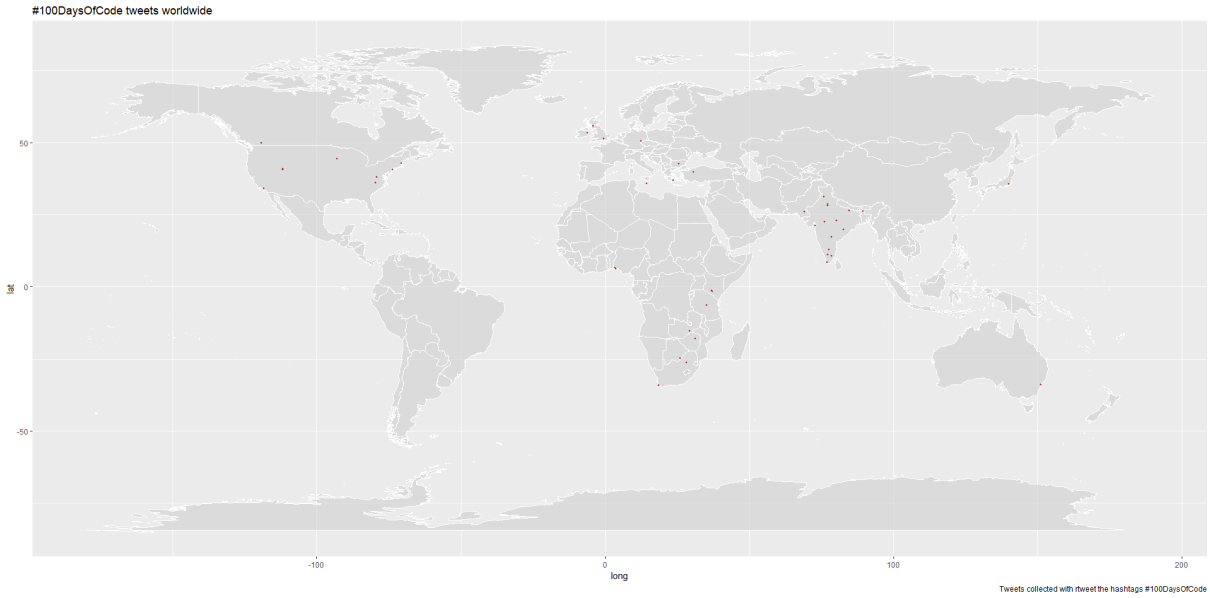
It can be seen from the node graph that the words with more occurrences are basically important nodes, and they are closely related to each other. The core word *daysofcode* has a deep connection with each node. In addition, for example, people who mention *JavaScript* usually mention *programming*, and people who mention *Python* often mention *devcommunity.* Explain that they all belong to the same "community".



We did a sentiment analysis to show everyone's tweet mood. Most people show positive attitudes on this topic, and many people also show trust that they can learn code well and look forward to learning and communicating.



We tried to explore the geographic location of the people who tweeted that with #100DaysOfCode, however, there are few tweets have latitude and longitude information, only 45 of our 9923 tweets have it. According to this map, we speculated that more people in India participated in this code event.



4. Conclusion:

From time plot we can see that there are many users participating in #100dayofdode every day. By visualizing #100DaysOfCode's Twitter content, we can see that most of the people following this topic are interested in the computer field. Most people are interested in JavaScript and Python, and many of them are working on development, projects, or machine learning. So maybe we can engage in industries such as web production, data analysis, and artificial intelligence product development. We are also seeing a lot of new entrants in programming learning, which is consistent with the rapid development of data science. Interestingly, we found that a lot of female coder also actively tweets on this topic, indicating that the gender gap in the industry may be gradually decreasing.

Most of the people involved in hashtag maintain an optimistic and positive attitude towards code or communication. People who are enjoying fun and looking forward to learning than those people in a state of confusion or sadness.

Limitations and improvements:

According to the current word cloud, we still cannot judge whether users belong to code learners or workers, nor can we see their specific employment industries.

We need to do single-person Twitter content analysis for more users, so that we can analyze the individual in a comprehensive way, then we can know their learning purpose or employment direction.