Tweets are downloaded from twitter #Uriattack for 10000 recent tweets on 11th October

Click here for code of integrating with twitter: Downloading Tweet

Code and Output for Text Analysis:

```
#/Uploading file in R/#: tweets.df<-

read.csv(file.choose(),header=TRUE)

tweets.df$created <- as.Date(tweets.df$created, format= "%d-%m-%y")
```

```
#/Remove character string between < >/#

tweets.df$text <- regex(tweets.df$text,"<",">")
```

```
#/Create document corpus with tweet text /#

myCorpus<- Corpus(VectorSource(tweets.df$text))

inspect(myCorpus[250])
```

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

[[1]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 103
```

```
#/Convert to Lower case/#

myCorpus <- tm_map(myCorpus, tolower)

inspect(myCorpus[250])
```

```
[1] vistara gives free tickets to kin of soldiers injured in uri attack
https://t.co/p1io92x8qh #uriattack
```

```
#/Remove the links URL/#

myCorpus <- tm_map(myCorpus, removeURL)

inspect(myCorpus[250])
```

```
[1] vistara gives free tickets to kin of soldiers injured in uri attack
#uriattack
```

```
#/Remove anything except the english language and space/#
```

```
myCorpus <- tm_map(myCorpus, removePunctuation)

inspect(myCorpus[250])
```

```
vistara gives free tickets to kin of soldiers injured in uri attack
uriattack
```

```
#/ Remove Stopwords /#

myStopWords<- c((stopwords('english')),c("rt", "use", "used", "via", "amp"))

myCorpus<- tm_map(myCorpus,removeWords , myStopWords)

inspect(myCorpus[250])
```

```
[1] vistara gives free tickets kin soldiers injured uri attack
uriattack
```

```
#/ Remove Single Letter Words #/

myCorpus <- tm_map(myCorpus,removeSingle)

inspect(myCorpus[250])
```

```
[1] vistara gives free tickets kin soldiers injured uri attack
uriattack
```

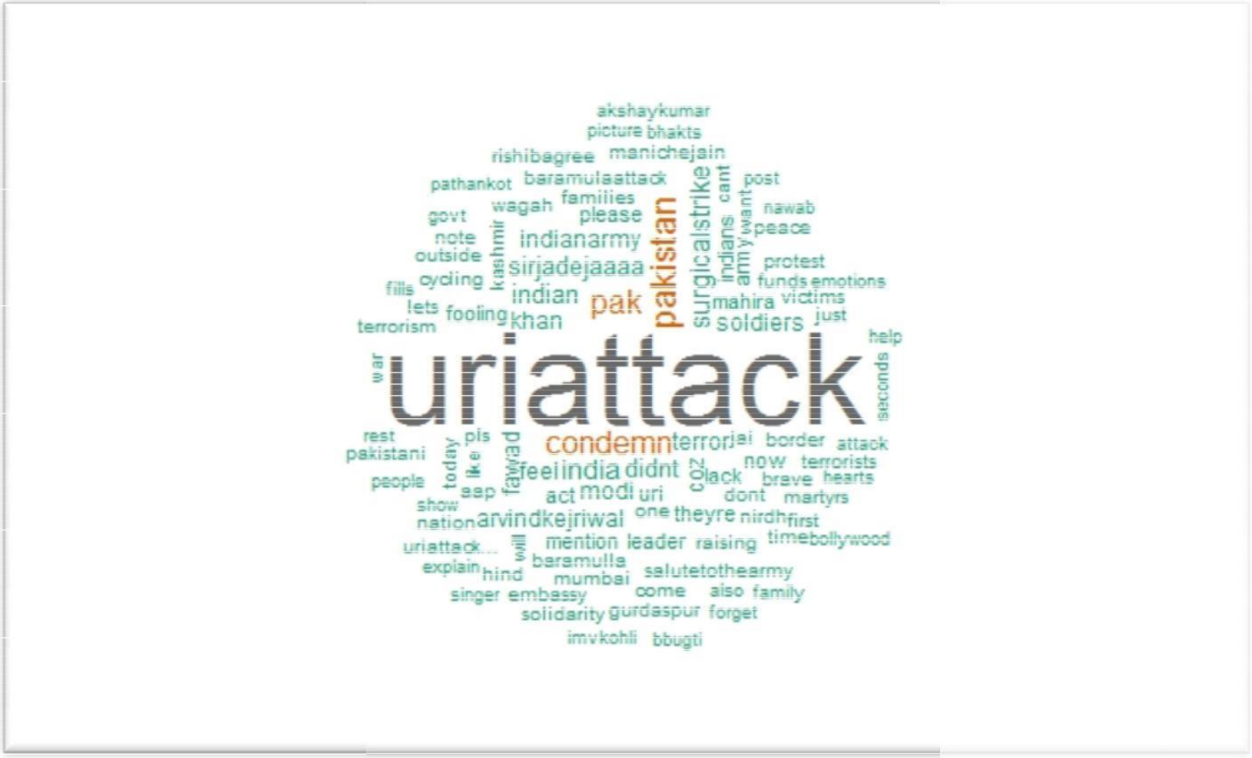
```
#/Remove Extra Whitespaces/#

myCorpus<- tm_map(myCorpus, stripWhitespace)

inspect(myCorpus[250])
```

```
[1] vistara gives free tickets kin soldiers injured uri attack uriattack
```

```
#/Stem words in the corpus/#

myCorpus<-tm_map(myCorpus, stemDocument)

inspect(myCorpus[250])
```

```
[1] vistara gives free tickets kin soldiers injured uri attack uriattack
```

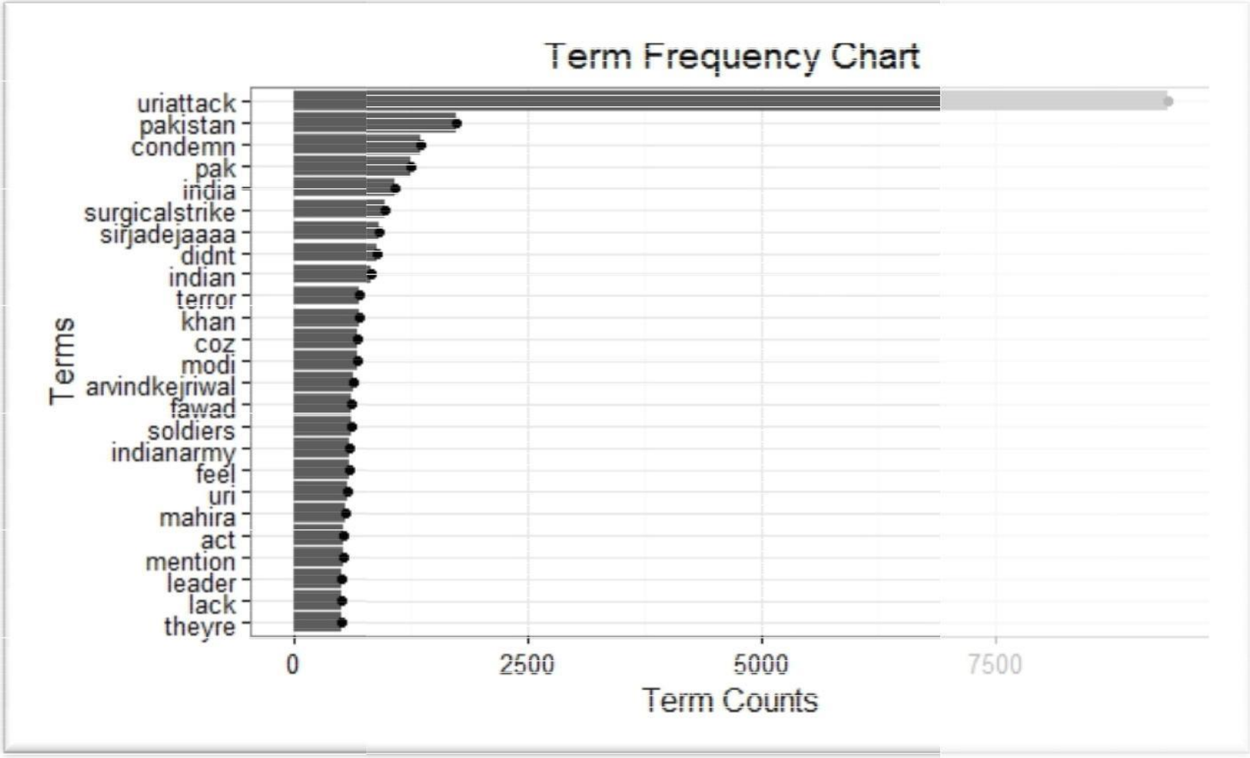
```
#/Converting to Plain Text and Term Document Matrix/#

myCorpusPT<-tm_map(myCorp us,PlainTextDocument)

dtm <- DocumentTermMatrix(myCorpusPT)

dtm2 <- as.matrix(dtm)


#/Wordcloud/#

frequency <- colSums(dtm2)

frequency <- sort(frequency, de creasing= TRUE)

install.packages("wordcloud")

library('RColorBrewer')

library('wordcloud')

words<-names(frequency)

pal<- brewer.pal(8, "Dark2")

wordcloud(words[1:100],frequency[1:100],random.order = F,colors = pal)
```



Explanation for Chart: *The chart above is showing the frequency of terms occurred in tweets. The higher frequent terms font are bigger and less frequent terms font are lower.Fro m the chart it is directly inferred that* ***uri attack is related with Pakistan, uri attack is condemen ed, uriattack is***

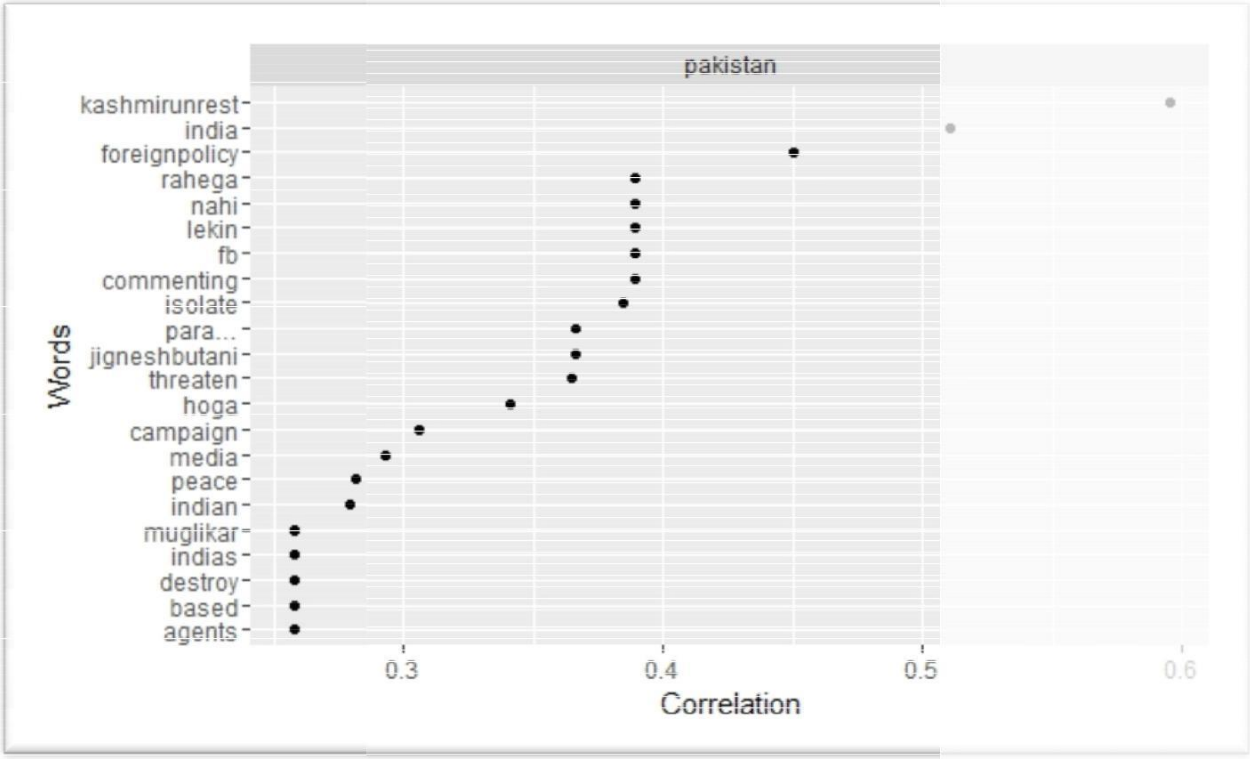
***related with surgical strike** etc, also here the words appeared are those whose f requency is more than100.*

```
idx <- which(dimnames(dtm)$Te rms %in% c("uriattack", "pakistan"))

as.matrix(dtm[idx,91:100])


#/Find the terms used most freq uently/#

freq.terms <- findFreqTerms(dt m, lowfreq = 200)

term.freq <- colSums(as.matrix( dtm))

term.freq <- subset(term.freq, t erm.freq > 500)

df <- data.frame(term = names(t erm.freq), freq= term.freq)


#/plotting the graph of frequent terms/#

p <- ggplot(df, aes(reorder(term, freq),freq)) + theme_bw() + geom_bar(stat = "i dentity")
+ coord_flip() +labs(list(title="Ter m Frequency Chart", x="Terms", y="Term Counts"))

p <- p+ geom_point()

print(p)
```



Explanation for the chart: *The chart is showing the Frequency of Terms in bar chart whose frequency is greater than 500 ie it means t he number of appearance is greater 500.*

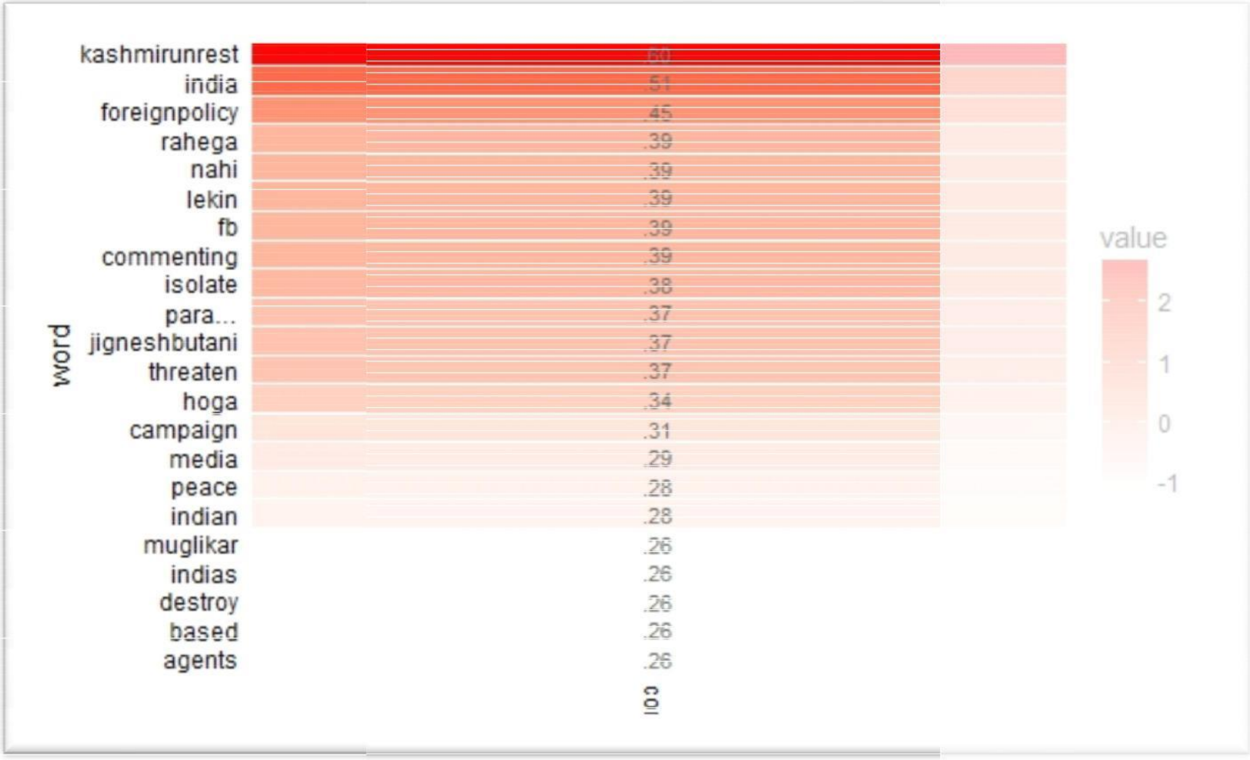
```
# Identify and plot word correlations. For example - pakistan#

WordCorr <- apply_as_df(myCorpus[1:500], word_cor, word = "pakistan", r=.25)

plot(WordCorr)
```



Explanation for the chart: *This chart is showing the correlation of word Pakistan with other words in recent 1500 tweets where correlation is more than 0.25. Correlation in text analysis means how those words are coming together in tweets. Correlation value of 1 means the word is always going together while correlation of 0 means the words never comes together. Here* ***Pakistan*** *and* ***Kashmir Unrest*** *are highly correlated showing correlation of 0.6. It means that after converting it into document term matrix form presence of words in a same row(tweet) showing correlation so correlation of 0.6 means 60% of recent 1500 tweets contains both* ***Pakistan*** *and* ***Kashmir unrest.***

Heat Chart for showing Correlattion :

```
(vect2df(WordCorr[[1]], "word", "cor"), values=TRUE, high="red",
   digits=2, order.by ="cor", plot = FALSE) + coord_flip()
```



Explanation of the chart: *The heeat chart is also depicting correlation between the word* ***Pakistan*** *and other given words. As the correlation increases its colours are getting deep red shhowing correlation is increasing.*

# Tweets with word - pakistan

```
df <- data.frame(text=sapply(myyCorpusPT, `[[`, "content"), stringsAsFactors=FALSE)

head(unique(df[grep("pakistan", df$text), ]), n=10)
```

```
[1] "URIATTACK SURGICALSTRIIKES BE1 CHINA GAME GET GREATER CONTROL PAKISTAN
INDIA ECONOMIES ROUTINE SKIRMISH THINK"
 [2] "WHATS UR TAKE BOYCOTTTING PAKISTANI ARTISTS INDIA URIATTACK IA MSRK ASKSRK"
 [3] "URIATTACK BARAMULLAAATTACK FIRINGCONTINUES CLEARLY SHOWS DAT PAKISTAN
DOESNT WANT PEACE WANTS REST PEACE HAPPYDUSSEHRA"

 [4] "PLEASE TELL IM ONE SEES IRONY INDIA CHINA PAKISTAN URIATTACK "
 [5] " MUGLIKAR GIVE PEACE CHANCE DESTROY PAKISTAN INDIA BASED MEDIA AGENTS
URIATTACK"                                                        [6] " INDIAN
BCHA PARTY CONDDEMN URIATTACK DUSSEHRA PAKISTANARMY NARENDRAMODI ADGPI HIGHTS
CREATIVITY "
```

```
[7] " PRASHANTEMPIRES HERES PAKISTANI PEOPLE THINK INDIA SURGICALSTRIKES MODI
MNS WORLDMENTALHEALTHDAY URIATTACK "
[8] "HERES PAKISTANI PEOPLE THINK INDIA SURGICALSTRIKES MODI MNS
WORLDMENTALHEALTHDAY URIATTACK "
[9] " MEDIA REPORTING DIRECTLY INDOPAK BORDER LOC PAKISTAN DOESNT NEED INTEL
PAMPORE URIATTACK"
[10] " ANINEWS MANOJ TIWARI HOLDS PROTEST NEAR PAKISTAN HIGH COMMISSION
DELHI URIATTACK "
```

Find association with a specific keyword in the tweets – #uriattack of “pakistan”

findAssocs(dtm, "pakistan", 0.2)

```
$pakistan
     gurdaspur   rishibagree        blamed         hours investigation
without
          0.32          0.32          0.30          0.28          0.27
0.27
      accepted    akbackspak        invite kashmirunrest         nawaz
popular
          0.26          0.26          0.26          0.26          0.26
0.26
        reason          nana      patekars    terrorists        within
sharif
          0.26          0.25          0.25          0.25          0.24
0.23
      display…     exclusive        vulgar          inch        peace…
          0.22          0.22          0.22          0.21           0.2
```

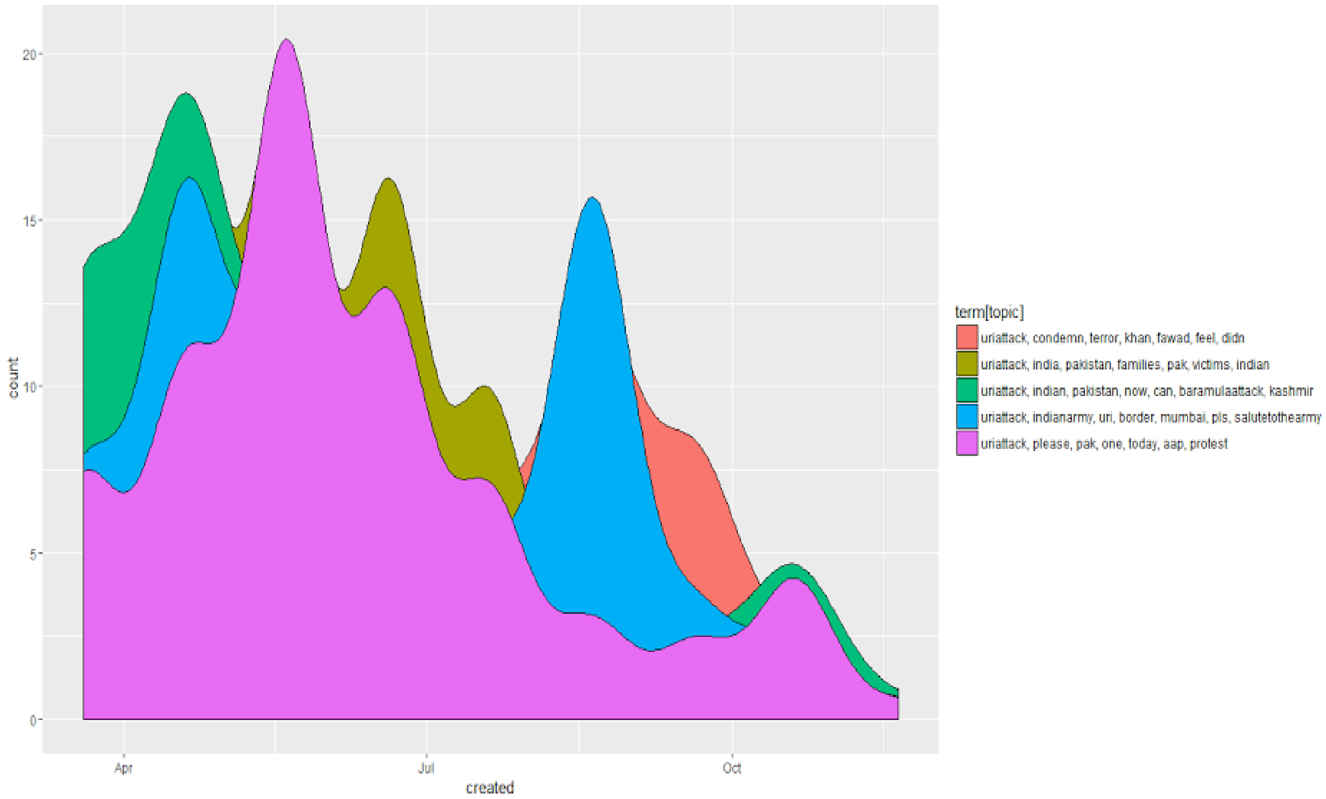
The above value shows that for entire 10000 tweets Pakistan is highly correlated to gurdaspur and rishibagree.

```
#Topic Modelling to identify latent/hidden topics using LDA technique

dtm <- as.DocumentTermMatrix(dtm)

rowTotals <- apply(dtm , 1, sum)

NullDocs <- dtm[rowTotals==0, ]
dtm    <- dtm[rowTotals> 0, ]

if (length(NullDocs$dimnames$Docs) > 0) {
  tweets.df <- tweets.df[-as.numeric(NullDocs$dimnames$Docs),]
}

lda <- LDA(dtm, k = 5) # find 5 topic
term <- terms(lda, 7) # first 7 terms of every topic (term
<- apply(term, MARGIN = 2, paste, collapse = ", "))
```

```
"uriattack, pak, arvindkejriwal, modi, surgicalstrike, pakistan, didnt"
    Topic 2 "uriattack, uri, indianarmy, pakistan, border, india, pls"
                                                                Topic 3

                  "uriattack, victims, kashmir, cant, brave, dont, india"
```

```
                                                        Topic 4
       "uriattack, condemn, terror, khan, sirjadejaaaa, fawad, feel"
                                                        Topic 5
              "uriattack, pakistan, indian, come, lets, one, time"
```

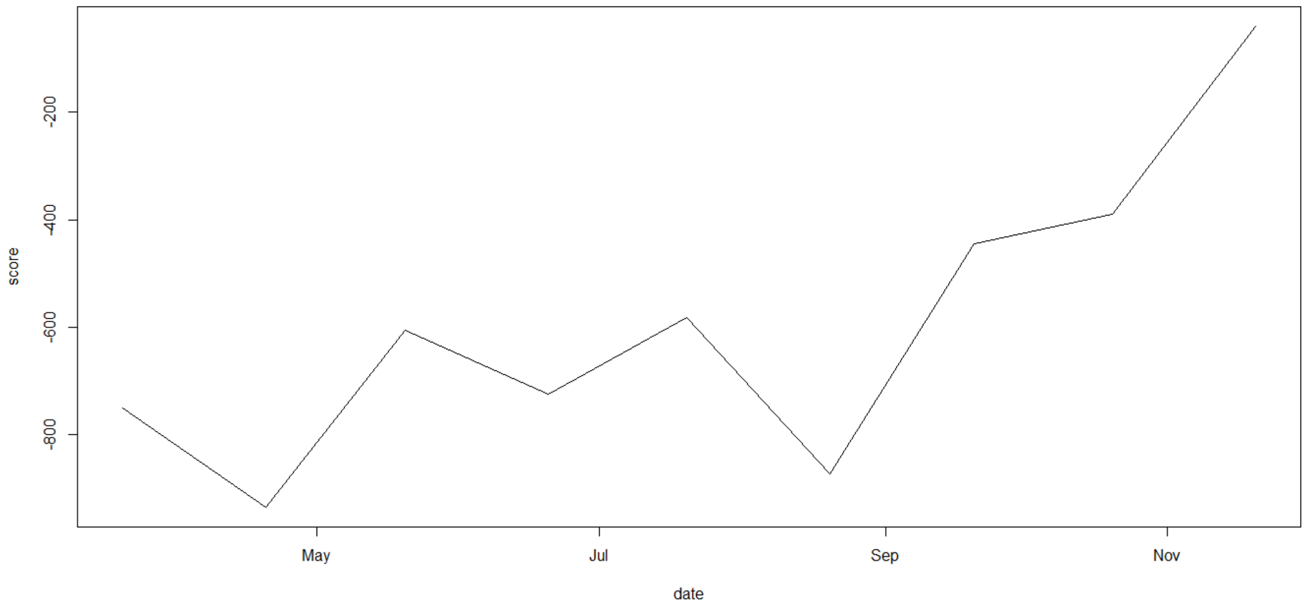
```
topics<- topics(lda)

created <- as.Date(tweets.df$created, "%m-%d-%y")

topics<- data.frame(created, topic = topics)

qplot (created, ..count.., data=topics, geom ="density", fill= term[topic], position="stack")
```



```
#Sentiment Analysis

sentiments <- polarity(tweets.df$text)

sentiments <- data.frame(sentiments$all$polarity)
sentiments[["polarity"]] <- cut(sentiments[[ "sentiments.all.polarity"]], c(-5,0.0,5), labels =
("negative","positive"))
```

```
table(sentiments$polarity)


tweets.df$created
install.packages("data.table")
library(data.table)
sentiments$score<- 0
sentiments$score[sentiments$polarity == "positive"]<-1
sentiments$score[sentiments$polarity == "negative"]<- -1


sentiments$date <- as.IDate(tweets.df$created, "%m-%d-%y")
result <- aggregate(score ~ date, data = sentiments, sum)
plot(result, type = "l")
```

```
negative positive 
    7672     2328 
```



```
#Stream Graph for sentiment by date

Data<-data.frame(sentiments$polarity)
colnames(Data)[1] <- "polarity"
Data$Date <- tweets.df$created
```

```
Data$text <- NULL
Data$Count <- 1
attach(Data)
graphdata <- aggregate(Count ~ polarity + as.character.Date(Date),data=Data,FUN=length)
colnames(graphdata)[2] <- "Date"
str(graphdata)
```

```
'data.frame':	18 obs. of  3 variables:
 $ polarity: Factor w/ 2 levels "negative","positive": 1 2 1 2 1 2 1 2
1 2 ...
 $ Date    : chr  "03-20-10" "03-20-10" "04-20-10" "04-20-10" ...
 $ Count   : int  1017 267 1492 557 1316 710 1001 277 769 187 ...
```