Fun Text Mining Project

Xi Fang

6/23/2020

A survey was conducted among sports fan about their impressions on convention sports and e-sport. This project uses modified data from Hui Du at UGA.

Let's dig into their response to see what are some common impressions

```
df1 <- read.csv("words.csv")
# combine columns 2 to 21
col <- colnames(df1)[2:21]
df1 $text <- apply(df1 [,col], 1, paste, collapse = " ")
df2 <- df1[,c(1,22)]
# head(df2)

# formatting

library(tm)
```

```
## Loading required package: NLP
```

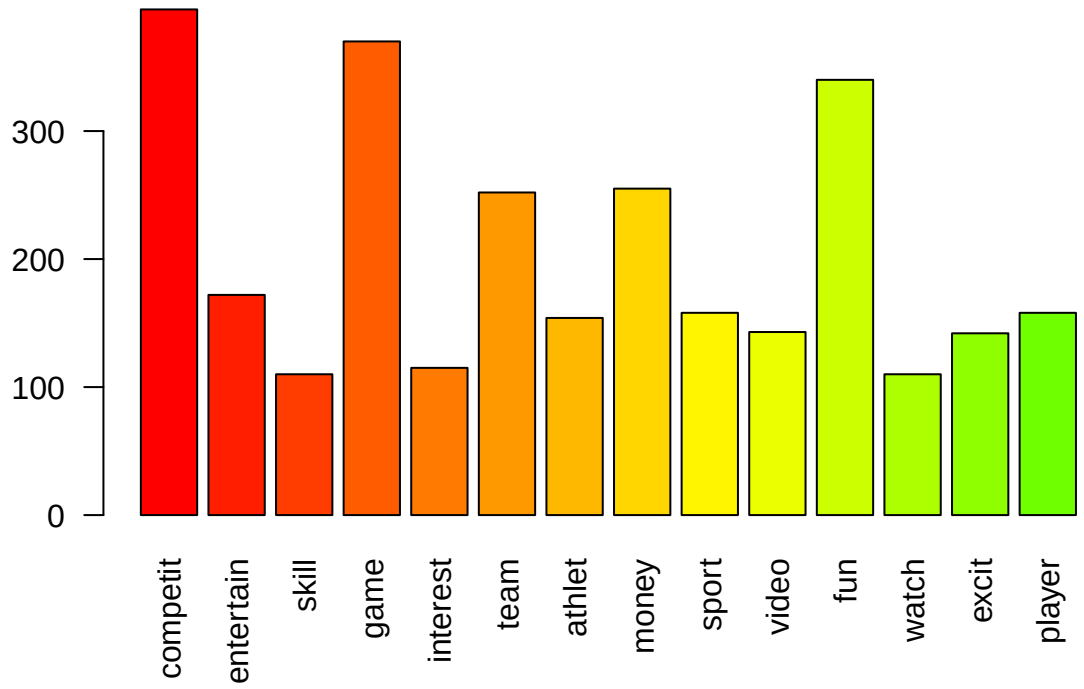
```
library(quanteda)
```

```
## Package version: 2.0.1

## Parallel computing: 2 of 8 threads used.

## See https://quanteda.io for tutorials and examples.

##
## Attaching package: 'quanteda'

## The following objects are masked from 'package:tm':
##
##     as.DocumentTermMatrix, stopwords

## The following objects are masked from 'package:NLP':
##
##     meta, meta<-

## The following object is masked from 'package:utils':
##
##     View
```

```r
corpus <- corpus(df2, docid_field = "ID",
                 text_field = "text",
                 unique_docnames = TRUE)
corpus <- Corpus(VectorSource(corpus))

# clean text
corpus <- tm_map(corpus, tolower)
corpus <- tm_map(corpus, removePunctuation)
corpus <- tm_map(corpus, removeNumbers)
corpus <- tm_map(corpus, stemDocument)
clean <- tm_map(corpus, removeWords, stopwords('english'))
# stopwords('english')
clean <- tm_map(clean, removeWords, c('veri'))
# inspect(corpus[1:3])

# Inspect word frequency
freq <- TermDocumentMatrix(clean)
freqset <- as.data.frame(as.matrix(freq))
write.table(freqset, file='wordfreq_all.csv', sep = ',')
# freq

# findFreqTerms(freq,lowfreq = 5)

# delete words appears less than 0.05 % frequency

word <- removeSparseTerms(freq, 0.995)
word1 <- as.data.frame(as.matrix(word))
rownames(word1)<-make.names(rownames(word))
word <- as.matrix(word1)
# head(word1)
# dim(word1)
write.table(word1, file='wordfreq_trimmed.csv', sep=',')
```

Visualization

```r
# frequency visualization
c <- rowSums(word1)
c <- subset(c, c >= 100)
barplot(c,
        las = 2,
        col = rainbow(50))
```




```
# Word cloud
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

```
v <- sort(rowSums(word1), decreasing = TRUE)
wordcloud(words = names(v),
          freq = v,
          max.words = 60,
          random.order = F,
          min.freq = 20,
          colors = brewer.pal (8, 'Dark2'), scale = c(3, 0.4))
```


championship
stadium
practic
teamwork
watch
athlet
hard
new
activ
real
sport
entertain
nfl
bet
tournament
play
focus
game
win
crowd
player
physic
work
footbal
good
competit
popular
make
like
time
big
organ
enjoy
fun
team
skill
train
happi
comput
fan
fast
friend
high
money
interest
best
lotonlin
healthi
live
video
excit
profession
health
peopl
leagu
can
fit
challeng
esport