ADS project 1

yang xing

09/16/2018

## Part I

This part is all about some data exploration from HappyDB

library(tm)

## Loading required package: NLP

library(tidytext)  
library(tidyverse)

## -- Attaching packages ---------------------------------- tidyverse 1.2.1 --

## √ ggplot2 3.0.0 √ purrr 0.2.5  
## √ tibble 1.4.2 √ dplyr 0.7.5  
## √ tidyr 0.8.1 √ stringr 1.3.1  
## √ readr 1.1.1 √ forcats 0.3.0

## -- Conflicts ------------------------------------- tidyverse\_conflicts() --  
## x ggplot2::annotate() masks NLP::annotate()  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(DT)  
library(stringr)  
library(scales)

##   
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':  
##   
## discard

## The following object is masked from 'package:readr':  
##   
## col\_factor

library(wordcloud)

## Loading required package: RColorBrewer

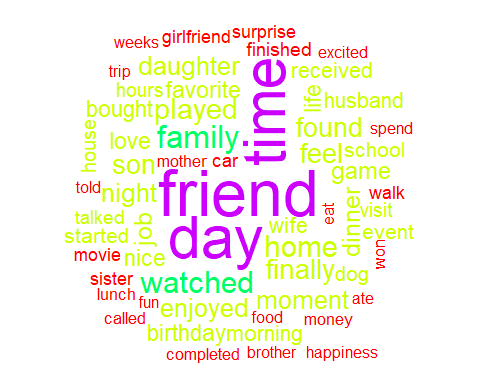
library(topicmodels)

urlfile<-'https://raw.githubusercontent.com/rit-public/HappyDB/master/happydb/data/demographic.csv'  
demo\_data <- read.csv(urlfile,as.is = TRUE)  
  
hm\_data <- read.csv("processed\_moments.csv",as.is = TRUE)  
hm\_data <- hm\_data %>%  
 inner\_join(demo\_data, by = "wid")

wordcloud(words = hm\_data$text, min.freq = 1500, random.order = FALSE, colors = rainbow(5))

## Warning in tm\_map.SimpleCorpus(corpus, tm::removePunctuation):  
## transformation drops documents

## Warning in tm\_map.SimpleCorpus(corpus, function(x) tm::removeWords(x,  
## tm::stopwords())): transformation drops documents



poswords <- scan("positive-words.txt", what = 'character', comment.char = ";")  
negwords <- scan("negative-words.txt", what = 'character', comment.char = ";")  
sum(!is.na((match(x = hm\_data$text, poswords))))

## [1] 94

sum(!is.na((match(x = hm\_data$text, negwords))))

## [1] 28

ratio <- sum(!is.na((match(x = hm\_data$text, poswords)))) / sum(!is.na((match(x = hm\_data$text, negwords))))  
ratio

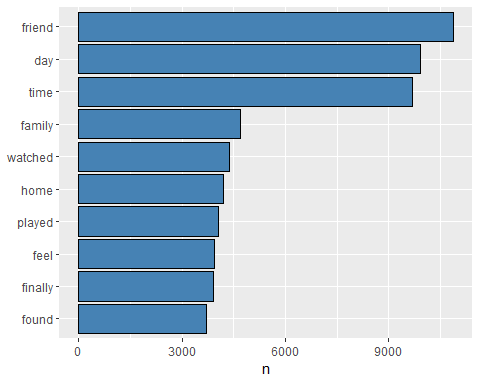
## [1] 3.357143

As we could expect, this is a positive article.

hm\_data\_tokenize <- hm\_data %>%  
 unnest\_tokens(word,text) %>%  
 count(word,sort = TRUE)  
head(hm\_data\_tokenize,10)

## # A tibble: 10 x 2  
## word n  
## <chr> <int>  
## 1 friend 10892  
## 2 day 9930  
## 3 time 9692  
## 4 family 4692  
## 5 watched 4385  
## 6 home 4211  
## 7 played 4058  
## 8 feel 3946  
## 9 finally 3922  
## 10 found 3720

hm\_data\_ggplot\_tokenize <- hm\_data\_tokenize %>%  
 head(10) %>%  
 mutate(word = reorder(word,n))%>%  
 ggplot(aes(word,n))+  
 geom\_col(color="black", fill="steelblue")+  
 xlab(NULL)+  
 coord\_flip()  
hm\_data\_ggplot\_tokenize



hm\_data\_bigram <- hm\_data %>%  
 unnest\_tokens(bigram, text, token = "ngrams", n = 2) %>%  
 count(bigram,sort = TRUE)  
head(hm\_data\_bigram,10)

## # A tibble: 10 x 2  
## bigram n  
## <chr> <int>  
## 1 <NA> 3154  
## 2 spend time 825  
## 3 video game 736  
## 4 moment life 590  
## 5 ice cream 531  
## 6 watched movie 503  
## 7 birthday party 468  
## 8 mother day 445  
## 9 spent time 329  
## 10 friend mine 319

hm\_data\_trigram <- hm\_data %>%  
 unnest\_tokens(trigram, text, token = "ngrams", n = 3) %>%  
 count(trigram,sort = TRUE)  
head(hm\_data\_trigram,10)

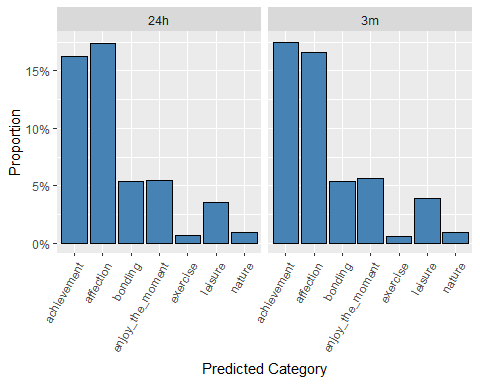
## # A tibble: 10 x 2  
## trigram n  
## <chr> <int>  
## 1 <NA> 13827  
## 2 played video game 289  
## 3 offsite colleagues fun 110  
## 4 colleagues fun stimulating 109  
## 5 fun stimulating discussions 109  
## 6 friend birthday party 108  
## 7 smart phone surprise 97  
## 8 spend time family 90  
## 9 person person time 72  
## 10 judgements person wellbeing 68

As we could see, “played” or “watched” some activities with our “friend” and “family” plays an important role in making people happy. Have some “ice cream” to eat is useful for good mood as well.

## Part II

This part is talking about the details changing for seven predicted category if we seperate the data into different groups. #Predicted category VS Reflection Period

ggplot(hm\_data, aes(x=as.factor(hm\_data$predicted\_category))) +   
 facet\_wrap(~hm\_data$reflection\_period) +   
 geom\_bar(aes(y=..count../sum(..count..)),color = "black",fill= "steelblue") +  
 scale\_y\_continuous(labels=percent\_format())+  
 xlab("Predicted Category") +   
 ylab("Proportion") +  
 theme(axis.text.x = element\_text(angle = 60, hjust = 1))



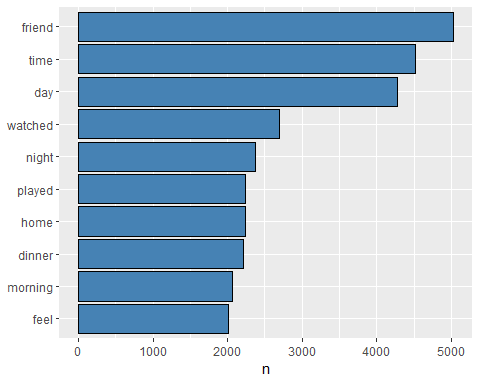
hm\_data\_24h <- hm\_data[hm\_data$reflection\_period == "24h", ]  
hm\_data\_3m <- hm\_data[hm\_data$reflection\_period == "3m", ]  
hm\_data\_24h\_tokenize <- hm\_data\_24h %>%  
 unnest\_tokens(word,text) %>%  
 count(word,sort = TRUE)  
head(hm\_data\_24h\_tokenize,10)

## # A tibble: 10 x 2  
## word n  
## <chr> <int>  
## 1 friend 5035  
## 2 time 4524  
## 3 day 4284  
## 4 watched 2694  
## 5 night 2374  
## 6 played 2236  
## 7 home 2234  
## 8 dinner 2208  
## 9 morning 2065  
## 10 feel 2010

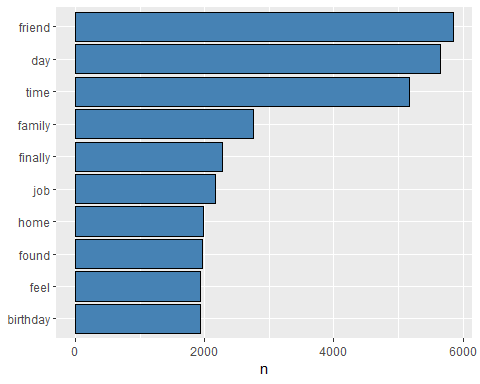
hm\_data\_3m\_tokenize <- hm\_data\_3m %>%  
 unnest\_tokens(word,text) %>%  
 count(word,sort = TRUE)  
head(hm\_data\_3m\_tokenize,10)

## # A tibble: 10 x 2  
## word n  
## <chr> <int>  
## 1 friend 5857  
## 2 day 5646  
## 3 time 5168  
## 4 family 2760  
## 5 finally 2270  
## 6 job 2162  
## 7 home 1977  
## 8 found 1969  
## 9 feel 1936  
## 10 birthday 1934

hm\_data\_24h\_ggplot <- hm\_data\_24h %>%  
 unnest\_tokens(word,text) %>%  
 count(word,sort = TRUE) %>%  
 head(10) %>%  
 #filter(n>2000) %>%  
 mutate(word = reorder(word,n))%>%  
 ggplot(aes(word,n))+  
 geom\_col(color="black", fill="steelblue")+  
 xlab(NULL)+  
 coord\_flip()  
hm\_data\_24h\_ggplot



hm\_data\_3m\_tokenize <- hm\_data\_3m %>%  
 unnest\_tokens(word,text) %>%  
 count(word,sort = TRUE) %>%  
 head(10) %>%  
 #filter(n>2000) %>%  
 mutate(word = reorder(word,n))%>%  
 ggplot(aes(word,n))+  
 geom\_col(color="black", fill="steelblue")+  
 xlab(NULL)+  
 coord\_flip()  
hm\_data\_3m\_tokenize



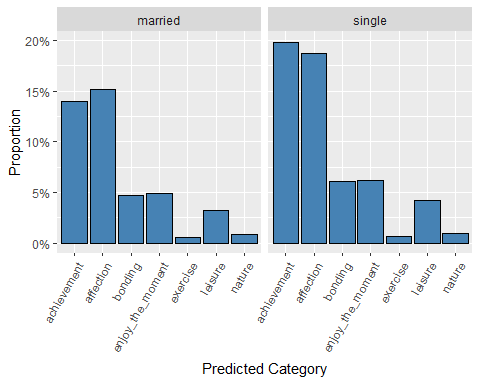
table(hm\_data\_24h$predicted\_category) - table(hm\_data\_3m$predicted\_category)

##   
## achievement affection bonding enjoy\_the\_moment   
## -3131 -1464 -286 2149   
## exercise leisure nature   
## 320 1207 305

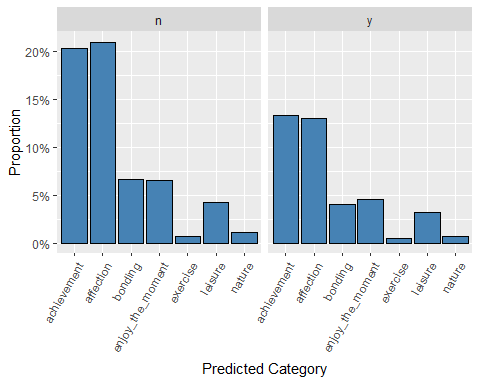
From the above infomation in different reflection period, something about exercise and leisure appears more frequently in 24 hours than in 3 months. However the achievent and affection appear more often in 3m relection period. It indicates that the longer the reflection, the more people tend to recall the big events rather than food and entertainment

# Predicted category VS Marital status

hm\_data\_marital <- hm\_data[hm\_data$marital == "married" |hm\_data$marital == "single",]  
ggplot(hm\_data\_marital, aes(x=as.factor(hm\_data\_marital$predicted\_category))) +   
 facet\_wrap(~hm\_data\_marital$marital) +   
 geom\_bar(aes(y=..count../sum(..count..)),color = "black",fill= "steel blue") +  
 scale\_y\_continuous(labels=percent\_format())+  
 xlab("Predicted Category") +   
 ylab("Proportion") +  
 theme(axis.text.x = element\_text(angle = 60, hjust = 1))

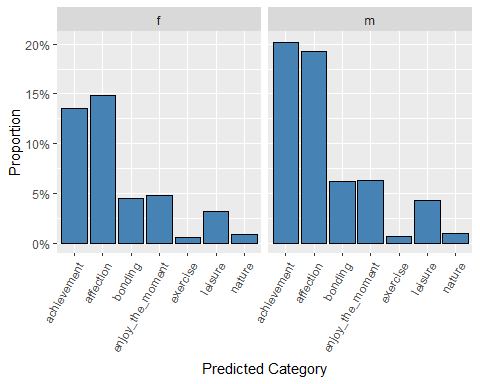
 As we could see, the sigle person more like to talk about the achievement they made in the past. But the married person prefer to take about their affection, which is activity with family members and loved ones. #Predicted category VS Parenthood

hm\_data\_parenthood <- hm\_data[hm\_data$parenthood == "n" |hm\_data$parenthood == "y",]  
ggplot(hm\_data\_parenthood, aes(x=as.factor(hm\_data\_parenthood$predicted\_category))) +   
 facet\_wrap(~hm\_data\_parenthood$parenthood) +   
 geom\_bar(aes(y=..count../sum(..count..)),color = "black",fill= "steelblue") +  
 scale\_y\_continuous(labels=percent\_format())+  
 xlab("Predicted Category") +   
 ylab("Proportion") +  
 theme(axis.text.x = element\_text(angle = 60, hjust = 1))



As we could see, the sigle person more like to talk about the achievement they made in the past. But the married person prefer to take about their affection, which is activity with family members and loved ones. #Predicted category VS Gender

hm\_data\_gender <- hm\_data[hm\_data$gender == "f" |hm\_data$gender == "m",]  
ggplot(hm\_data\_gender, aes(x=as.factor(hm\_data\_gender$predicted\_category))) +   
 facet\_wrap(~hm\_data\_gender$gender) +   
 geom\_bar(aes(y=..count../sum(..count..)),color = "black",fill= "steelblue") +  
 scale\_y\_continuous(labels=percent\_format())+  
 xlab("Predicted Category") +   
 ylab("Proportion") +  
 theme(axis.text.x = element\_text(angle = 60, hjust = 1))

 As we could see, the male more like to talk about the achievement they made in the past. But the female prefer to take about their affection, which is activity with family members and loved ones.

## Conclusion

Basically, this project contains two parts. The first part is about the general picture of the data. The wordcloud, sentimental analysis and frequency words, etc. I found that is a good idea to spend sometime doing activities with friends and familis is good for our mood. Second part is about some detail analysis about predicted category respect to different group. It seems like people with 24 hours relection period more like to recall their “affection” and “leisure” compare to people with 3 month reflection period. Moreover, male people and parents without kids will foucus on their “achievement” and “affection” compare to others.