HappyDB_Huiyu_Zhang

Topic: How people from different age groups differ in describing their happy moments

First, load necessary packages

```r
library(tm)
library(tidyverse)
library(tidytext)
library(ngram)
library(plyr)
library(dplyr)
library(data.table)
library(tidytext)
library(ggplot2)
library(ggcorrplot)
library(base)
library(DataCombine)
library(colorspace)
library(magrittr)
library(multipanelfigure)
```

Dataset Loading

```r
hm<-read_csv("../output/processed_moments.csv")
url<-'https://raw.githubusercontent.com/rit-public/HappyDB/master/happydb/data/demographic.csv'
demo<-read_csv(url)
```

Data Combing and data cleaning

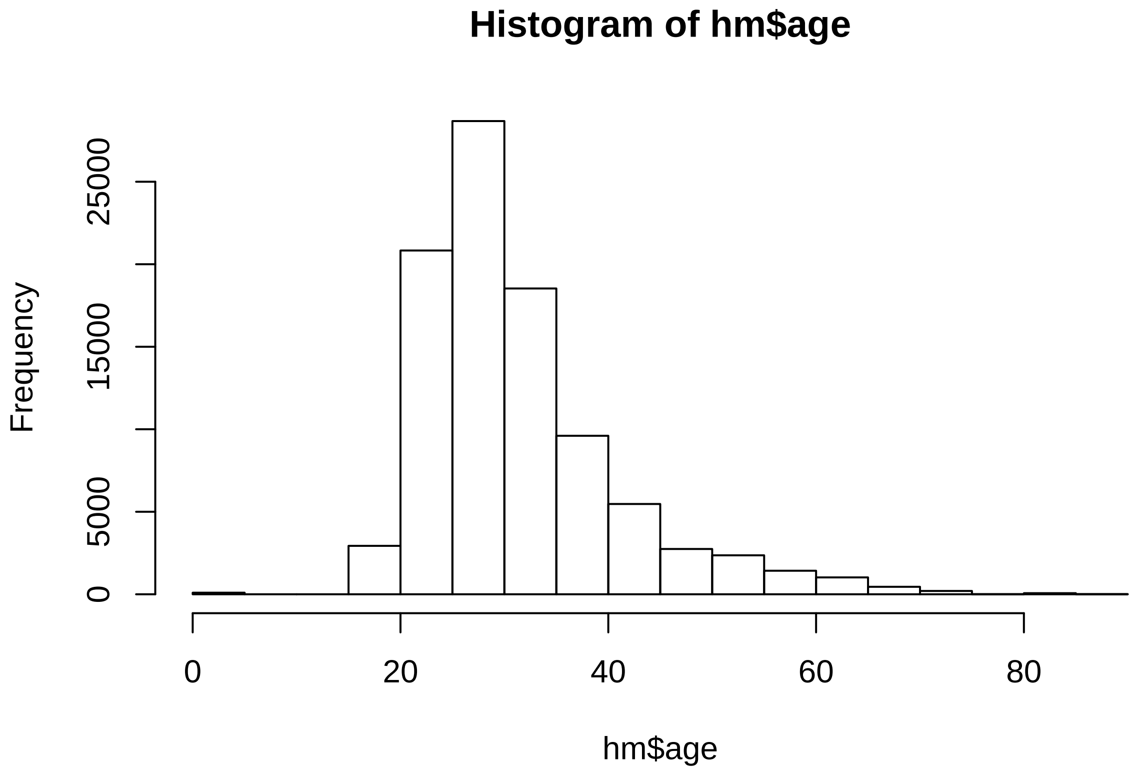
```r
# Combine dataset hm and dataset demo by their common wid
hm<- inner_join(hm,demo,by="wid")
hm<- select(hm,wid,original_hm,gender,marital,parenthood,reflection_period,age,country,predicted_catego
# Transfer the age from a string to a numeric number
hm$age<-as.numeric(hm$age)
# Add a column calculating the number of words
hm<- mutate(hm,count=sapply(hm$original_hm, wordcount))
# Filer out dirty data
hm<- filter(hm, gender %in% c("m","f"))
hm<- filter(hm, marital %in% c("single","married"))
hm<- filter(hm, parenthood %in% c("n","y"))
hm<- filter(hm, reflection_period %in% c("24h","3m"))
```

Explore basic information about different age groups

```
table(hm$age)
```

```
##
##    2    3   17   18   19   20   21   22   23   24   25   26   27   28   29
##   15   81    6  463 1002 1464 2509 3452 4488 4434 5950 6022 5493 5651 6065
##   30   31   32   33   34   35   36   37   38   39   40   41   42   43   44
## 5446 4130 4447 3158 3642 3156 2559 2140 1956 1399 1552 1433 1080 1052 1053
##   45   46   47   48   49   50   51   52   53   54   55   56   57   58   59
##  852  522  582  560  646  434  443  558  438  563  359  355  356  245  246
##   60   61   62   63   64   65   66   67   68   69   70   71   72   73   74
##  225  381  264  113  117  147  141   51   81   96   84   30   63   18   84
##   75   78   83   84   88   95  227  233
##    6    3   63    3    6    3    9   51
```

```
# Since there is only few people older than 90, which is not very useful for analysis. I am gonna filte
hm<- filter(hm,age<90)
# Take a look on distribution of ages
hist(hm$age)
```



```
# Break them into 9 age groups
agebreaks<- c(0,10,20,30,40,50,60,70,80,90)
agelabels<- c("0-9","10-19","20-29","30-39","40-49","50-59","60-69","70-79","80-89")
setDT(hm)[,agegroups:=cut(age,breaks=agebreaks,right=FALSE,labels=agelabels)]
# Take a look on the distribution of agegroups
table(hm$agegroups)
```

```
##
##   0-9 10-19 20-29 30-39 40-49 50-59 60-69 70-79 80-89
##    96  1471 45528 32033  9332  3997  1616   288    72
```

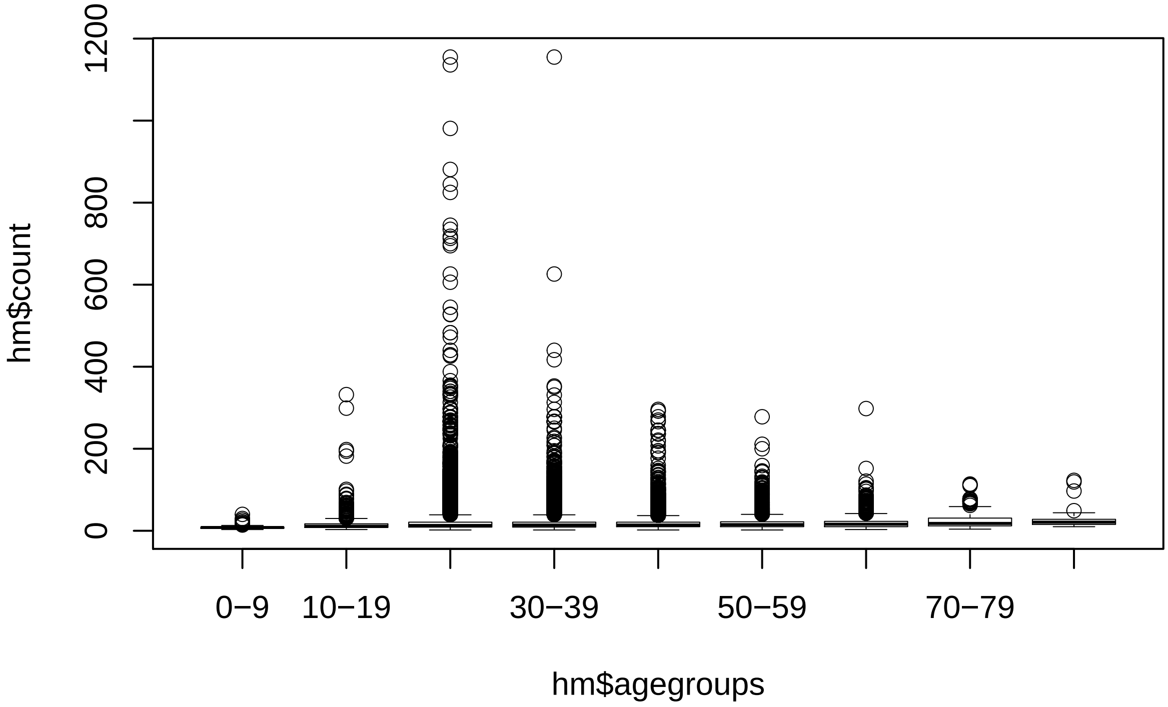
```
# Explore the basic relationship between agegroups and other variables
table(hm$gender,hm$agegroups)
```

```
##
##       0-9 10-19 20-29 30-39 40-49 50-59 60-69 70-79 80-89
##   f      6   408 17207 12584  4533  2848   926   132     9
##   m     90  1063 28321 19449  4799  1149   690   156    63
```

```
table(hm$country,hm$agegroups)
```

```
##
##         0-9 10-19 20-29 30-39 40-49 50-59 60-69 70-79 80-89
##   AFG     0     0     5     0     0     0     0     0     0
##   ALB     6     0    42     0     0     0     0     0     0
##   ARE     0     0    30     3     0     0     0     0     0
##   ARG     0     0     3     3     0     0     0     0     0
##   ARM     0     0     0    15     0     0     0     0     0
##   ASM     0     0    13     0     0     0     0     0     0
##   AUS     0    18    33    48    12     6     0     0     0
##   AUT     0     0    15     2     0     0     0     0     0
##   BEL     0     0     0     6     3     0     0     0     0
##   BGD     0     0    66     3     0     0     0     0     0
##   BGR     0     0    63     0     4     0     0     0     0
##   BHS     0     0     0     3     0     0     0     0     0
##   BRA     0     0   105    18     0     0     0     0     0
##   BRB     0     0     3     3     0     0     0     0     0
##   CAN     0    36   279   162    18     9    27     0     0
##   CHL     0     0     3     0     3     0     0     0     0
##   COL     0     0     6    23     3     0     0     0     0
##   CRI     0     0     0     0     3     0     0     0     0
##   CYP     0     0     0     3     0     0     0     0     0
##   CZE     0     0     6     0     0     0     0     0     0
##   DEU     0     0    39    45     0     0     0     0     0
##   DNK     0     0    51     0     0     0     0     0     0
##   DOM     0     0     3    48     0     0     0     0     0
##   DZA     0     0     3     9     0     0     0     0     0
##   ECU     0     0     0     3     0     0     0     0     0
##   EGY     0     0    57     0     0     0     0     0     0
##   ESP     0     0     6     2     3     3     0     0     0
##   EST     0     0     6     0     0     0     0     0     0
##   ETH     0     0     0     3     0     0     0     0     0
##   FIN     0     6    12     0     3     0     0     0     0
##   FRA     0    15     9    12     9     6     0     0     0
##   GBR     3    15   196    48    72    18     3     0     0
##   GHA     0     0     0     0     3     0     0     0     0
##   GMB     0     0     0     6     0     0     0     0     0
##   GRC     0     0    27    12     3     0     0     0     0
##   GTM     0     0     6     0     0     0     0     0     0
##   HKG     0     0     3     0     0     0     0     0     0
##   HRV     0     0     3     3     0     0     0     0     0
##   IDN     0     0    18    12     3     0     0     0     0
##   IND    75   142 10286  4831   993   191   105     3     0
##   IRL     0     0     3    24     3     0     0     0     0
##   IRQ      0     0     0     3     0     0     0     0     0
##   ISL      0     0     3     6     0     0     0     0     0
##   ISR      0     0     0     0     3     0     0     0     0
##   ITA      0     9     9     6     9     3     0     0     0
##   JAM      0    36    15     3     0     6     0     0     0
##   JPN      0     0     0    15     0     0     0     0     0
##   KAZ      0     3     0     0     0     0     0     0     0
##   KEN      0     0    33     0     0     0     0     0     0
##   KNA      0     0     9     0     0     0     0     0     0
##   KOR      0     0     0     0     6     0     0     0     0
##   KWT      0     0    18     0     0     0     0     0     0
##   LKA      0     0    12     0     0     0     0     0     0
##   LTU      0     0    42     0     0     0     0     0     0
##   LVA      0     0     0     3     0     0     0     0     0
##   MAR      0     0     6     0     0     0     0     0     0
##   MDA      0     0     0    36     0     0     0     0     0
##   MEX      0     3    69    45     0     3     3     0     0
##   MKD      0     0     0   102     0     0     0     0     0
##   MLT      0     0     3     6     0     0     0     0     0
##   MUS      0     0     3     0     0     0     0     0     0
##   MYS      0     3     3     3     6     0     0     0     0
##   NGA      0     0    27    48     6     0     0     0     0
##   NIC      0     0    12     0     0     3     0     0     0
##   NLD      0     3     0     0    12     0     0     0     0
##   NOR      0     0     0     3     0     0     0     0     0
##   NPL      0     0     0     6     0     0     0     0     0
##   NZL      0     0    24     6     6     0     0     0     0
##   PAK      0     0     9     3    27     0     0     0     0
##   PER      0     0    24    10     0     0     0     0     0
##   PHL      0    27   213    24    12     0     0     0     0
##   POL      0     3     6     3     3     0     0     0     0
##   PRI      0     0    27     3     0     0     0     0     0
##   PRT      0     6     3    72     3     0     0     0     0
##   ROU      0     0    43     3     0     0     0     0     0
##   RUS      0     0     0    30     0     0     0     0     0
##   SAU      0     0     3     0     0     0     0     0     0
##   SGP      0     3    12     9     0     0     0     0     0
##   SLV      0     0     3     0     0     0     0     0     0
##   SRB      0     0    81    12     3     0     0     0     0
##   SUR      0     0     0     3     0     0     0     0     0
##   SVN      0     0     6     0     0     0     0     0     0
##   SWE      0     0     0    27     0     0     0     0     0
##   TCA      0     0     0     3     0     0     3     0     0
##   THA      0     0     0     0    84     0     0     0     0
##   TTO      0     0     3    24     3     0     0     0     0
##   TUN      0     0     3     0     0     0     0     0     0
##   TUR      0    12    24     6     9     0     0     0     0
##   TWN      0     0     9     0     0     0     0     0     0
##   UGA      0     0    15     0     0     0     0     0     0
##   UKR      0     0     0     3     0     0     0     0     0
##   UMI      0     0    12     0     3     0     0     0     0
##   URY      0     0     0     0     0     0    42     0     0
##   USA     12  1128 32823 25941  7945  3710  1427   279    72
##   VEN      0     0   369   114    36    24     3     0     0
##   VIR     0     0     3     0     0     0     0     0     0
##   VNM     0     0    89    36     0     0     0     0     0
##   ZAF     0     0    18     0     0     3     0     0     0
##   ZMB     0     0     3     0     0     0     0     0     0
```

```
# Distribution of length of words in different agegroups.
plot(hm$count~hm$agegroups,type="p",lwd=0.5)
```



It is very interesting to see how does people from different agegroups differ in length of happy momemnt description, and to see the distribution of their genders and countries

Create bag of words

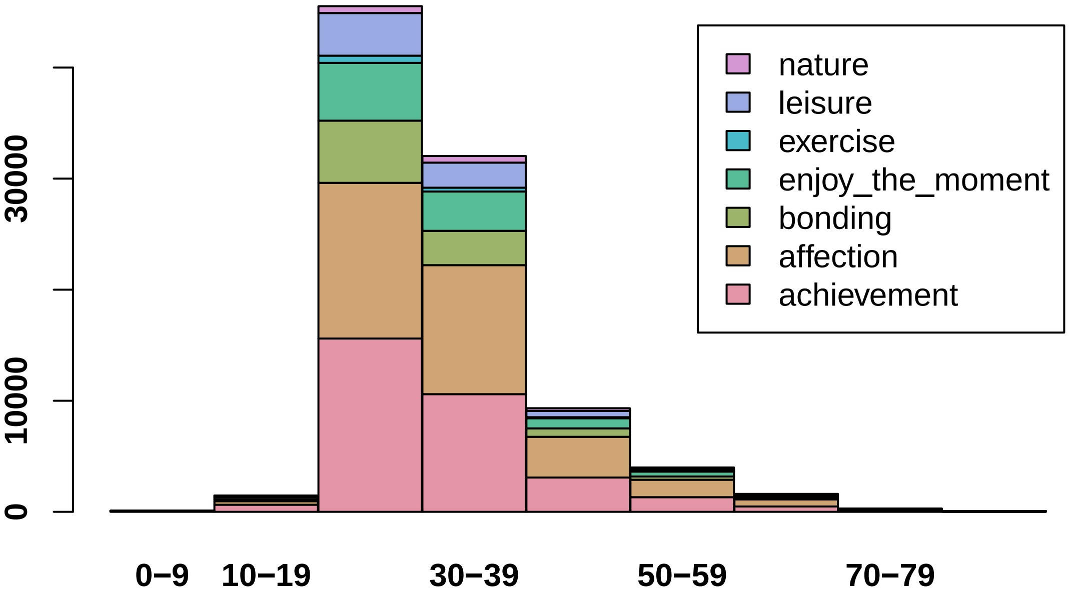
```
bow<- unnest_tokens(hm,word,text)
word_count<- dplyr::count(bow,word,sort=TRUE)
```

Relationship between agegroups and predicted_category

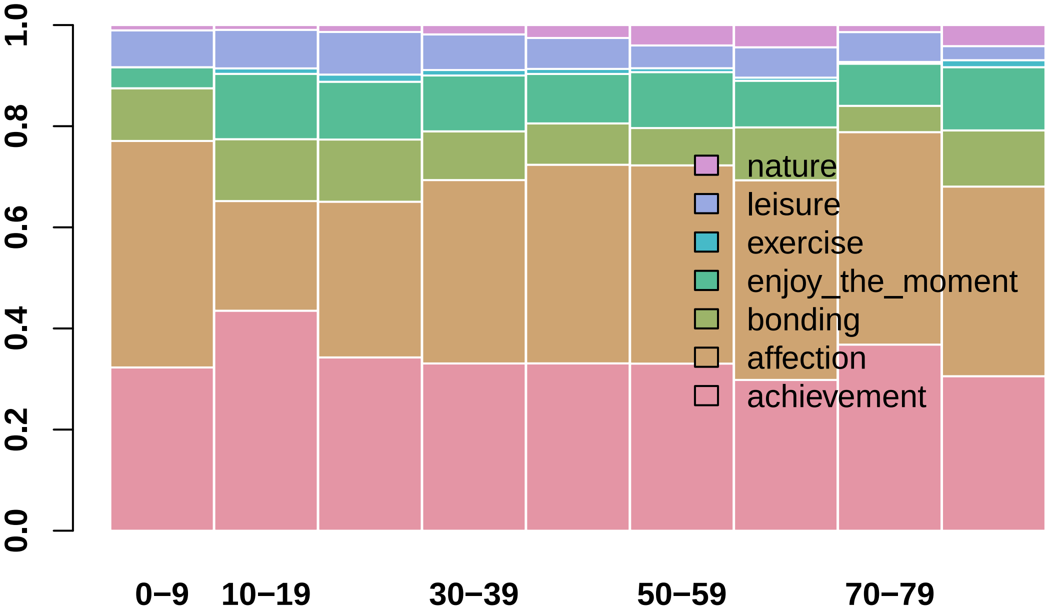
Let's find out how people from different age groups differ in the predicted category of their happy moments

```
# Count predicted categories for every agegroup
category_agegroups<- ddply(hm,.(hm$predicted_category,hm$agegroups),nrow)
names(category_agegroups)<- c("predicted_category","agegroups","counts")
# Insert a new row where 0-9 age groups didn't mention exercise at all
category_agegroups<- InsertRow(category_agegroups,c("exercise","0-9",0),37)
category_agegroups$counts<- as.numeric(category_agegroups$counts)
data<- matrix(category_agegroups$counts,nrow=7,byrow = T)
rownames(data)<- c("achievement","affection","bonding","enjoy_the_moment","exercise","leisure","nature"
colnames(data)<- c("0-9","10-19","20-29","30-39","40-49","50-59","60-69","70-79","80-89")
#count(hm,vars=c("predicted_category","agegroups"))
```

```
theme_set(theme_classic())
barplot(data,col=rainbow_hcl(7),legend=rownames(data),space = 0.005,font.axis=2)
```



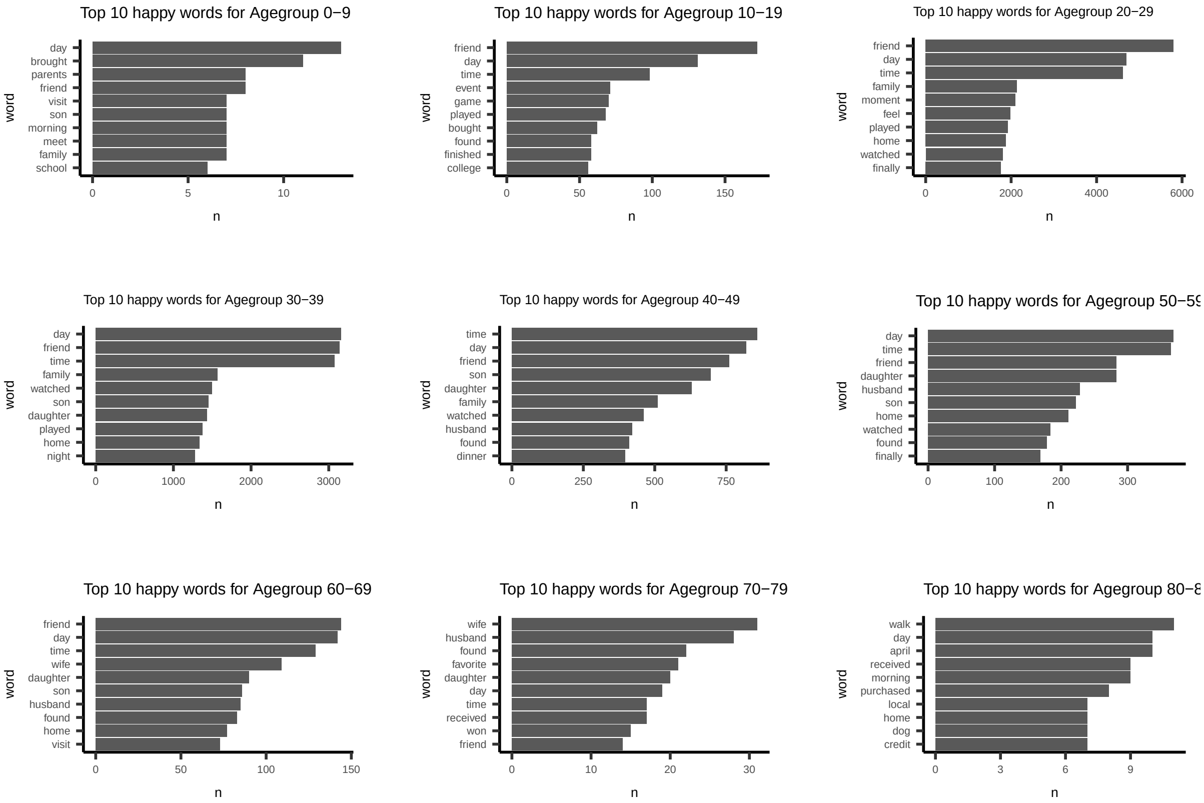
```
data_percentage<- apply(data,2,function(x){x/sum(x)})
barplot(data_percentage,col=rainbow_hcl(7),border="white",legend.text=rownames(data),space = 0.005,font
```



Looking at the first plot, we are able to see what kind of happy moments are people from different agegroups mainly taking about. But since there's large difference of population between groups, it is hard to tell how the percentage of category differs between agegroups. That's the reason why I created the second plot. According to the second plot, it's much more easier to find out the category percentage difference between agegroups.

Find top 10 popular happy words in every agegroup

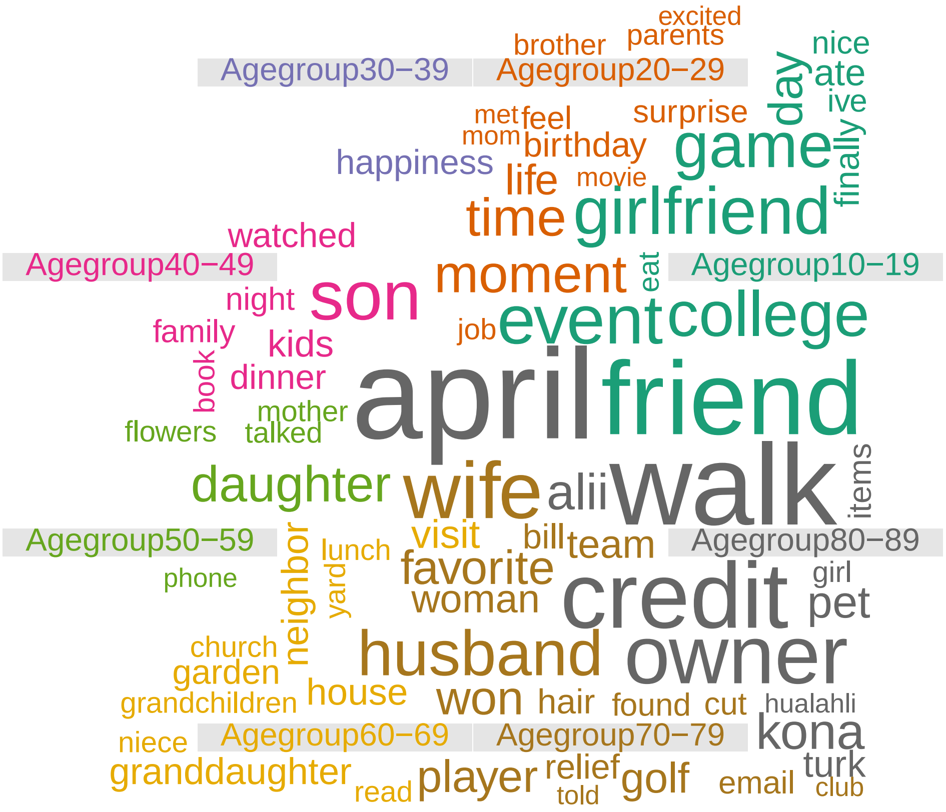
```
bow<- unnest_tokens(hm,word,text)
bow_0<- bow[bow$agegroups=="0-9",]
bow_10<-bow[bow$agegroups=="10-19",]
bow_20<-bow[bow$agegroups=="20-29",]
bow_30<-bow[bow$agegroups=="30-39",]
bow_40<-bow[bow$agegroups=="40-49",]
bow_50<-bow[bow$agegroups=="50-59",]
bow_60<-bow[bow$agegroups=="60-69",]
bow_70<-bow[bow$agegroups=="70-79",]
bow_80<-bow[bow$agegroups=="80-89",]
word_count<- filter(dplyr::count(bow,word,sort=TRUE),n!=1)
word_count0<- filter(dplyr::count(bow_0,word,sort=TRUE),n!=1)[1:10,]
word_count10<- filter(dplyr::count(bow_10,word,sort=TRUE),n!=1)[1:10,]
word_count20<- filter(dplyr::count(bow_20,word,sort=TRUE),n!=1)[1:10,]
word_count30<- filter(dplyr::count(bow_30,word,sort=TRUE),n!=1)[1:10,]
word_count40<- filter(dplyr::count(bow_40,word,sort=TRUE),n!=1)[1:10,]
word_count50<- filter(dplyr::count(bow_50,word,sort=TRUE),n!=1)[1:10,]
word_count60<- filter(dplyr::count(bow_60,word,sort=TRUE),n!=1)[1:10,]
word_count70<- filter(dplyr::count(bow_70,word,sort=TRUE),n!=1)[1:10,]
word_count80<- filter(dplyr::count(bow_80,word,sort=TRUE),n!=1)[1:10,]
p1<- word_count0 %>%
  mutate(word=fct_reorder(word,n)) %>%
  ggplot(aes(x=word,y=n))+geom_bar(stat = "identity")+coord_flip()+labs(title="Top 10 happy words for Ag
  theme(text = element_text(size=5))
p2<- word_count10 %>%
  mutate(word=fct_reorder(word,n)) %>%
  ggplot(aes(x=word,y=n))+geom_bar(stat = "identity")+coord_flip()+labs(title="Top 10 happy words for Ag
  theme(text = element_text(size=5))
p3<- word_count20 %>%
  mutate(word=fct_reorder(word,n)) %>%
  ggplot(aes(x=word,y=n))+geom_bar(stat = "identity")+coord_flip()+labs(subtitle="Top 10 happy words fo
  theme(text = element_text(size=5))
p4<- word_count30 %>%
  mutate(word=fct_reorder(word,n)) %>%
  ggplot(aes(x=word,y=n))+geom_bar(stat = "identity")+coord_flip()+labs(subtitle="Top 10 happy words fo
  theme(text = element_text(size=5))
p5<- word_count40 %>%
  mutate(word=fct_reorder(word,n)) %>%
  ggplot(aes(x=word,y=n))+geom_bar(stat = "identity")+coord_flip()+labs(subtitle="Top 10 happy words fo
  theme(text = element_text(size=5))
p6<- word_count50 %>%
  mutate(word=fct_reorder(word,n)) %>%
  ggplot(aes(x=word,y=n))+geom_bar(stat = "identity")+coord_flip()+labs(title="Top 10 happy words for Ag
  theme(text = element_text(size=5))
p7<- word_count60 %>%
  mutate(word=fct_reorder(word,n)) %>%
  ggplot(aes(x=word,y=n))+geom_bar(stat = "identity")+coord_flip()+labs(title="Top 10 happy words for Ag
  theme(text = element_text(size=5))
p8<- word_count70 %>%
  mutate(word=fct_reorder(word,n)) %>%
  ggplot(aes(x=word,y=n))+geom_bar(stat = "identity")+coord_flip()+labs(title="Top 10 happy words for Ag
  theme(text = element_text(size=5))
p9<- word_count80 %>%
  mutate(word=fct_reorder(word,n)) %>%
  ggplot(aes(x=word,y=n))+geom_bar(stat = "identity")+coord_flip()+labs(title="Top 10 happy words for Ag
  theme(text = element_text(size=5))
figure<- multi_panel_figure(columns = 3,rows = 3,panel_label_type = "none")
figure %>%
  fill_panel(p1,column = 1,row = 1) %>%
  fill_panel(p2,column = 2,row = 1) %>%
  fill_panel(p3,column = 3,row = 1) %>%
  fill_panel(p4,column = 1,row = 2) %>%
  fill_panel(p5,column = 2,row = 2) %>%
  fill_panel(p6,column = 3,row = 2) %>%
  fill_panel(p7,column = 1,row = 3) %>%
  fill_panel(p8,column = 2,row = 3) %>%
  fill_panel(p9,column = 3,row = 3)
```



Here are the top 10 happy words for differnt agegroups.

Create Comparison Word Cloud

```
#since the comparison word cloud only allows 8 groups for campare, I deleted the Agegroup 0-9
corpus<- c(paste(bow[bow$agegroups=="10-19",]$word,collapse=" "),paste(bow[bow$agegroups=="20-29",]$word
co<- Corpus(VectorSource(corpus))
tdm<- TermDocumentMatrix(co)
m<- as.matrix(tdm)
colnames(m)<- c("Agegroup10-19","Agegroup20-29","Agegroup30-39","Agegroup40-49","Agegroup50-59","Agegrou
wordcloud::comparison.cloud(m,title.size = 1,match.colors = T)
```



This comparison word cloud shows the most common happy words among various agegroups. We can clearly see that some of agegroups share the same happy words but also differs in other happy words.