5243HW1

Peiu Zhang

9/19/2018

My project topic is *Happy Moment for Different People: what makes people happy, and how to be happy.* To find out what makes people happy, I count the word frequency so that I can find the happy thing shows most often. To find out how to be happy. I use two words relationships to describe. Further, the word correlation shows how likely two words written together. That was another way to explain how to be happy. For different people, I split people into various different country, gender, marriage statu, parenthood. Then explained the difference of happy moments from different kind of people

In general, what makes people happy

1.1 Load the data I need

```
urlfile<-'Fall2018-Proj1-PuleiPai/data/cleaned_hm.csv'
hm_data <- read_csv(urlfile)
```

```
## Parsed with column specification:
## cols(
##   hmid = col_integer(),
##   wid = col_integer(),
##   reflection_period = col_character(),
##   original_hm = col_character(),
##   cleaned_hm = col_character(),
##   modified = col_character(),
##   num_sentence = col_integer(),
##   ground_truth_category = col_character(),
##   predicted_category = col_character()
## )
```

```
urlfile2<-'Fall2018-Proj1-PuleiPai/data/demographic.csv'
demo_data <- read.csv(urlfile2)
```

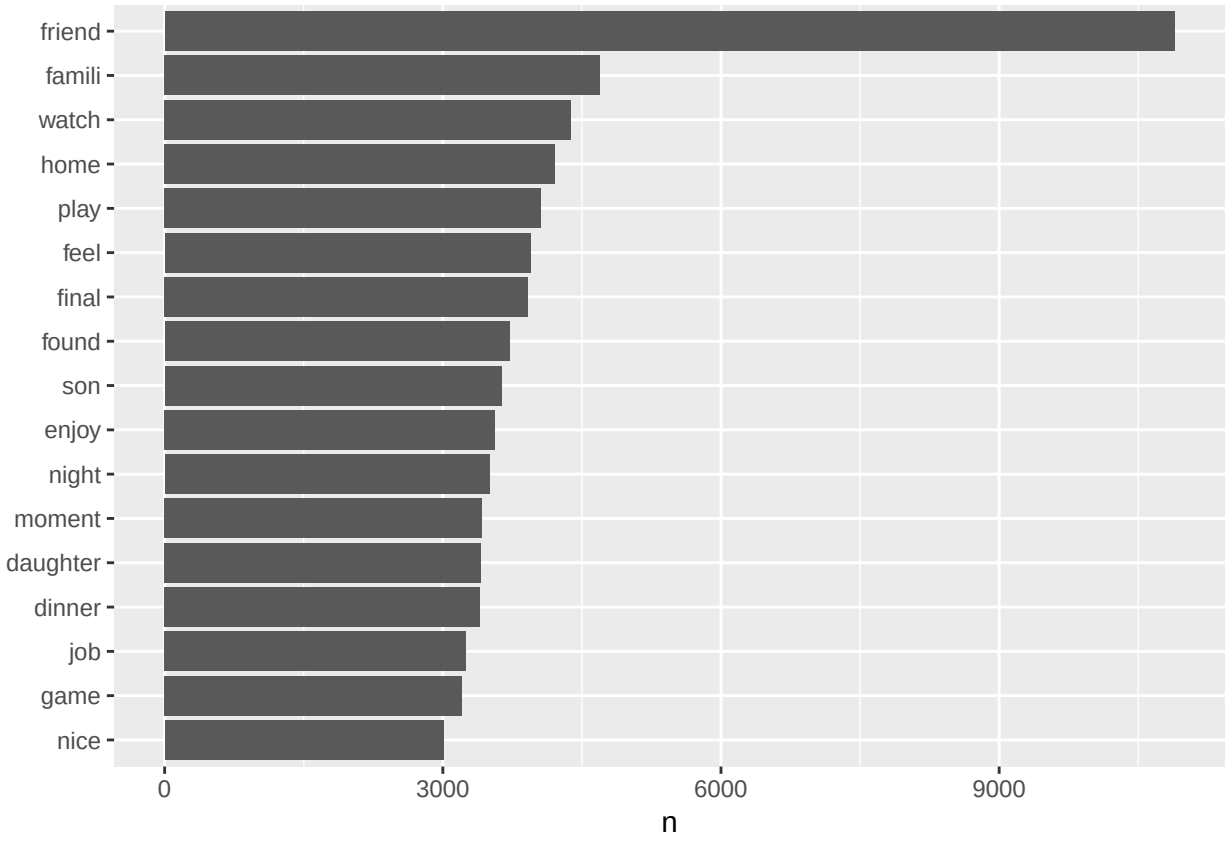
1.2 Clean the text and then convert it to one-token-per-document-per-row

```
corpus <- VCorpus(VectorSource(hm_data$cleaned_hm))%>%
  tm_map(content_transformer(tolower))%>%
  tm_map(removePunctuation)%>%
  tm_map(removeNumbers)%>%
  tm_map(removeWords, character(0))%>%
  tm_map(stripWhitespace)
# munate every word to it's stem form
stemmed <- tm_map(corpus, stemDocument) %>%
  tidy() %>%
  select(text)
# make original corpus one-token-per-document-per-row
dict <- tidy(corpus) %>%
```

```
  select(text) %>%
  unnest_tokens(dictionary, text)
# clean the stop word out
data("stop_words")
word <- c("happy","ago","yesterday","lot","today","months","month",
                  "happier","happiest","last","week","past","time","day")
stop_words <- stop_words %>%
  bind_rows(mutate(tibble(word), lexicon = "updated"))
# bind stem word column to dictionary
completed <- stemmed %>%
  mutate(id = row_number()) %>%
  unnest_tokens(stems, text) %>%
  bind_cols(dict) %>%
  anti_join(stop_words, by = c("dictionary" = "word"))
```

1.3 Word Frequency

```
completed %>%
  count(stems, sort = TRUE) %>%
  filter(n > 3000) %>%
  mutate(stems = reorder(stems, n)) %>%
  ggplot(aes(stems, n)) +
  geom_col() +
  xlab(NULL) +
  coord_flip()
```



```
library(wordcloud)
```

```
## Warning: package 'wordcloud' was built under R version 3.4.4
```

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

```
set.seed(0)
completed %>%
  count(stems, sort = TRUE) %>%
  filter(n > 300) %>%
  mutate(stems = reorder(stems, n)) %>%
  with(wordcloud(stems,n,colors = brewer.pal(6,"Dark2"),max.word = 100))
```

We can see from graphs above, in general, friend, family, home is top four things makes people happy. Friend is the most important source of happiness.

1.4 Reconstruct tidy text with stem words

```
completed1 <- completed %>%
  group_by(stems) %>%
  count(dictionary) %>%
  mutate(word = dictionary[which.max(n)]) %>%
  ungroup() %>%
  select(stems, word) %>%
  distinct() %>%
  right_join(completed) %>%
  select(-stems)
```

```
## Joining, by = "stems"
```

```
completed2 <- completed1 %>%
  group_by(id) %>%
  summarise(text = str_c(word, collapse = " ")) %>%
  ungroup()

hm_data <- hm_data %>%
  mutate(id = row_number()) %>%
  inner_join(completed2)
```

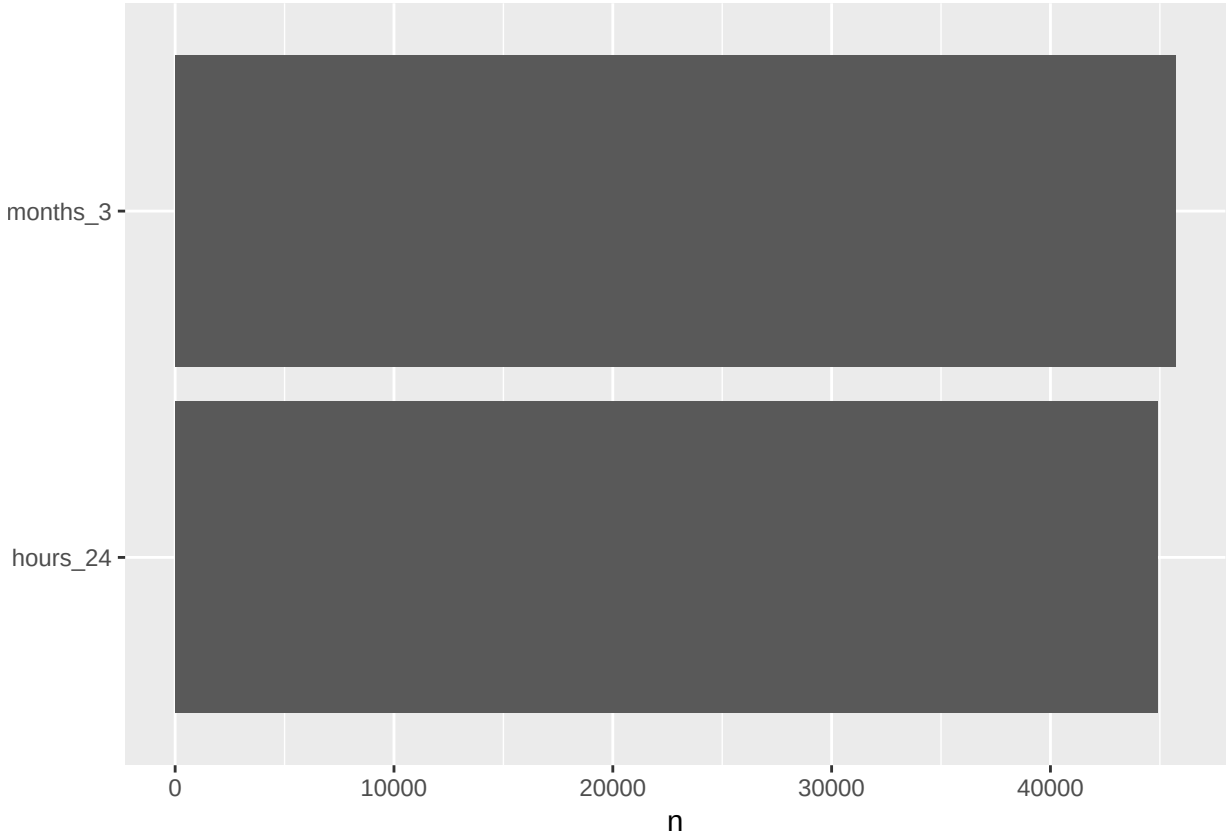
```
## Joining, by = "id"
```

1.5 Combined text data to demographic data

```
hm_data1 <- hm_data %>%
  inner_join(demo_data, by = "wid") %>%
  select(wid,
         id,
         original_hm,
         gender,
         marital,
         parenthood,
         reflection_period,
         age,
         country,
         ground_truth_category,
         predicted_category,
         text) %>%
  mutate(count = sapply(hm_data$text, wordcount)) %>%
  filter(gender %in% c("m", "f")) %>%
  filter(marital %in% c("single", "married")) %>%
  filter(parenthood %in% c("n", "y")) %>%
  filter(reflection_period %in% c("24h", "3m")) %>%
  mutate(reflection_period = fct_recode(reflection_period,
                                        months_3 = "3m", hours_24 = "24h"))%>%
  filter(country %in% c("USA", "IND","VEN"))
```

1.6 Happy moment reflection time

```
hm_data1 %>%
  count(reflection_period, sort = TRUE) %>%
  mutate(stems = reorder(reflection_period, n)) %>%
  ggplot(aes(reflection_period, n)) +
  geom_col() +
  xlab(NULL) +
  coord_flip()
```

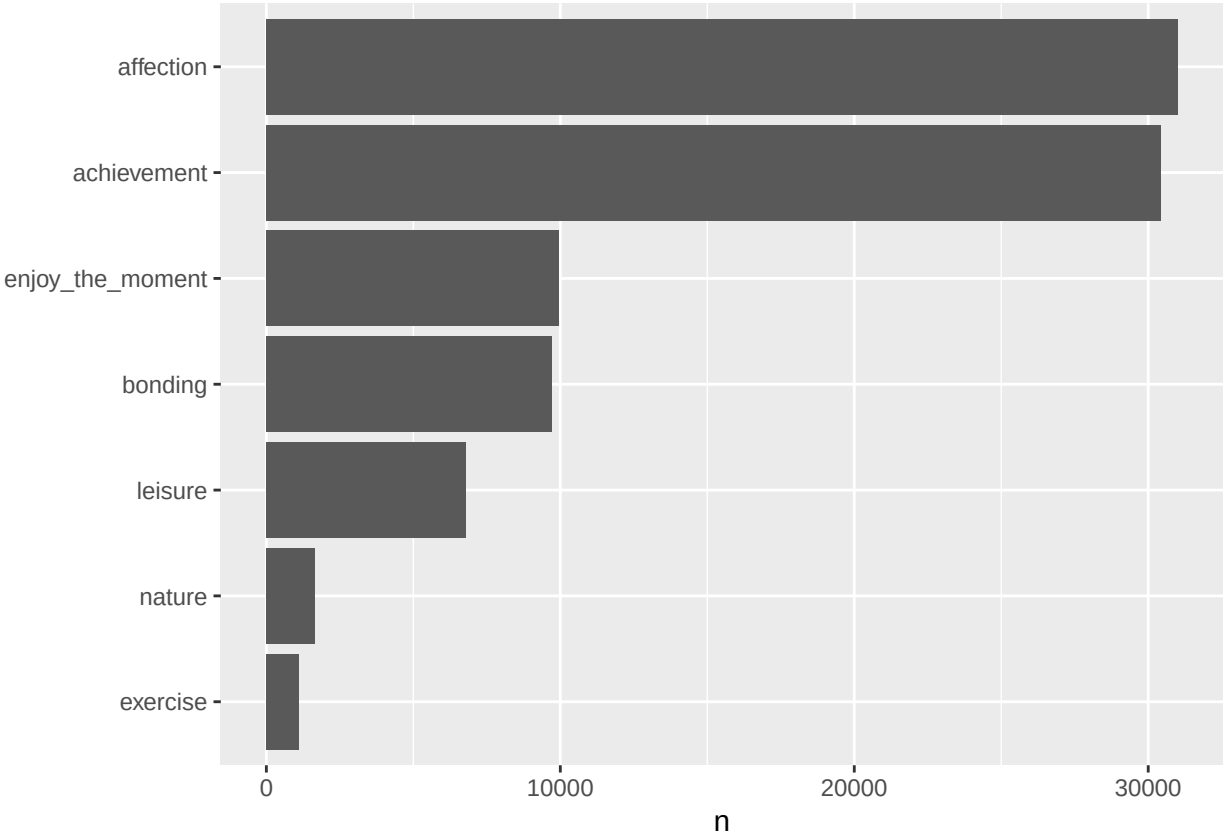


Half people reflect their last happy moment in last 3 months. We guess that is because some people are not easily to be happy nowadays.

1.7 Category frequency

```
cate <- hm_data1 %>%
  count(predicted_category, sort = TRUE)

cate%>%
  mutate(predicted_category = reorder(predicted_category, n)) %>%
  ggplot(aes(predicted_category, n)) +
  geom_col() +
  xlab(NULL) +
  coord_flip()
```



From this picture we can see people mostly feel happy in affection, achievement and enjoy_the_moment categories. This means our relationships and career are equally important.

Different kind of person's different happy moments

2.1 Word frequency within each happy category

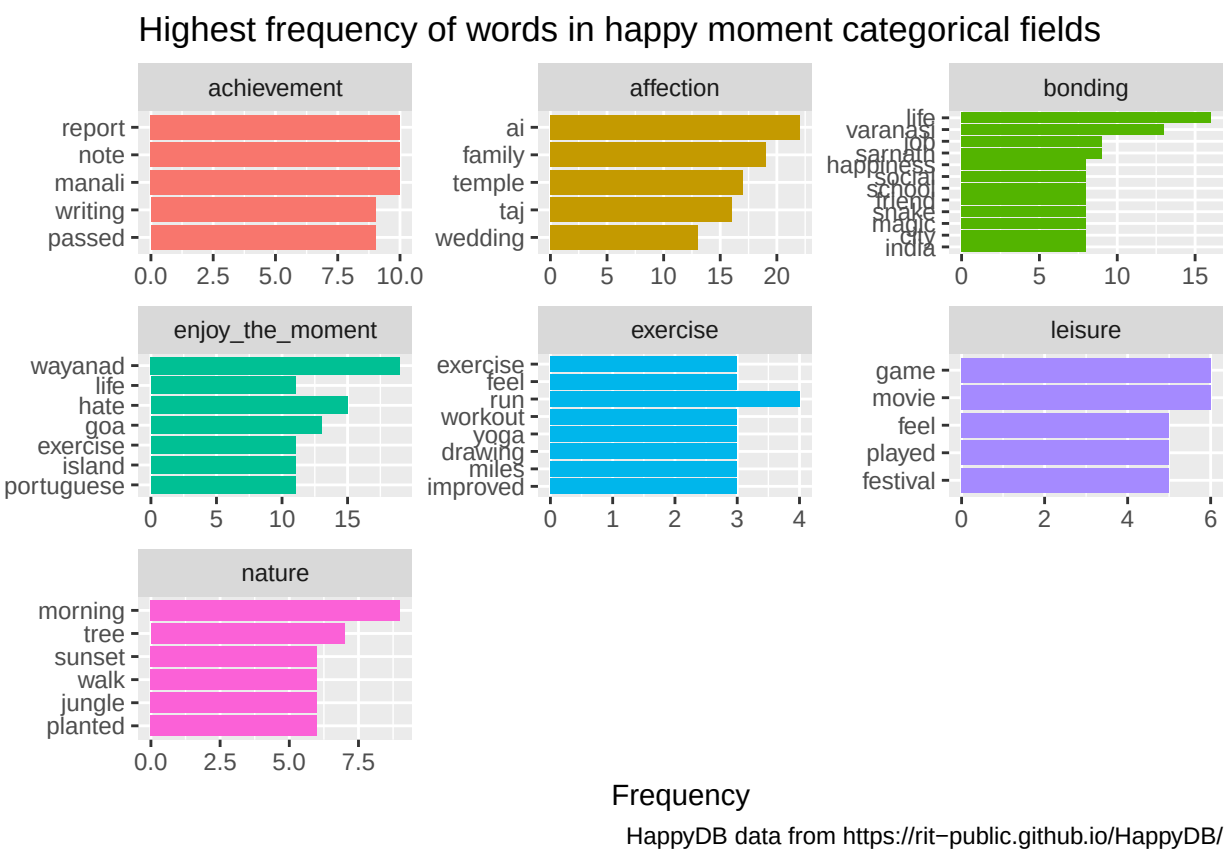
```
highnum1 <- hm_data1 %>%
  select(id,
         predicted_category,
         text) %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  count(id, word, sort = TRUE) %>%
  ungroup()
```

```
## Joining, by = "word"
```

```
cate2 <- hm_data1 %>%
  select(id,
         predicted_category)
highnum1 <- full_join(highnum1, cate2)
```

```
## Joining, by = "id"
```

```r
# Graph
 highnum1 %>%
  arrange(desc(n)) %>%
  group_by(predicted_category) %>%
  distinct(word, predicted_category, .keep_all = TRUE) %>%
  top_n(5, n) %>%
  ungroup() %>%
  mutate(word = factor(word, levels = rev(unique(word)))) %>%
  ggplot(aes(word, n, fill = predicted_category)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~predicted_category, ncol = 3, scales = "free") +
  coord_flip() +
  labs(title = "Highest frequency of words in happy moment categorical fields",
       caption = "HappyDB data from https://rit-public.github.io/HappyDB/",
       x = NULL, y = "Frequency")
```

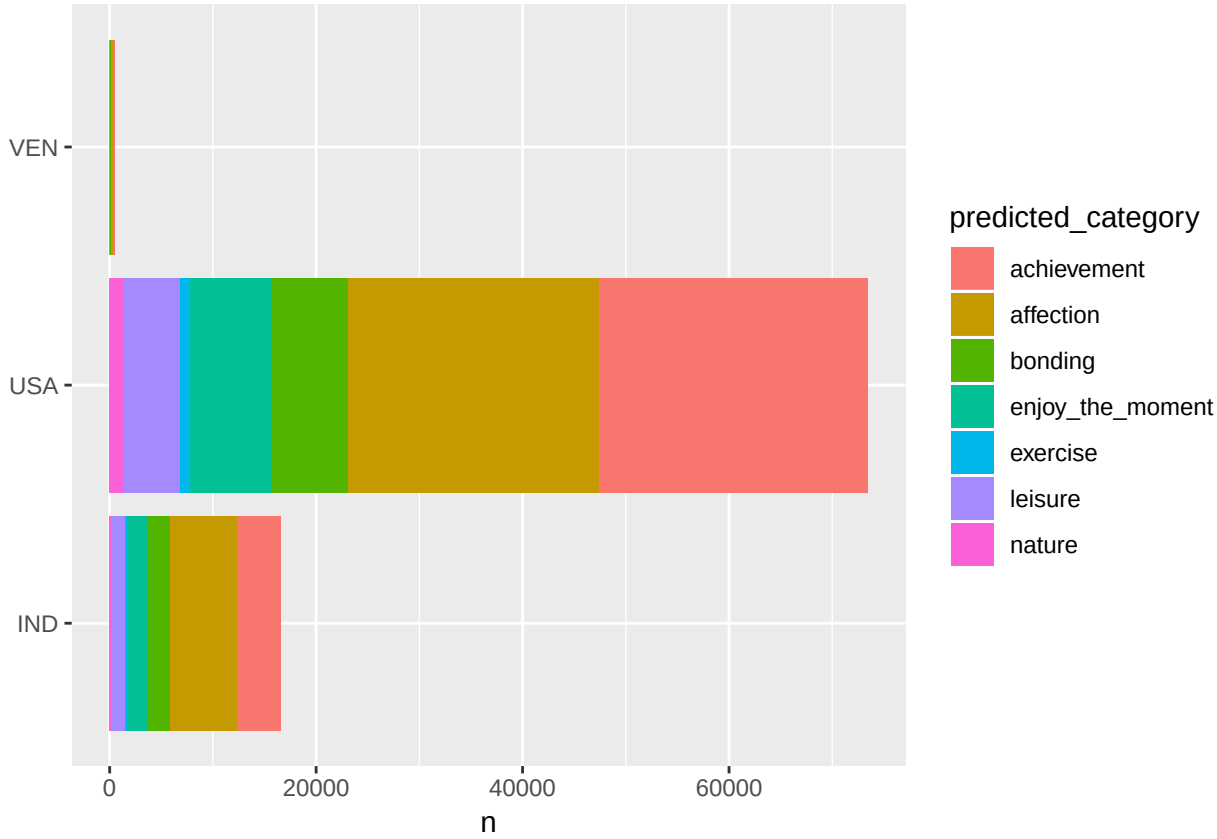


Temple for achievement, ai and family for affection, wedding and job for bonding, sleep and eat for enjoy_the_momeng, life and project for exercise, island and travel for leisure, morning and dog for nature. But we also can see the predicted category "achievement" is not that precise.

2.2 Country

```r
cate1 <- hm_data1 %>%
  group_by(country) %>%
  count(predicted_category)
```

```
cate1%>%
  ggplot(aes(country,n))+
  geom_col(aes(fill = predicted_category))+
  xlab(NULL) +
  coord_flip()
```



Achievement and Affection is the top common happy categories among those countries. We can see the difference in exercise and nature, which Indian people seldom feel happy while excercising or they don't exercise as frequent as American.

2.3 Gender

```
gen1 <- hm_data1 %>%
  group_by(gender) %>%
  count(predicted_category)

set.seed(0)
gen1 %>%
  acast(predicted_category ~ gender, value.var = "n", fill = 0) %>%
  comparison.cloud(colors = c("gray20", "gray80"),
                   max.words = 100)
```

We can see from this wordcloud that male's happy moment most from achievement, female's happy moment mostly from affection. This accords with common sense that females are more sensitive than male. And we can guess that male cares about their career and achievements more.

2.4 Marital

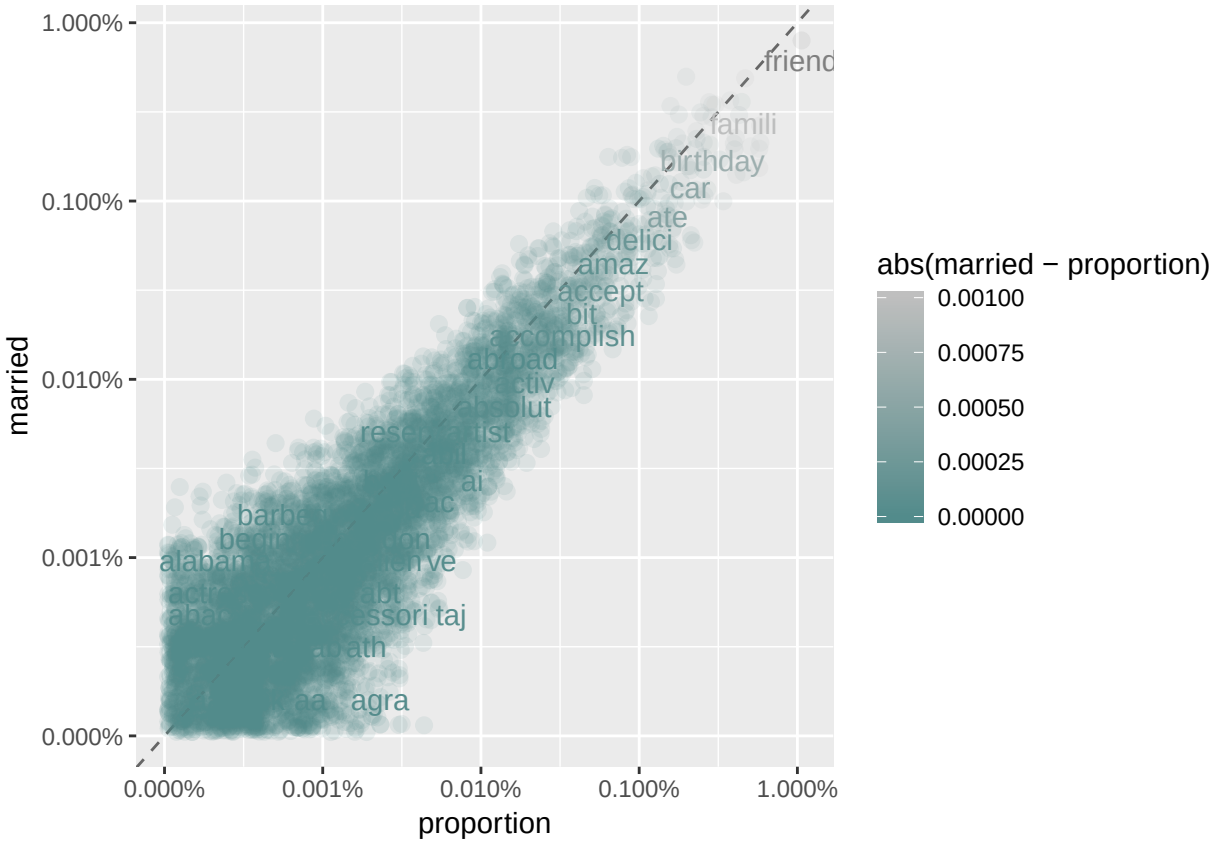
```
mp1 <- hm_data1 %>%
  select(id,
         marital,
         parenthood,
         text) %>%
  filter(marital %in% c("single", "married")) %>%
  filter(parenthood %in% c("n", "y"))
#stem dataframe
corpusmp1<- VCorpus(VectorSource(mp1$text))
stemmedmp1 <- tm_map(corpusmp1, stemDocument) %>%
  tidy() %>%
  select(text)
compmp1 <- stemmedmp1 %>%
  mutate(id = row_number())
compmp2 <- mp1 %>%
  inner_join(compmp1, by = "id")%>%
  mutate(text = text.y)%>%
  select(-c(text.x,text.y))
tidmp1 <- compmp2 %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words)
```

```
## Joining, by = "word"
```

```
##Calculate word frequency by marital
mar1 <- tidmp1 %>%
  mutate(word = str_extract(word, "[a-z']+")) %>%
        count(marital,word, sort = TRUE)%>%
  mutate(proportion = n/sum(n))%>%
  group_by(marital)%>%
  select(-n)%>%
  spread(marital,proportion)%>%
  gather(marital, proportion, `single`)
# ggplot picture
ggplot(mar1, aes(x = proportion, y = `married`, color = abs(`married` - proportion)))+
  geom_abline(color = "gray40", lty = 2) +
  geom_jitter(alpha = 0.1, size = 2.5, width = 0.3, height = 0.3) +
  geom_text(aes(label = word), check_overlap = TRUE, vjust = 1.5) +
  scale_x_log10(labels = percent_format()) +
  scale_y_log10(labels = percent_format()) +
  scale_color_gradient(limits = c(0, 0.001), low = "darkslategray4", high = "gray75")
```

```
## Warning: Removed 9723 rows containing missing values (geom_point).
```

```
## Warning: Removed 9723 rows containing missing values (geom_text).
```



The words are approximately located on the abline, no obvious outliers. We guess that marriage does not affect happy moment that much. No matter married or single, friend makes people happy most.

2.5 Prenthood

```
# calculate word frequencey by parenthood
tidmp2 <- mp1 %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words)
```

```
## Joining, by = "word"
```

```
par1 <- tidmp2 %>%
  group_by(parenthood) %>%
  count(word,sort= TRUE)

set.seed(0)
par1 %>%
  acast(word ~ parenthood, value.var = "n", fill = 0) %>%
  comparison.cloud(colors = c("gray20", "gray80"),
                   max.words = 100)
```

For peple who have kids, we can see “daughter”,“son”,“family” and “husband” is most shown on the wordcloud. Their happy moments occurs more in family life. For people who don’t have kids, we can see girlfriend, boyfriend and dog are important happiness source.

How to be happy

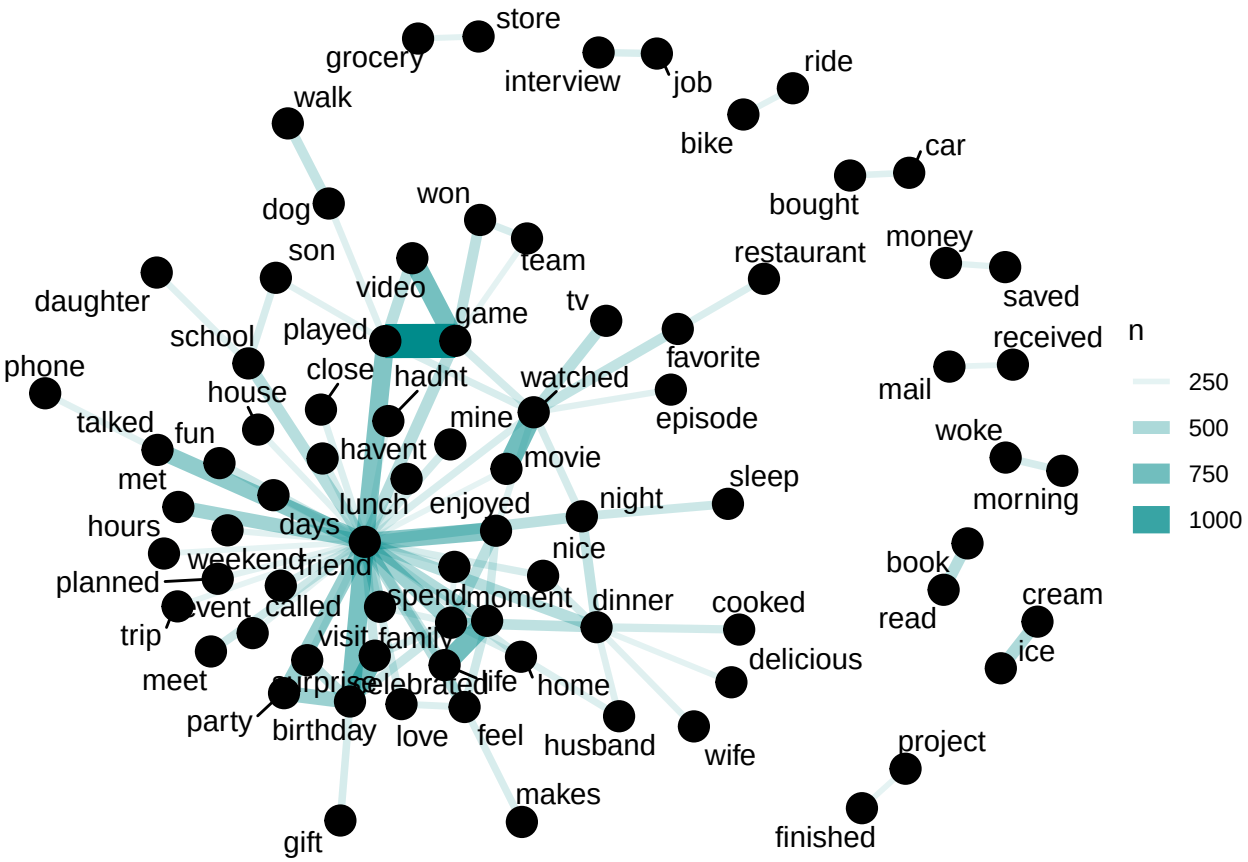
```
head(completed1)
```

```
## # A tibble: 6 x 3
##   word            id dictionary
##   <chr>        <int> <chr>
## 1 successfully     1 successful
## 2 date             1 date
## 3 sympathy         1 sympathy
## 4 connected        1 connection
## 5 son              2 son
## 6 marks            2 marks
```

```
word_pairs <- completed1 %>%
  pairwise_count(word, id, sort = TRUE, upper = FALSE)
```

3.1 Co-occurring words

```
set.seed(1234)
word_pairs %>%
  filter(n >= 250) %>%
  graph_from_data_frame() %>%
  ggraph(layout = "fr") +
  geom_edge_link(aes(edge_alpha = n, edge_width = n), edge_colour = "cyan4") +
  geom_node_point(size = 5) +
  geom_node_text(aes(label = name), repel = TRUE,
                 point.padding = unit(0.2, "lines")) +
  theme_void()
```

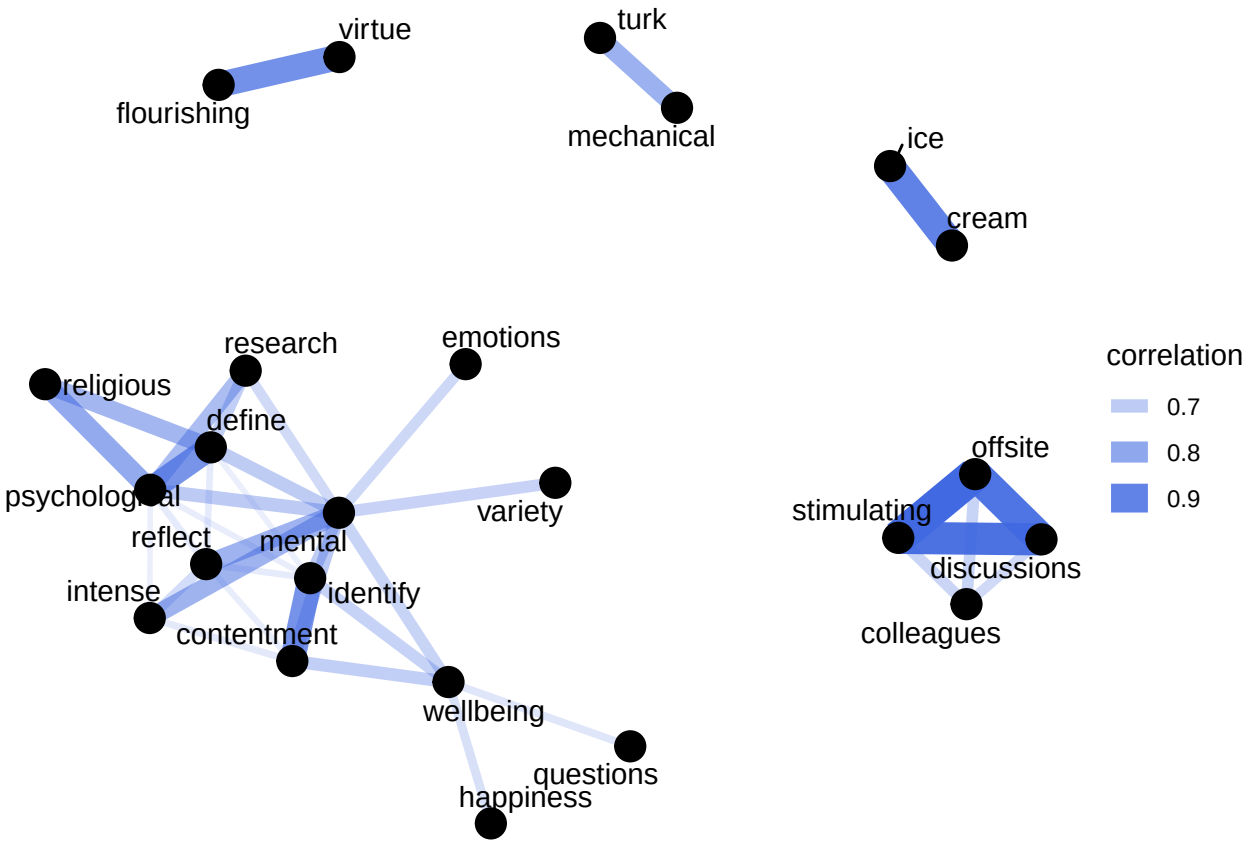


We see some clear clustering in this network of happy moment words; words in happy moments sentences are largely organized into several families of words that tend to go together. Most people tend to have similar happy moment discription, such as “friend time”, “game played”, and “celebrate birthday”, etc. Participating in those events are ways to be happy.

3.2 Words correlation

```
word_cors <- completed1 %>%
  group_by(dictionary) %>%
  filter(n() >= 100) %>%
  pairwise_cor(dictionary, id, sort = TRUE, upper = FALSE)

# graph the relationship
set.seed(1234)
word_cors %>%
  filter(correlation > .6) %>%
  graph_from_data_frame() %>%
  ggraph(layout = "fr") +
  geom_edge_link(aes(edge_alpha = correlation, edge_width = correlation), edge_colour = "royalblue") +
  geom_node_point(size = 5) +
  geom_node_text(aes(label = name), repel = TRUE,
                 point.padding = unit(0.2, "lines")) +
  theme_void()
```



This network above appears much different than the co-occurence network. The difference is that the co-occurrence network asks a question about which word pairs occur most often, and the correlation network

asks a question about which words occur more often together than with other words. This word networks tend to explain how to be happy.

```
write_csv(hm_data1, "/Users/peiluzhang/Documents/GitHub/Fall2018-Proj1-PuleiPai/output/merged_data.csv")
write_csv(word_cors, "/Users/peiluzhang/Documents/GitHub/Fall2018-Proj1-PuleiPai/output/wordcor_data.csv
write_csv(word_pairs, "/Users/peiluzhang/Documents/GitHub/Fall2018-Proj1-PuleiPai/output/wordpar_data.cs
```