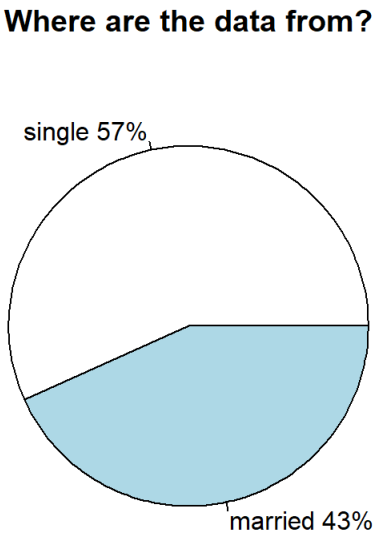
Project1_new

Sizhu Chen sc4248

What will be changed by marriage?

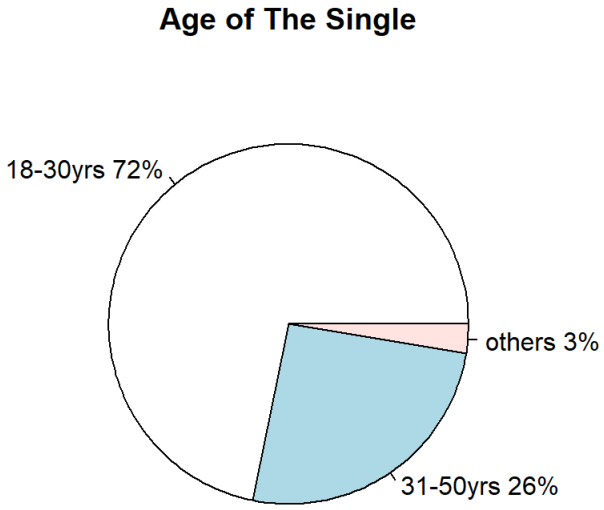
```
sing_wt_marr<-data.frame(lb=c("single","married"),dt=c(sum(hm_data$marital=="single")/nrow(hm_data),
                                    sum(hm_data$marital=="married")/nrow(hm_data)))

lbls1 <- paste(sing_wt_marr$lb, round(sing_wt_marr$dt*100)) # add percents to labels
lbls1 <- paste(lbls1,"%",sep="") # ad % to labels
pie(sing_wt_marr$dt,labels = lbls1,main="Where are the data from?")
```



We got more HappyDB data from single people than married ones. But there's no huge disparity. So we can regard them as half-to-half roughly, thus our analysis can go on without big deviation from sample collecting.

Who is single now?

```
# Who is single now?
single_group<-hm_data[hm_data$marital=="single","age"]
single_group[is.na(single_group)]<-"prefer not to say"

age1<-as.numeric(single_group$age[single_group$age!="prefer not to say"& single_group$age!="60yrs"])

age_of_single<-data.frame(age_label=c("18-30yrs","31-50yrs","others"),
                          age_count=c(sum(age1>17&age1<=30)/nrow(single_group),
                                      sum(age1>30&age1<=50)/nrow(single_group),
                                      (sum(age1<=17)+sum(age1>50)+sum(single_group$age=="prefer not to say"
))/nrow(single_group)))

pct <- round(age_of_single$age_count*100)
lbls <- paste(age_of_single$age_label, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(age_of_single$age_count,labels = lbls,main="Age of The Single")
```

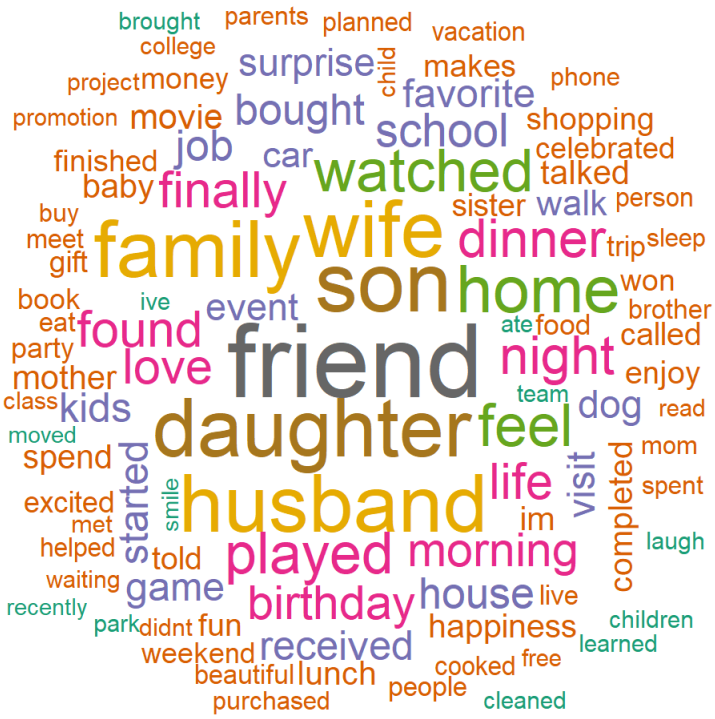


Just as we imagine, most of the single's data are from people between 18 to 30 years old, which means young people. They take nearly three quarters. Middle-aged people contribute a quarter of the samples.

What the keywords of happiness?

```
###Part 1. compare the word clouds
wordcloud2(word_count_single[1:100,],size=0.8,shape="circle",color="random-light",backgroundColor = "black")
```

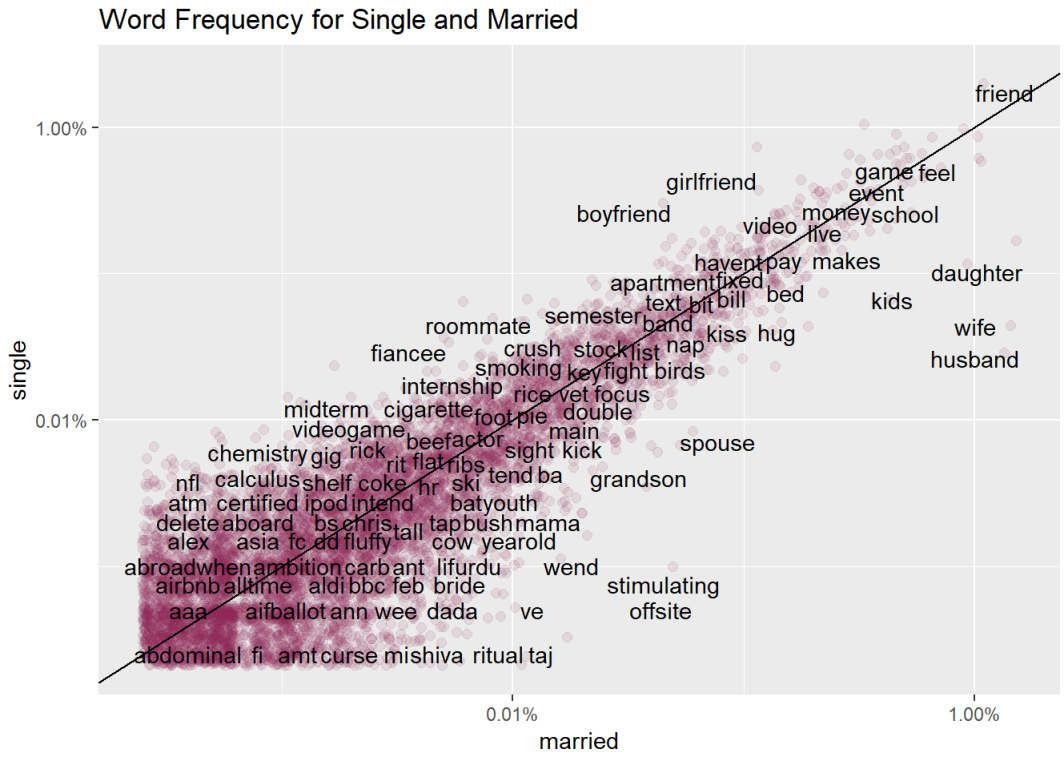

```
library(wordcloud)

wordcloud(words =word_count_married$word[1:100],
          freq = word_count_married$n[1:100],
          random.order=FALSE,colors=brewer.pal(8, "Dark2") )
```



The keywords of happiness for the single are shown with the cooler black background. The biggest, striking word is FRIEND. Around it we can see “watched”, “played”, “family”,“job”,etc. When it comes to the married people, FRIEND is still the biggest but, I have to mention, some other words become almost as large as “friend”: “son”,“daughter”,“husband”,“wife”,“family”.

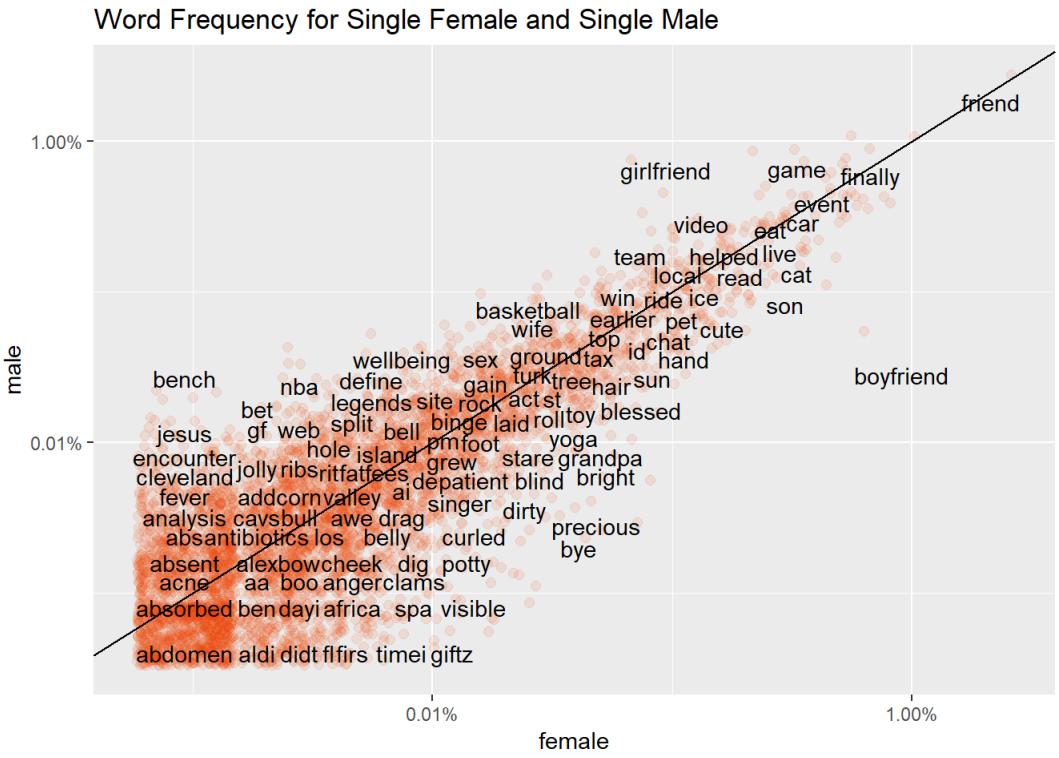
It’s not difficult to distinguish these two atmosphere. Wordcloud for the single is more relaxing, energetic and free. While the second wordcloud is warm and linked by kinship.

```
###Part 2. word frequency

## total: btw single & married
frequency_marital <- bag_of_words %>%
  group_by(marital) %>%
  count(word, sort = TRUE) %>%
  left_join(bag_of_words %>%
              group_by(marital) %>%
              summarise(total = n())) %>%
  mutate(freq = n/total)

frequency_marital<- frequency_marital %>%
  select(marital, word, freq) %>%
  spread(marital, freq) %>%
  arrange(married,single)

ggplot(frequency_marital, aes(married,single)) +
  geom_jitter(alpha = 0.1,color= "violetred4", size = 2, width = 0.2, height = 0.2) +
  labs(title="Word Frequency for Single and Married",x="married",y="single")+
  geom_text(aes(label = word), check_overlap = TRUE, vjust = 1.5) +
  scale_x_log10(labels = percent_format()) +
  scale_y_log10(labels = percent_format()) +
  geom_abline(color = "black")
```



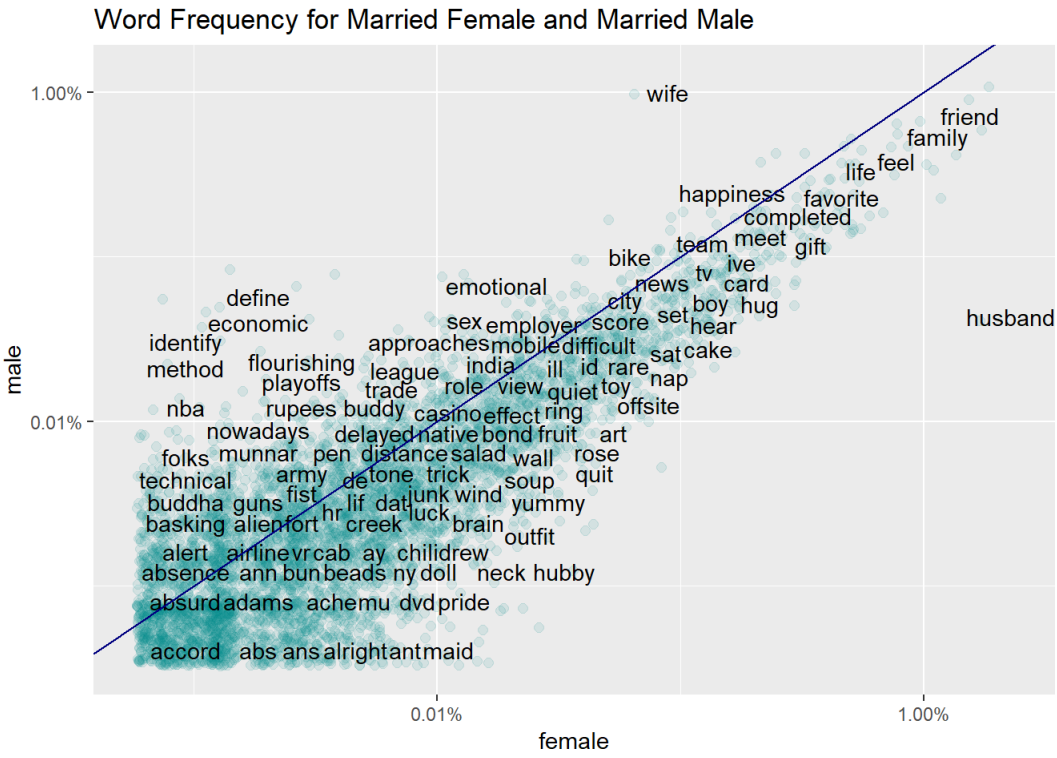
Let's explain it more clearly. The plot above is word frequency of the single and the married. "friend" is prominent at the highest position, near the divider line. Married people care more about their companions than single people, and their kids are one of the most important reasons to happiness.

```
##branch: btw female & male in single people

frequency_single <- bag_of_words_single %>%
  group_by(gender) %>%
  count(word, sort = TRUE) %>%
  left_join(bag_of_words_single %>%
              group_by(gender) %>%
              summarise(total = n())) %>%
  mutate(freq = n/total)

frequency_single<- frequency_single %>%
  select(gender, word, freq) %>%
  spread(gender, freq) %>%
  arrange(f,m)

ggplot(frequency_single, aes(f,m)) +
  geom_jitter(alpha = 0.1,color= "orangered2", size = 2, width = 0.2, height = 0.2) +
  labs(title="Word Frequency for Single Female and Single Male",x="female",y="male")+
  geom_text(aes(label = word), check_overlap = TRUE, vjust = 1.5) +
  scale_x_log10(labels = percent_format()) +
  scale_y_log10(labels = percent_format()) +
  geom_abline(color = "black")
```



Among the single people, obviously women and men have different hapiness. Besides “boyfriend”, single women enjoy the feeling of “love”, “share”, “blessed”, and “cat”,“grandparent”,“purse” can make them delighted. Single men put more focus on “game”, “video”, “team”, “basketball”, “smoking”.

```
##branch: btw female & male in married people

frequency_married <- bag_of_words_married %>%
  group_by(gender) %>%
  count(word, sort = TRUE) %>%
  left_join(bag_of_words_single %>%
              group_by(gender) %>%
              summarise(total = n())) %>%
  mutate(freq = n/total)

frequency_married<- frequency_married %>%
  select(gender, word, freq) %>%
  spread(gender, freq) %>%
  arrange(f,m)

ggplot(frequency_married, aes(f,m)) +
  geom_jitter(alpha = 0.1,color= "cyan4", size = 2, width = 0.2, height = 0.2) +
  labs(title="Word Frequency for Married Female and Married Male",x="female",y="male")+
  geom_text(aes(label = word), check_overlap = TRUE, vjust = 1.5) +
  scale_x_log10(labels = percent_format()) +
  scale_y_log10(labels = percent_format()) +
  geom_abline(color = "navy")
```

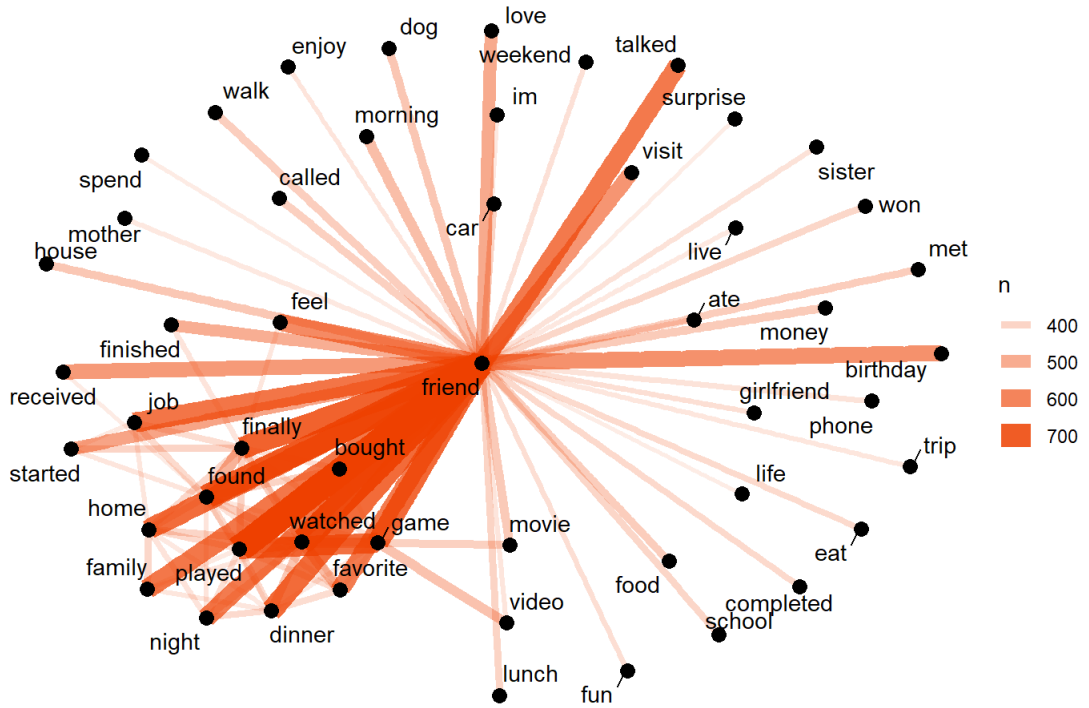


After getting married, people changed. Women put more weights on her family, and men get more mature. “video”, “game” don’t take that high places anymore.

Network Analysis (Word co-ocurrences and correlations)

```
###Part 3. network analysis/Word co-ocurrences and correlations
word_pairs_single <- bag_of_words_single %>%
  pairwise_count(word, wid, sort = TRUE, upper = FALSE)

set.seed(1234)
word_pairs_single[1:100,] %>%
  graph_from_data_frame() %>%
  ggraph(layout = "fr") +
  geom_edge_link(aes(edge_alpha = n, edge_width = n), edge_colour = "orangered2") +
  geom_node_point(size = 3) +
  geom_node_text(aes(label = name), repel = TRUE,point.padding = unit(0.2, "lines")) +
  labs(title="Network for Single People")+
  theme_void()
```

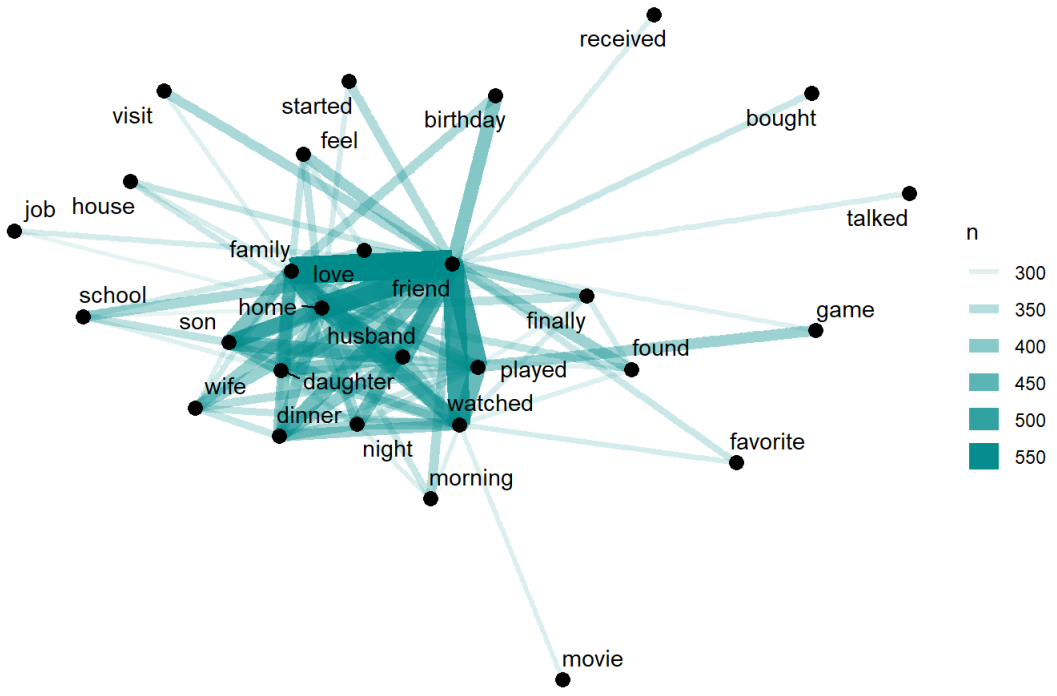
Network for Single People



From the orange dandelion on the plot, we can learn that the world of singles is fairly simple. The center of happiness is “friend”. They are willing to do almost everything with their friends, thus get a lot of good moments.

```
word_pairs_married <- bag_of_words_married %>%
  pairwise_count(word, wid, sort = TRUE, upper = FALSE)

word_pairs_married[1:100,] %>%
  graph_from_data_frame() %>%
  ggraph(layout = "fr") +
  geom_edge_link(aes(edge_alpha = n, edge_width = n), edge_colour = "cyan4") +
  geom_node_point(size = 3) +
  geom_node_text(aes(label = name), repel = TRUE,point.padding = unit(0.2, "lines")) +
  labs(title="Network for Married People")+
  theme_void()
```

Network for Married People

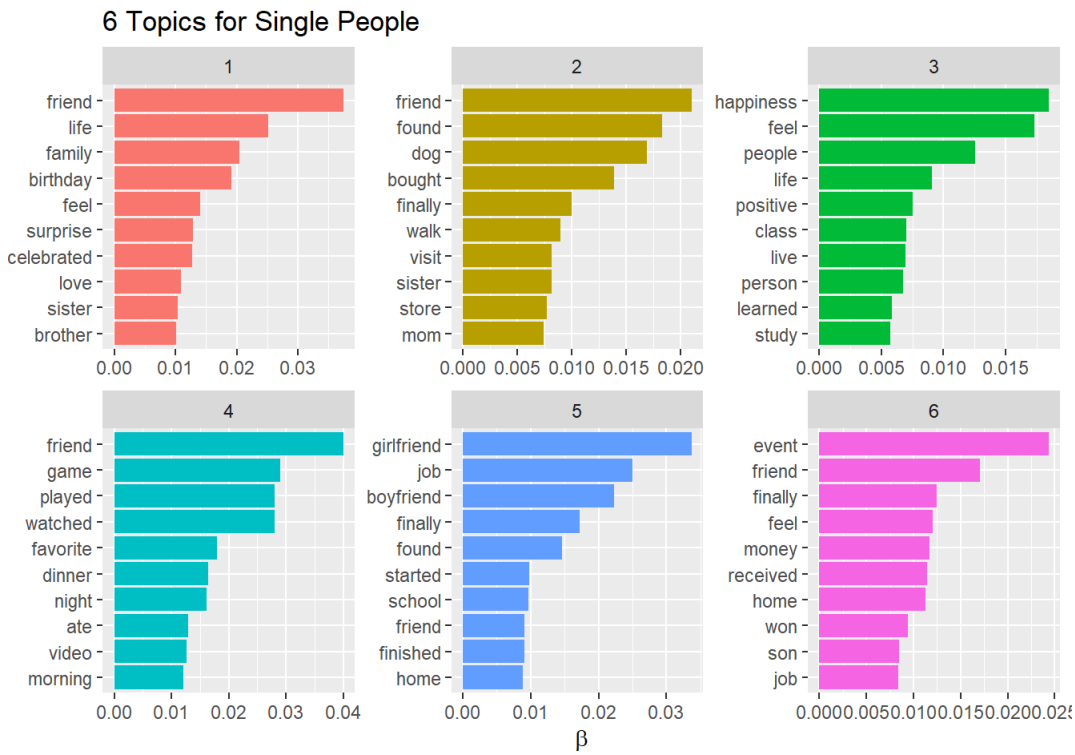


The hapiness of married people is getting complicated. The divergent part is from “friend”, as we can understand. Meanwhile the keywords from family are interactional. Relatively they spend less time with friends than before. However, they prefer to combine “friend” with “family”

to create happy memories together.

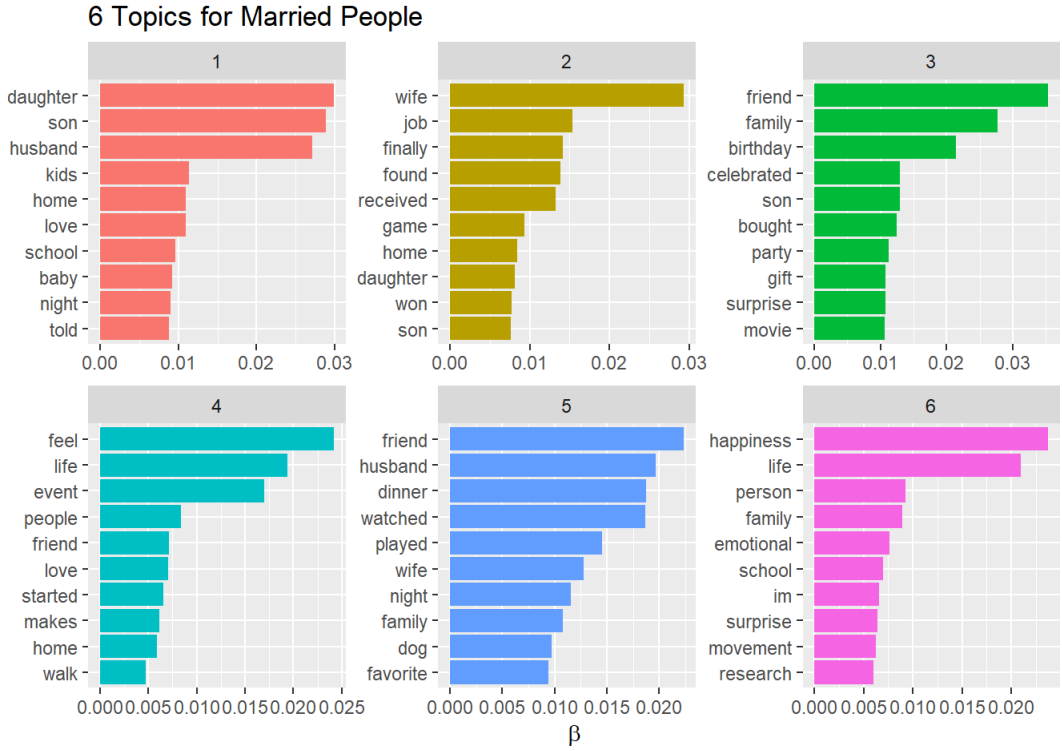
Topic Modeling

```
###Part 4.Topic Modeling

library(topicmodels)

##for single people
word_count_new_single <- bag_of_words_single %>%
  count(wid, word, sort = TRUE) %>%
  ungroup()

word_dtm_single <- word_count_new_single %>%
  cast_dtm(wid, word, n)

word_lda_single <- LDA(word_dtm_single, k = 6, control = list(seed = 1234))

tidy_lda_single<-tidy(word_lda_single)

top_terms_single <- tidy_lda_single %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

top_terms_single %>%
  mutate(term = reorder(term, beta)) %>%
  group_by(topic, term) %>%
  arrange(desc(beta)) %>%
  ungroup() %>%
  mutate(term = factor(paste(term, topic, sep = "__"),
                       levels = rev(paste(term, topic, sep = "__")))) %>%
  ggplot(aes(term, beta, fill = as.factor(topic))) +
  geom_col(show.legend = FALSE) +
  coord_flip() +
  scale_x_discrete(labels = function(x) gsub("__.+$", "", x)) +
  labs(title = "6 Topics for Single People",
       x = NULL, y = expression(beta))+
  facet_wrap(~ topic, scales = "free",ncol = 3)
```



We can approximately summarize the topics of the single's happy moments: birthday, dog and siblings, study, games, relationship, and money.

```
##for married people
word_count_new_married <- bag_of_words_married %>%
  count(wid, word, sort = TRUE) %>%
  ungroup()

word_dtm_married<- word_count_new_married %>%
  cast_dtm(wid, word, n)

word_lda_married <- LDA(word_dtm_married, k = 6, control = list(seed = 1234))

tidy_lda_married<-tidy(word_lda_married)

top_terms_married <- tidy_lda_married %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

top_terms_married %>%
  mutate(term = reorder(term, beta)) %>%
  group_by(topic, term) %>%
  arrange(desc(beta)) %>%
  ungroup() %>%
  mutate(term = factor(paste(term, topic, sep = "__"),
                       levels = rev(paste(term, topic, sep = "__")))) %>%
  ggplot(aes(term, beta, fill = as.factor(topic))) +
  geom_col(show.legend = FALSE) +
  coord_flip() +
  scale_x_discrete(labels = function(x) gsub("__.+$", "", x)) +
  labs(title = "6 Topics for Married People",
       x = NULL, y = expression(beta))+
  facet_wrap(~ topic, scales = "free",ncol = 3)
```

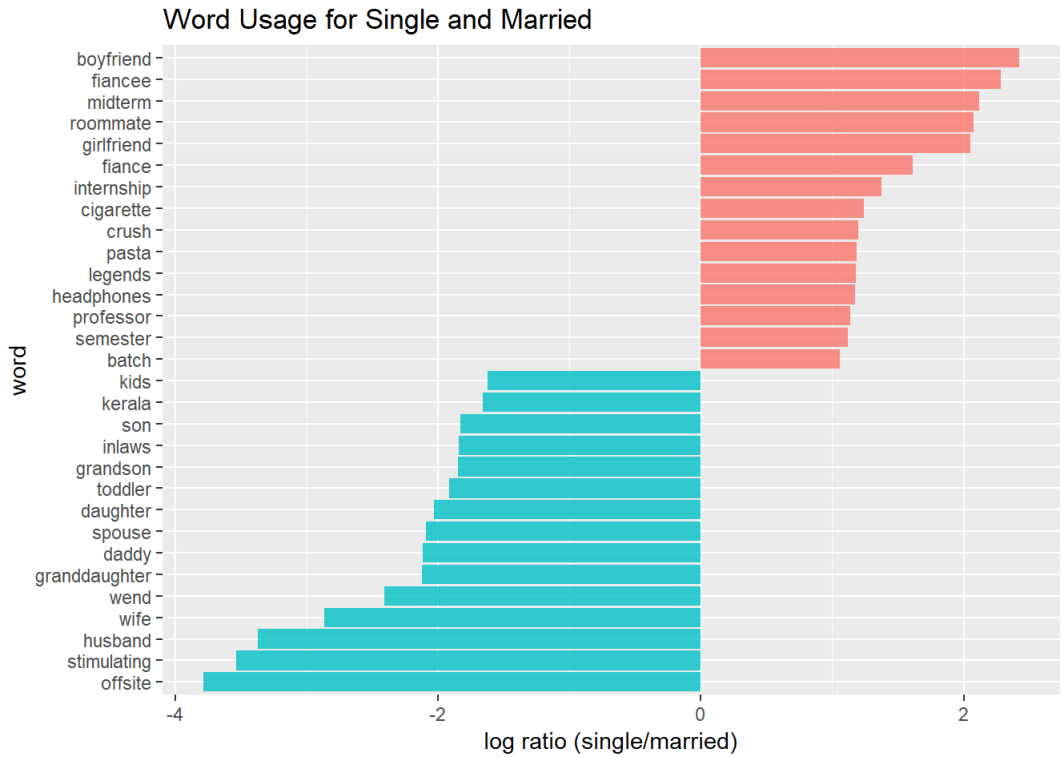


The 6 main topics for married people are: kids, career, brithday, friend events, friends with family, study.

It's harder to summarize the topics for the latter ones, because the same collection of words keep showing up on and on. Possibly we can draw a conclusion that marriage makes people more difficult to be happy. Precisely, we can not only blame on marriage, but also aging. When you are young, games can make you cheerful, but it will be not enough after several years.

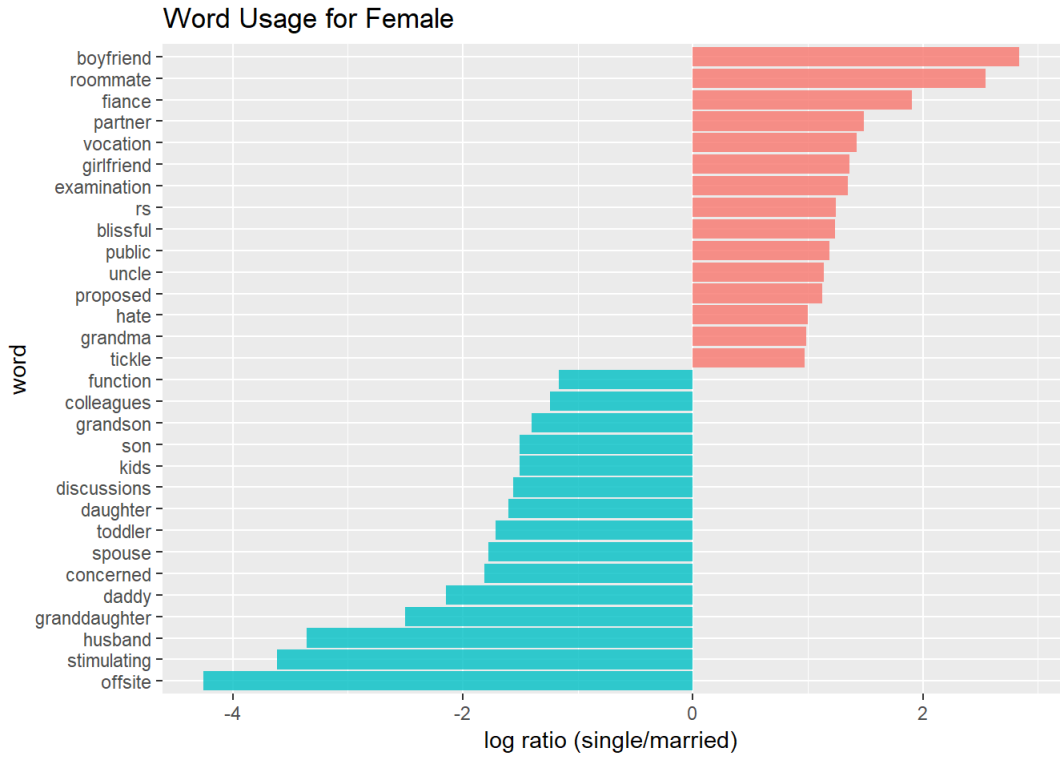
Word Usage Compare

```
###Part 5. word usage

##Total: btw single & married
word_ratios_marital <- bag_of_words %>%
  count(word, marital) %>%
  group_by(word) %>%
  filter(sum(n) >= 50) %>%
  ungroup() %>%
  spread(marital, n, fill = 0) %>%
  mutate_if(is.numeric, funs((. + 1) / (sum(.) + 1))) %>%
  mutate(logratio = log(single / married)) %>%
  arrange(desc(logratio))


word_ratios_marital %>%
  group_by(logratio < 0) %>%
  top_n(15, abs(logratio)) %>%
  ungroup() %>%
  mutate(word = reorder(word, logratio)) %>%
  ggplot(aes(word, logratio, fill = logratio < 0)) +
  geom_col(alpha=0.8,show.legend = FALSE) +
  coord_flip() +
  labs(title="Word Usage for Single and Married")+
  ylab("log ratio (single/married)") +
  scale_fill_discrete(name = "", labels = c("single", "married"))
```



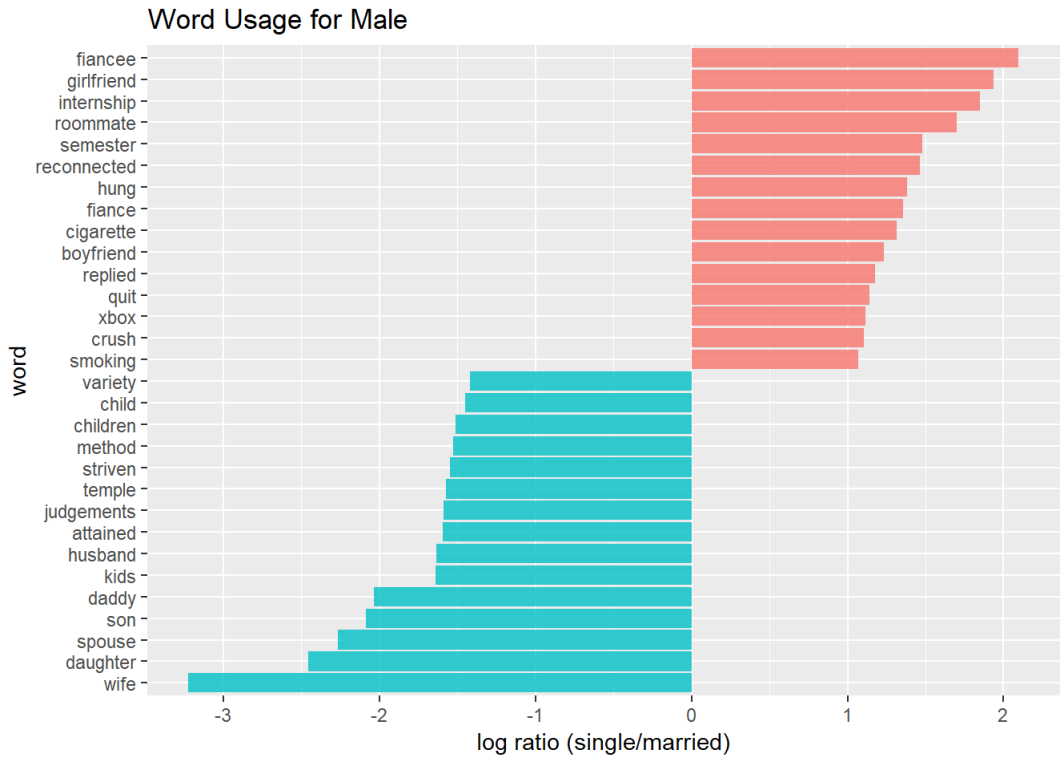
From the bar plot above we can get that the main change after getting married is the role change in relationship. Also some words will not be used anymore as age grows, like “roommate”, “internship”, “semester”.

```
##branch: the female change
bag_of_words_f<-bag_of_words[bag_of_words$gender=="f",]
bag_of_words_m<-bag_of_words[bag_of_words$gender=="m",]


word_ratios_f <- bag_of_words_f %>%
  count(word, marital) %>%
  group_by(word) %>%
  filter(sum(n) >= 50) %>%
  ungroup() %>%
  spread(marital, n, fill = 0) %>%
  mutate_if(is.numeric, funs((. + 1) / (sum(.) + 1))) %>%
  mutate(logratio = log(single /married )) %>%
  arrange(desc(logratio))

word_ratios_f %>%
  group_by(logratio < 0) %>%
  top_n(15, abs(logratio)) %>%
  ungroup() %>%
  mutate(word = reorder(word, logratio)) %>%
  ggplot(aes(word, logratio, fill = logratio < 0)) +
  geom_col(alpha=0.8,show.legend = FALSE) +
  coord_flip() +
  labs(title="Word Usage for Female")+
  ylab("log ratio (single/married)") +
  scale_fill_discrete(name = "", labels = c("single", "married"))
```



The married women will care more about families, including their children as well as their parents. Single ladies are always happy on various friends and emotional things.

```
##branch:male

word_ratios_m <- bag_of_words_m %>%
  count(word, marital) %>%
  group_by(word) %>%
  filter(sum(n) >= 50) %>%
  ungroup() %>%
  spread(marital, n, fill = 0) %>%
  mutate_if(is.numeric, funs((. + 1) / (sum(.) + 1))) %>%
  mutate(logratio = log(single /married )) %>%
  arrange(desc(logratio))

word_ratios_m %>%
  group_by(logratio < 0) %>%
  top_n(15, abs(logratio)) %>%
  ungroup() %>%
  mutate(word = reorder(word, logratio)) %>%
  ggplot(aes(word, logratio, fill = logratio < 0)) +
  geom_col(alpha=0.8,show.legend = FALSE) +
  coord_flip() +
  labs(title="Word Usage for Male")+
  ylab("log ratio (single/married)") +
  scale_fill_discrete(name = "", labels = c("single", "married"))
```



For males, they become hardly to get hapiness from their cigarettes and xbox, or to be more accurate, there are other delightful moments worthy of being remembered, which are more memorable.

Summary

To sum up, there are three things to be changed after getting married:

1. The freedom of your time. You definitely will lose some your time with friends and games. But cleverly many married people choose to invite friends over to share happy moments together with families.
2. The role in relationship. You will never have girlfriends or boyfriends, but instead you will get wife or husband. And you can upgrade to mommy or daddy, or even grandma or grandpa. Familiy makes a great contribution to your happiness.
3. The maturity in psychology. Games or TV cannot satisfy you anymore. You need bigger success in substance, emotion and interpersonal relationship.

No matter what's your age, married or not, enjoy the present. And take your time.