GR5243 Project 1: Do I Belong to Happiness?

Code **▼**

Jingwen Wang (jw3667)

Step 0: Basic Set-ups and Questions Raising

Brief Introduction on the Data:

The Data that we are using is *HappyDB* which is a corpus of 100,000 crowd-sourced happy moments via Amazon's Mechanical Turk. You can read more about it on https://arxiv.org/abs/1801.07746 (https://arxiv.org/abs/1801.07746)

According to the methods introduced in the lectures (credit to "Arpita Shah and Tian Zheng"), we are able to clean our data. Since the process is fully developed in the file "Text_Processing.Rmd", and "processed_moments.csv" (located in "output") is going to be used as the initial data file for my analysis.

The Research Question & Analysis Logic Flow:

Happy_DB provides 100,000 happy moments of people from diversified background. It is very common to ask a question:

Will someone of a certain type be happier?

Everyone knows the term happiness, but it is very difficult for us to define what happiness truely is. Philosophers, psychologists, artists, and even economists and politicians can give us various definition about happiness based on their theory and measurement. Someone may look happy everyday, but someone does not. So what makes us different? What makes people feel happy? Are we in the particular group of happiness people?

In this analysis, we may want to figure if there is a/some demographic characteristics will affect people's happiness. At the very end of this analysis, we may provide some insights into this problem.

If so, what are the special factors that affect the happiness of this group?

The analysis will follow steps below to examine how demographic characteristics will affect people's happiniess:

- **Step 1**: Combine the cleaned data with demographic information, including gender, marital status, parenthood, age, and country.
- Step 2: Use the Sentimental Analysis to quantify happiness level and to visualize it based on demographics
- Step 3: Based on the sentimental analysis results, conduct a behavior analysis on the selected demographic group(s) by Topic Modeling. This would answer if a certain group of people have some experience that would make them happier. (i.e. Do people with similar experience would be happy

because of similar activities)

Step 1: Combining demographic information to the data

Fist, we will check version and load packages that will be used in the analysis.

Hide

R.Version()\$version.string

[1] "R version 3.5.1 (2018-07-02)"

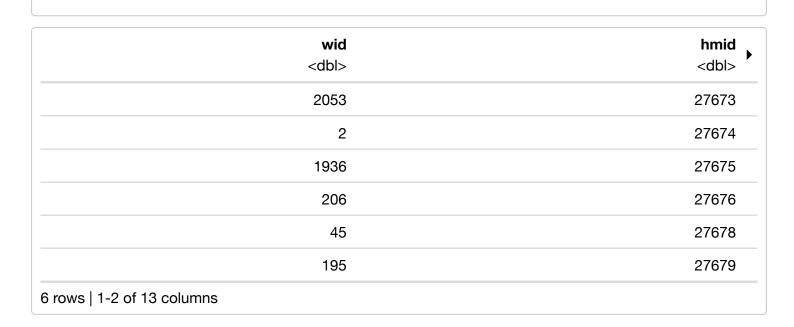
Load the processed text data along with demographic information on contributors:

We use the processed data for our analysis and combine it with the demographic information available. The file is directly read from output processed moments.csv.

Combine both the data sets and keep the required columns for analysis

We select a subset of the data that satisfies specific row conditions.

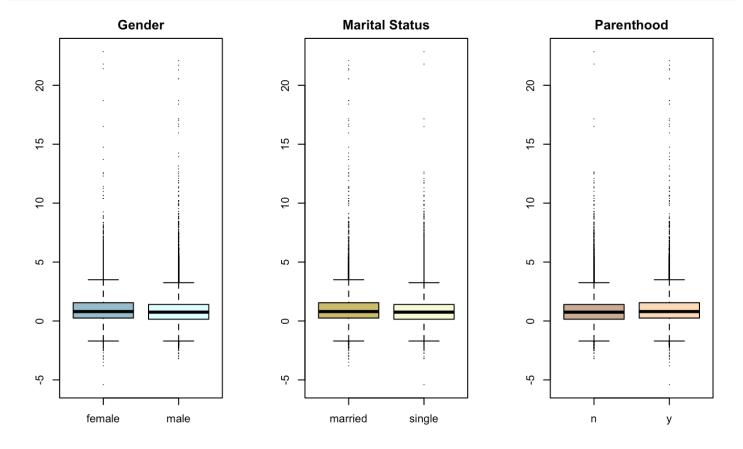
r r head(hm_data)



Step 2: Sentiment Analysis by Syuzhet

In the sentiment analysis, our goal is to evaluate the happiness level among all demographic groups. The

Visualizing happiniess level according to Gender, Marital Status, and Parenthood by Box-plot

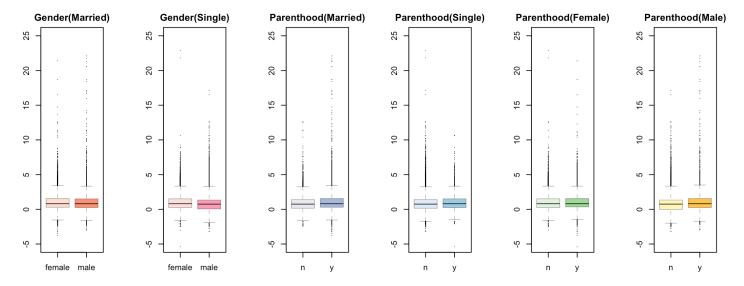


Five Number Summary of Gender, Marital Status, and Parenthood

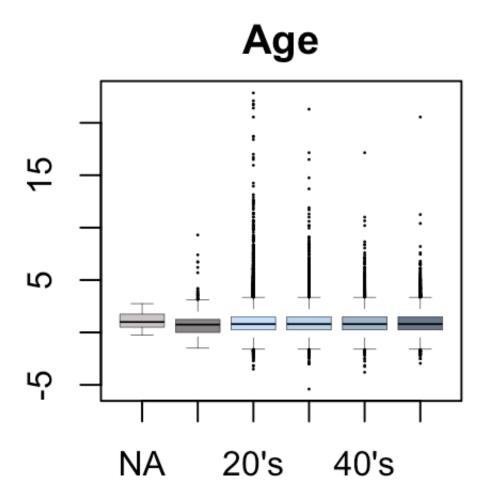
Gender(f/m) & Marrital(married/single) & Parenthood(n/y)
print(one.factor)

```
Min 1st Qu Median
                                     Mean 3rd Qu
                                                    Max
                   0.25
female
           -5.4
                          0.80 1.0343316
                                            1.55 22.85
male
           -3.2
                   0.15
                          0.75 0.9426862
                                            1.40 22.10
married
           -3.8
                   0.25
                          0.80 1.0369783
                                            1.55 22.10
                                            1.40 22.85
single
           -5.4
                   0.15
                          0.75 0.9368330
           -3.2
                   0.15
                          0.75 0.9310768
                                            1.40 22.85
parent
non-parent -5.4
                   0.25
                          0.80 1.0612110
                                            1.55 22.10
```

Visualizing Cross-groups by Box-plot



Visualizing Age group by the Box-plot



Conclusion of the Sentiment Analysis:

By performing the sentiment analysis, we look at the overall happiness level among all demographics, and we notice that there is no such a group whose happiness level is greatly higher than other groups. The overall hapiness index is around 1, suggesting that data we are collecting indeed showing happiness from people. Since the data is about happiness, we expect to view positive response, this explains why we have several extreme outliers on the positive direction but limited outliers on the negative direction.

Hightlights

- There is no clear pattern suggesting that age is a clear factor of hapiness.
- We notice that people with children are more likely to be happier than those without children.
- Female seems to be slightly happier than male.

Step 3: Study on Selected Demographic Group by Topic Modelling

According to our sentimental analysis results, we have the following observations:

- Married couples with kids seem to be happier than those without kids.
- Single with kids seems to be happier than bachelors without parenthood.

This leads to two following questions:

- 1. Does parenthood make people happier?
- 2. What experience/activities do parents/non-parents usually do? Is there any implication on their happiness?

According to *Thomas Hansen* in *Parenthood and Happiness: a Review of Folk Theories Versus Empirical Evidence*, the presence of children is considered as a widely accepted vital mode of happiness. Being a parent could make people feel fulfilled. Moreover, children are usually considered as the bond of a family which enhance the overall well-being by fostering greater family hapiness.

Therefore, we are going to apply *Topic Modelling* to figure if the patterns of events and/or relationships exist or not.

Visualization: How children will have an impact on people's happy moments

Wordcloud is used as a visualization tool to let us check what are the most frequent words in Parenthood(Y) and Parenthood(N) group. The results from the Wordcloud may indicate if parenthood takes a relatively big part of people's life.

```
# Wordcloud for Parenthood
par.data <- hm_data[hm_data$parenthood=="y",]

Hide

set.seed(2333)
cld.parent <- Corpus(VectorSource(par.data$text))
dtm.parent <- TermDocumentMatrix(cld.parent)
dtm.parent.ma <- as.matrix(dtm.parent)
dtm.parent.ma <- sort(rowSums(dtm.parent.ma),decreasing=TRUE)
d.parent.df <- data.frame(word = names(dtm.parent.ma),freq=dtm.parent.ma)

# Wordcloud for Non-Parenthood
no.par.data <- hm data[hm data$parenthood=="n",]</pre>
```

```
set.seed(2333)
cld.no.parent <- Corpus(VectorSource(no.par.data$text))
dtm.no.parent <- TermDocumentMatrix(cld.no.parent)
dtm.no.parent.ma <- as.matrix(dtm.no.parent)
dtm.no.parent.ma <- sort(rowSums(dtm.no.parent.ma),decreasing=TRUE)
d.no.parent.df <- data.frame(word = names(dtm.no.parent.ma),freq=dtm.no.parent.ma)</pre>
```

Visualization on People with Children:



Visualization on People with No Child:



Comment:

Apart from the top three words (friend, time, and day) shared by both "Patenhood(Y)" and "Parenthood(N)", we notice that words such as "son", "daughter", "family" are obviously high frenquency words in the group of "Patenhood(Y)". This suggests that most of happiness moments of this group come from their families, especially from their children.

Topic Modelling: Behavior Study on Parenthood(Y) Group (What experience are they sharing?)

Based on what we have from the Wordcloud, we want to further study what common experience/activities are shared with in Parenthood(Y) group. Later, we will conduct the same Topic Modelling approach on Parenthood(N) group, and to compare these two groups. The main purpose is to check if many happy moments of parents (i.e. activities that make people happy) are related to children.

```
# Group 1 Parenthood(Y)
# Cleaning Data
corpus.parent <- VCorpus(VectorSource(par.data$text))%>%
   tm_map(content_transformer(tolower))%>%
   tm_map(removePunctuation)%>%
   tm_map(removeNumbers)%>%
   tm_map(removeWords, character(0))%>%
   tm_map(stripWhitespace)
```

10 Topics in Parenthood(Y) Group (20 top words & words counts)

7	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic
[1,] [2,] [3,] iful"	"time" "family" "watched"	"finally" "found" "bought"	"home" "house" "baby"	"feel" "happiness" "makes"	"job" "received" "called"	"daughter" "son" "morning"	"walk" "trip" "beaut
[4,] ed"	"enjoyed"	"car"	"visit"	"live"	"completed"	"love"	"start
	"movie" "spend"	"money" "shopping"	"sister" "fun"	"people" "positive"	"finished" "book"	"dog" "started"	"bike" "templ
[7,]	"kids"	"purchased"	"brother"	"life"	"hours"	"told"	"plant
[8,]	"weekend"	"buy"	"helped"	"love"	"weeks"	"learned"	"garde
[9,]	"vacation"	"pay"	"excited"	"joy"	"read"	"laugh"	"takin
[10,]	"spent" "park"	"card" "store"	"mom" "child"	"world" "experience"	"project" "promotion"	"smile" "sleep"	"ride" "drive
[12,]	"person"	"hit"	"children"	"relatives"	"hard"	"hug"	"weath

!!							
	"relax"	"didnt"	"mother"	"person"	"ive"	"bed"	"flowe
rs" [14]	"song"	"saved"	"cleaned"	"emotional"	"company"	"run"	"decid
ed"	bong	buvcu	Officialica	Cinocionai	Company	1 411	acora
	"change"	"paid"	"moved"	"listening"	"office"	"woke"	"waiti
ng" [16,]	"planned"	"bill"	"parents"	"music"	"recently"	"cat"	"beach
11			_		-		
[17,] h"	"results"	"mturk"	"girl"	"research"	"proud"	"minutes"	"churc
[18,]	"movement"	"online"	"father"	"pleasant"	"couple"	"kiss"	"water
[19,] ed"	"waiting"	"extra"	"wedding"	"define"	"event"	"funny"	"stopp
[20,]	"study"	"fixed"	"boy"	"wellbeing"	"successfully"	"looked"	"hand"
	Topic 8	Topic 9	Topic 10				
[1,]	"day"	"friend"	"husband'	1			
[2,]	"moment"	"played"	"wife"				
[3,]	"school"	"birthday"	"night"				
[4,]	"life"	"game"	"dinner"				
[5,]	"surprise"	"son"	"nice"				
	"talked"	"won"	"favorite	e"			
	"event"	"celebrated					
	"class"	"party"	"lunch"				
	"phone"	"gift"	"eat"				
	"passed"	"enjoyed"	"food"				
	"dad"	"team"	"cooked"				
	"mother"	"college"	"brought'				
	"share"	"met"	"ate"				
	"dont"	"attended"	"evening'	ı			
		"marriage"					
	"date"						
		"video"					
		"graduation		, c			
		"function"					
	"meet"		"prepared	1 "			
[20,]	IIICC L	CTOPE	brebare(
	2 2	4 5 6	7 0	9 10			
1	2 3	4 5 6	/ 8	9 10			

```
# Find the most salient:
parent.terms.beta <- model.parent@beta
parent.terms.beta <- scale(parent.terms.beta)
parent.topics.terms <- NULL
for(i in 1:k){
   parent.topics.terms <- rbind(parent.topics.terms, model.parent@terms[order(parent.terms.beta[i,], decreasing = TRUE)[1:10]])
}</pre>
```

Most salient terms for each topic:

Visualization: Parenthood(Y) Word Frequency + Topics

By combining the results of Topic Modelling and the most salient terms, we manually summarize the following 10 topics:

- **Entertainment**: this topic contains joyful moments and activities in a person's daily life. It may vary from person to person. The activities in this group which lead to a certain type of feeling are more likely to be relaxed and at ease. It should be noticed that family seems to be involved here.
- **Shopping**: this topic contains purchasing activities.
- **Family**: this topic relates closely to parenthood. The family activities may includes many relationships such as parenthood and brotherhood.
- **Emotions**: this topic contains expressions relating to overall well-being. It may suggests emotions or activities in this topics relating to the physical and pschological healthness of a person.
- **Jobs**: this topic contains happy moments relating to jobs or achievements. It is interesting that word such as "son" also appears in this topic, which suggests that achievements from family members, especially kids, may affect people's happiness.
- **Kids**: this topic contains activities about playing with children.
- Outdoors: this topic contains outdoor activities and visiting natural sightseeings.
- **School**: this topic contains shoool activities; for those who are parents (i.e. not attending shoool themselves), this may suggest that children's activities and performance at shoool can bring them happiness.
- **Friends**: this topic is around friendship and special events sharing with friends (i.e. weddings, birthdays). Celebrating special events usually can bring people hapiness, especially when it happens with friends and family around.
- **Couple's**: this topic includes activities of couples. Words related to cooking pop-up very often. Enjoy meals with lovers can bring people happy.

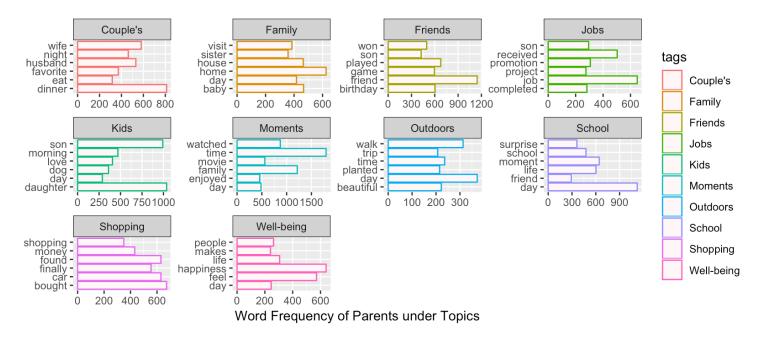
We are using <code>geom_col</code> to visualize the word frequency in each topics that we manually summarized. By visualizing this result, we can easily observe which group(s) of terms appear at most frequency as well as what key words appearing.

Hide

```
# Manually assign tags to each topics
parent.term.tags <- c("Moments", "Shopping", "Family", "Well-being", "Jobs", "Kids", "Outdo
ors", "School", "Friends", "Couple's")
# Re-grouping Parenthood(Y) Dataframe with Topics and their Probabilities
par.data$topics <- as.vector(parent.docs)
par.data$tags <- parent.term.tags [parent.docs]
colnames(parent.prob)=parent.term.tags
par.df.l=cbind(par.data, parent.prob)</pre>
```

Visualization: Hot Words Among each Topics

```
word count.mp <- par.df.1 %>%
  unnest tokens(word,text) %>%
  group by(tags) %>%
  count(word,sort=TRUE)
word_count.mp %>%
  slice(1:6) %>%
  mutate(word = reorder(word,n)) %>%
  ggplot(aes(x = word, y = n, color = tags))+
  xlab(NULL) + ylab("Word Frequency of Parents under Topics")+
  geom col(fill="snow")+
  scale fill brewer(palette = "Set3")+
  facet wrap(.~tags, scales = "free")+
  theme(
   strip.background = element rect(
     color="black", fill="lightgrey", size=0.4, linetype="solid"
     )
   ) +
  coord_flip()
```



Comment:

Obviously, many happy moments of parents are related to their children. Parents and children live together and they have time to do a lot of things together. Children, to a some extent, are part of their parents' lives. Therefore, children can be seen as a factor that can make parents happy.

Topic Modelling: Behavior Study on Parenthood(N) Group

```
# Group 2 Parenthood(N)
# Cleaning Data
corpus.no.parent <- VCorpus(VectorSource(no.par.data$text))%>%
   tm_map(content_transformer(tolower))%>%
   tm_map(removePunctuation)%>%
   tm_map(removeNumbers)%>%
   tm_map(removeWords, character(0))%>%
   tm_map(stripWhitespace)
```

10 Topics in Parenthood(N) Group (20 top words & words counts)

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Т
opic 7 [1,] "friend"	"feel"	"watched"	"dinner"	"finally"	"played"	"
time"	"momont"	"ondorrod"	"forrowito"	"started"	"found"	"
<pre>[2,] "talked" family"</pre>	"moment"	"enjoyed"	"favorite"	started	Touna	
[3,] "car" visit"	"event"	"night"	"ate"	"finished"	"game"	"
[4,] "school" trip"	"life"	"girlfriend"	"lunch"	"weeks"	"bought"	"
[5,] "met" weekend"	"people"	"movie"	"food"	"completed"	"won"	"
[6,] "fun" spend"	"makes"	"nice"	"eat"	"book"	"video"	"
[7,] "girl" planned"	"love"	"love"	"free"	"ive"	"purchased"	"
[8,] "couple" spent"	"share"	"boyfriend"	"delicious"	"class"	"buy"	"
[9,] "meet" moved"	"person"	"husband"	"restaurant"	"project"	"shopping"	"
[10,] "havent" live"	"dont"	"wife"	"cooked"	"hard"	"store"	"
[11,] "laugh" home"	"happened"	"relax"	"home"	"read"	"team"	"
[12,] "song" vacation"	"experience"	"date"	"drink"	"college"	"online"	"
[13,] "drive" house"	"special"	"amazing"	"coffee"	"exam"	"money"	"
	"world"	"beautiful"	"pizza"	"passed"	"tickets"	"
[15,] "listening"	"happiness"	"pretty"	"meal"	"successfully"	"saved"	"

```
summer"
[16,] "hours"
                   "remember"
                                  "season"
                                                "wife"
                                                              "lost"
                                                                               "dollars"
happiness"
[17,] "attended"
                   "proud"
                                  "funny"
                                                "brought"
                                                              "goal"
                                                                               "sale"
apartment"
[18,] "mine"
                   "change"
                                  "afternoon"
                                                "breakfast"
                                                                               "computer"
                                                              "gym"
city"
[19,] "close"
                   "wonderful"
                                  "episode"
                                                "ice"
                                                              "learned"
                                                                               "win"
beach"
[20,] "hadnt"
                   "bad"
                                  "planted"
                                                "cream"
                                                              "results"
                                                                               "items"
town"
      Topic 8
                 Topic 9
                               Topic 10
                                "job"
 [1,] "dog"
                 "day"
                 "birthday"
                                "received"
 [2,] "morning"
                 "surprise"
                                "money"
 [3,] "walk"
 [4,] "cat"
                 "brother"
                               "recently"
                 "mom"
                                "managed"
 [5,] "hours"
                 "sister"
 [6,] "home"
                                "graduation"
                 "celebrated"
                               "pay"
 [7,] "sleep"
 [8,] "run"
                 "phone"
                                "promotion"
                 "gift"
 [9,] "park"
                                "company"
[10,] "nice"
                 "mother"
                                "card"
[11,] "cleaned"
                 "parents"
                                "waiting"
[12,] "weather" "party"
                                "expected"
[13,] "woke"
                                "boss"
                 "house"
                 "excited"
                                "paid"
[14,] "rain"
                 "dad"
                                "called"
[15,] "taking"
[16,] "bike"
                 "called"
                                "offer"
[17,] "ride"
                                "mail"
                 "office"
[18,] "didnt"
                 "baby"
                                "accepted"
                 "smile"
                                "helped"
[19,] "puppy"
                                "interview"
[20,] "bed"
                 "told"
9229 5462 6838 6913 6102 6236 4219 5261 3702 4899
```

```
# Find the most salient:
no.parent.terms.beta <- model.no.parent@beta
no.parent.terms.beta <- scale(no.parent.terms.beta)
no.parent.topics.terms <- NULL
for(i in 1:k){
   no.parent.topics.terms <- rbind(no.parent.topics.terms, model.no.parent@terms[order (no.parent.terms.beta[i,], decreasing = TRUE)[1:10]])
}</pre>
```

Most salient terms for each topic:

Visualization: Parenthood(N) Word Frequency + Topics

By combining the results of Topic Modelling and the most salient terms, we manually summarize the following 10 topics:

- **Friends**: this topic contains key words around a person's daily life. It includes interactions and communications with firends.
- People: this topic is morea about the world and people around an individual.
- **Relationships**: this topic is mainly about relationships, especially romantic relationships.
- Foods: this topic contains a variety of food and activities around foods.
- Accomplish: this topic includes one's accomplishment.
- Entertainment: this topic contains daily entertainment activities, include shopping, gaming, playing basketball, etc.
- **Travel**: this topic contains outdoor activities and visiting natural sightseeings.
- **Lifestyle**: this topic contains lifestyles which make people feel comfortable and relaxed. Pet is a key word here.
- Celebrations: this topic is around special events or celebrations.
- **Works**: this topic is on works/shcool. It includes words that are not only emphasizing on the achievement but also covering the content of jobs.

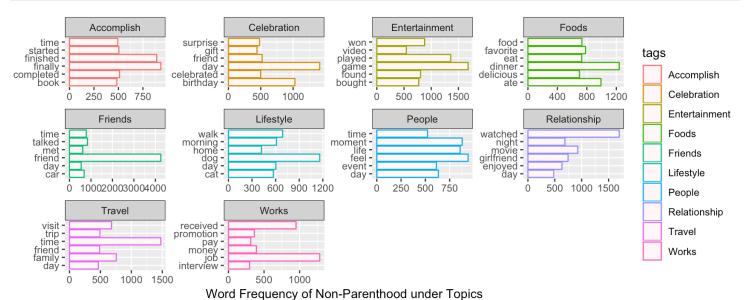
We are using <code>geom_col</code> to visualize the word frequency in each topics that we manually summarized. By visualizing this result, we can easily observe which group(s) of terms appear at most frequency as well as what key words appearing.

```
Hide
```

```
# Manually assign tags to each topics
no.parent.term.tags <- c("Friends","People","Relationship","Foods","Accomplish","Ente
rtainment","Travel","Lifestyle","Celebration","Works")
# Re-grouping Parenthood(Y) Dataframe with Topics and their Probabilities
no.par.data$topics <- as.vector(no.parent.docs)
no.par.data$tags <- no.parent.term.tags [no.parent.docs]
colnames(no.parent.prob)=no.parent.term.tags
no.par.df.l=cbind(no.par.data, no.parent.prob)</pre>
```

Visualization: Hot Words Among each Topics

```
word count.no.par <- no.par.df.1 %>%
  unnest tokens(word, text) %>%
  group_by(tags) %>%
  count(word,sort=TRUE)
word count.no.par %>%
  slice(1:6) %>%
  mutate(word = reorder(word,n)) %>%
  ggplot(aes(x = word, y = n, color = tags))+
  xlab(NULL) + ylab("Word Frequency of Non-Parenthood under Topics")+
  geom col(fill="snow")+
  scale fill brewer(palette = "Set3")+
  facet wrap(.~tags, scales = "free")+
  theme (
   strip.background = element_rect(
     color="black", fill="lightgrey", size=0.4, linetype="solid"
   ) +
  coord flip()
```



Comment:

People without children have more diverse lifestyles. Their daily activities may depend on their own personalities and lifestyles. Not raising children, people pay more attention to themselves. From word Frequency and Wordcloud, we notice that people without children are more self-centered, and unlike parents who are family-oriented.

Conclusion on Parenthood Study:

By performing visualization and topic modeling on parents and non-parents. We have the following observations:

- People with children are more likely to be together with family and children. From the <code>wordcloud</code> and <code>word Frequency under Topics</code> we can learn that activities that related to family and children take up a major part of happy moments of parents. Since <code>school</code> is also a topic there, we can guess that how children spend their days may have an impact on their parents' mood. In <code>Jobs</code> content, we also see word such as <code>son</code>, which would suggest that parents are likely to be happy because of their children's achievement.
- Even if both groups have key words that are related to food, they are demonstrating different patterns.
 People as parents are very likely married, and thus, meals and foods may be more related to family. In contrast, the food topic of non-parenthood group seems to be more food-related where more specific types of food names appear.
- People with no children demonstrate a very different lifestyle. They are more self-centered, and they are not spending as much time as parent group with family. It is interesting that, 'dog' and 'cat' are hot words in this topic, which indicating that people without children are likely to have a pet as companion.
- Works may be a major part of life for people without kids. The topic work of non-parent group suggests that people with no kids are more likely to be happy because of their jobs. Instead of taking of family, non-parents may focus more on self-promotion and accomplishment.

Step 4: Conclusion on Happy_DB Data Analysis

The sentiment analysis indicated that there is not a certain type of people who are actually happier than the most of the rest. We observed that people with children were relatively happier than those without kids.

Further, we studied the behavior of these two groups of people, Parenthood(Yes) and Parenthood(No). We observed that parents spent many of their time with family, especially with their children. The happiness related to children not only appeared in family activities, but also in outdoor activities, entertainment and jobs-related topics. Children indeed played an important part of parents' lives, and it is then very understandable that many happy moments of parents were about their family and children. On the other hand, people with no kids focused more on enjoy themselves. They might have pets as life companions, and they fulfilled themselves with works.

If we are supposed to give some explanation on why parenthood may result people happier. There are two reasons can be considered about:

- 1. Parents and children live together, and they are more likely to do things as a family.
- 2. Happiness, to some extent, is not bouncing between different joyments. Many people consider happiness as a positive mental status. Children, for many believes, bring fulfillment to parents and they make parents' lives more meaningful. Parenthood itself may not guarantee happiness, but it provides people a great opportunity to explore happiness with their beloved children.

Step 5: Reference

Hansen, Thomas. "Parenthood and Happiness: a Review of Folk Theories Versus Empirical Evidence."

Social Indicators Research, vol. 108, no. 1, 2011, pp. 29–64., doi:10.1007/s11205-011-9865-y (doi:10.1007/s11205-011-9865-y).