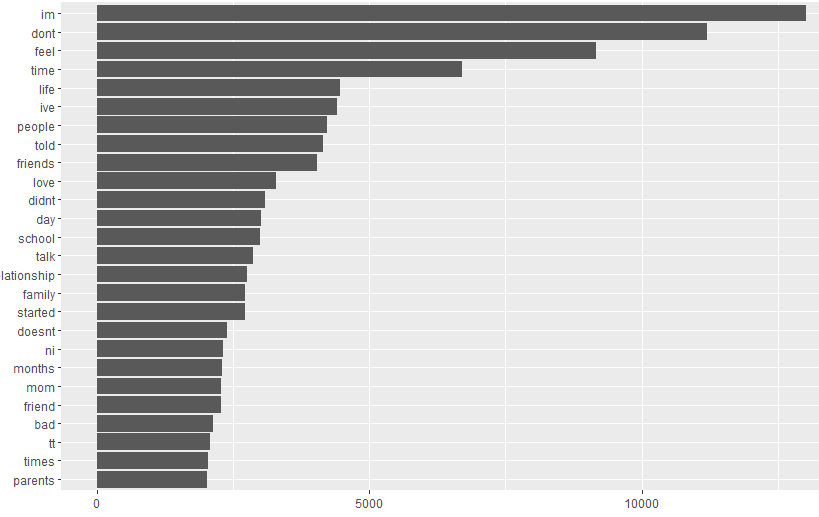
**Assignment 2: Text Analytics for National Institutes of Health (100 points)**

**Student Name:**

1. Read the file: data <- fread("psychcentral\_data.csv", sep=",", header=T, strip.white = T, na.strings = c("NA","NaN","","?"))
   1. **(1 point)** What are the column names in the data? “row”, “q\_content”, “q\_subject”, “answers”
   2. **(1 point)** How many rows does this data have? 8,360
2. Use libraries “dplyr” and “tidytext” to tokenize column q\_content. Then remove the stop-words. The count the number of tokens.
   1. **(2 points)** What are the top five tokens returned?

I’m, don’t, feel, time, life

* 1. **(2 points)** Use library “ggplot2” to create a visualization that shows the frequency of the tokens that appeared for at least 2000 times. (Hint: Change n in argument filter to 2000). Paste the visualization below:



* 1. **(2 points)** Based on the results in 4.2., would you suggest stemming on this text? Why? Bring one example from the visualization above that shows stemming should be done on this text?

Yes, I would recommend stemming on this text. Some examples that show stemming should be done on this text include don’t – didn’t – doesn’t AND time – times.

* 1. Install “SnowballC” package using install.packages("SnowballC", repos = "https://cran.r-project.org"). Use library “SnowballC” to stem q\_content using the code below:

library(SnowballC)

tidy\_text <- data %>%

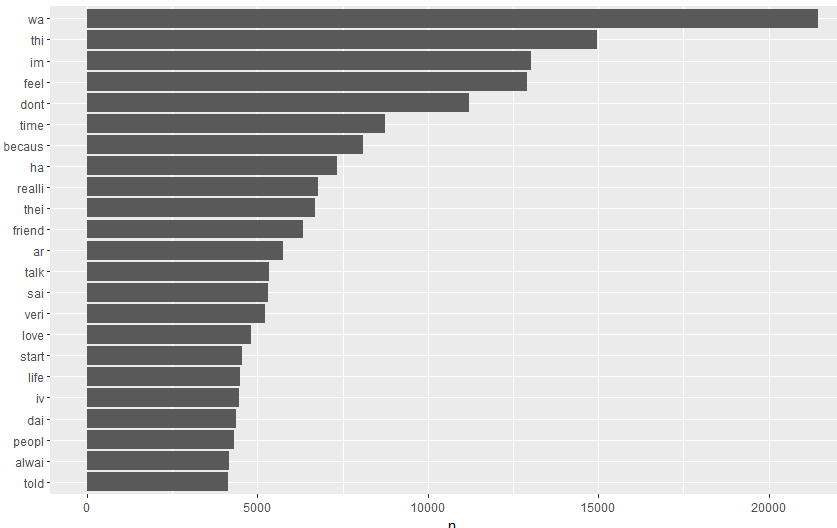
unnest\_tokens(word, q\_content) %>%

mutate(word = wordStem(word))

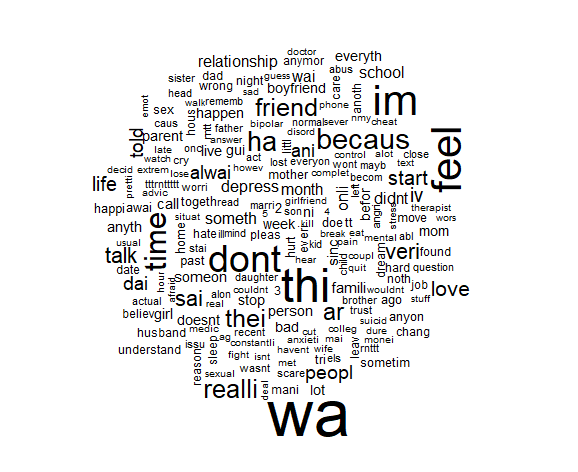
* + 1. **(2 points)** Then remove the stop-words. Now what are the top five tokens after stemming?

Wa, thi, im, feel, dont

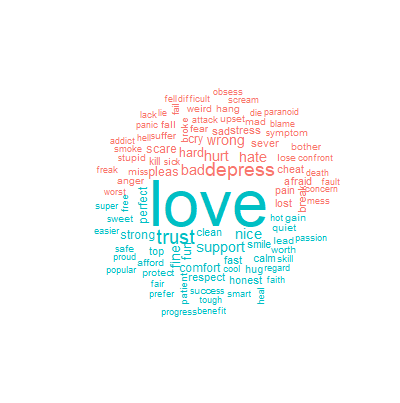
* + 1. **(2 points)** Use library “ggplot2” to create a visualization that shows the frequency of the tokens that appeared for at least 4000 times. (Hint: Change n in argument filter to 4000). Paste the visualization below:



* + 1. **(3 points)** Use library “wordcloud” to create a word cloud with the 200 most used tokens. Paste the visualization below:



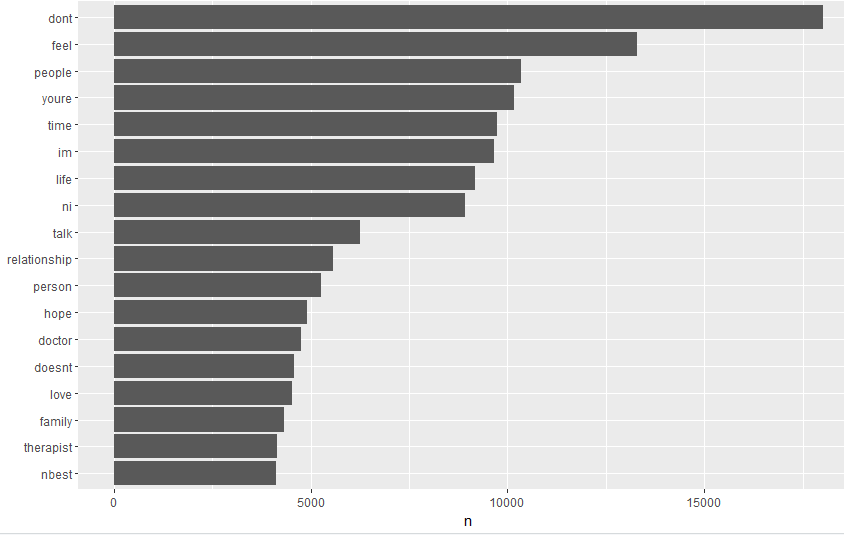
* + 1. **(5 points)** Create a color-coded word cloud based on sentiment. Use the most frequent 100 tokens for positive and negative words. Paste the word cloud in the space below:



1. Repeat all the steps in question 2 but this time for column answers.
   1. **(2 points)** What are the top five tokens returned?

dont, feel, people, youre, time

* 1. **(2 points)** Use library “ggplot2” to create a visualization that shows the frequency of the tokens that appeared for at least 4000 times. (Hint: Change n in argument filter to 4000). Paste the visualization below:



* 1. Install “SnowballC” package using install.packages("SnowballC", repos = "https://cran.r-project.org"). Use library “SnowballC” to stem answers using the code below:

library(SnowballC)

tidy\_text <- data %>%

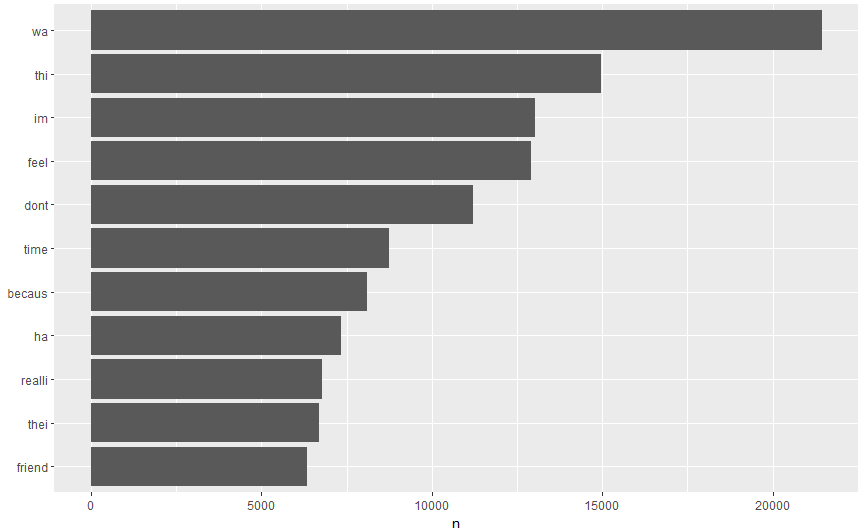
unnest\_tokens(word, answers) %>%

mutate(word = wordStem(word))

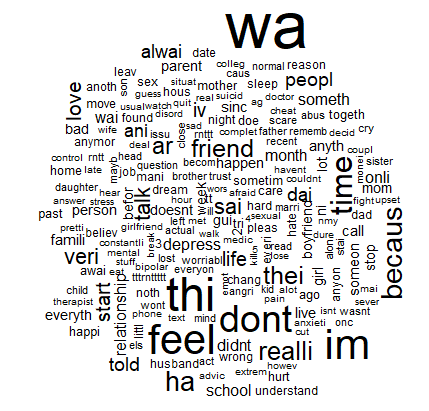
* + 1. **(2 points)** Then remove the stop-words. Now what are the top five tokens after stemming?

wa, thi, im, feel, don’t

* + 1. **(2 points)** Use library “ggplot2” to create a visualization that shows the frequency of the tokens that appeared for at least 6000 times. (Hint: Change n in argument filter to 6000). Paste the visualization below:



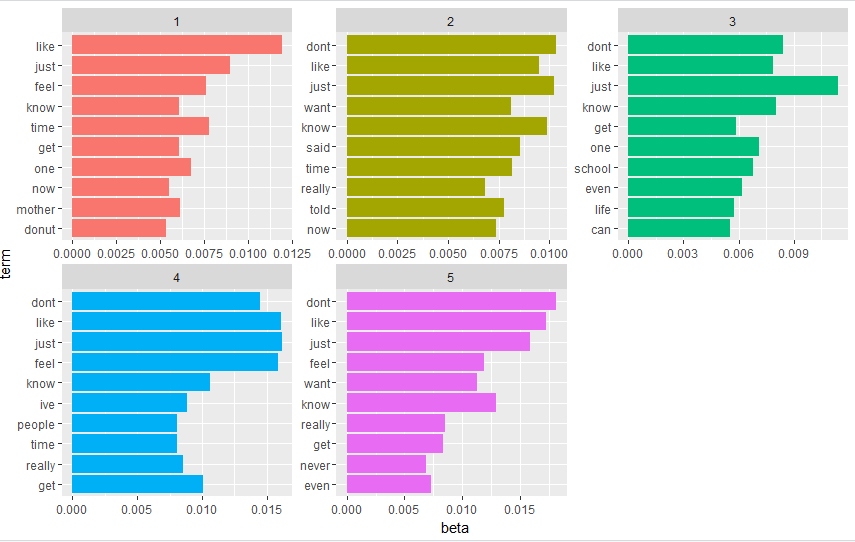
* + 1. **(6 points)** Use library “wordcloud” to create a word cloud with the 200 most used tokens. Paste the visualization below:



* + 1. **(6 points)** Create a color-coded word cloud based on sentiment. Use the most frequent 100 tokens for positive and negative words. Paste the word cloud in the space below:



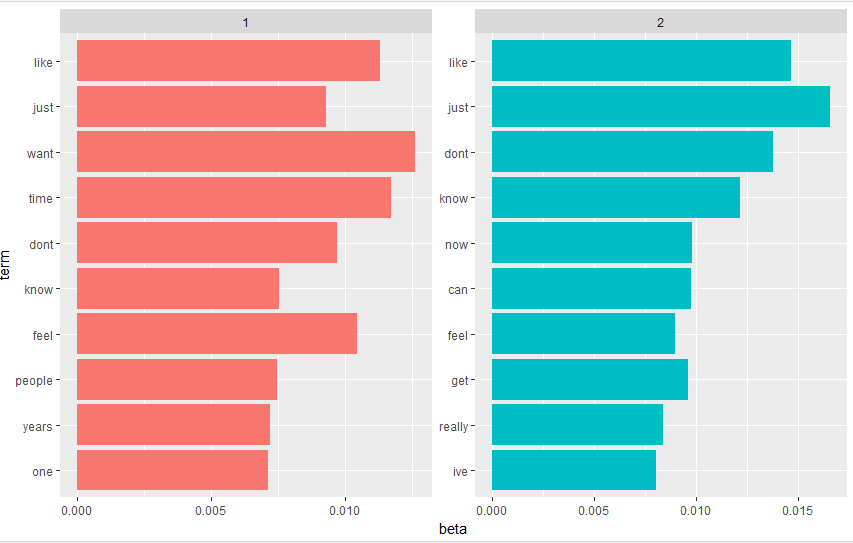
1. Use the following code to perform topic-modeling on q\_content:
   1. **(5 points)** The code above will create the beta scores for each document per topic (k = 5). Then create bar plots (similar to what we created in class) for each topic for 10 tokens (top\_n(10, beta)). Paste the visualization below.



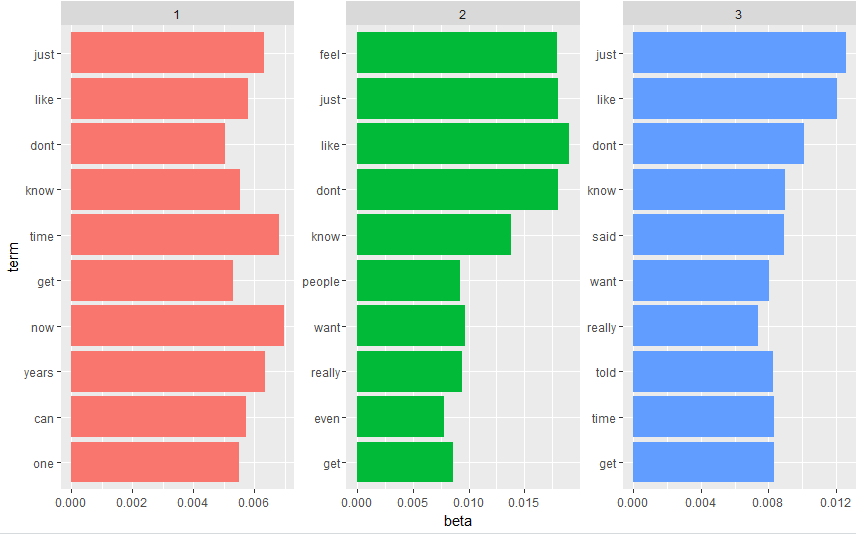
* 1. **(5 points)** Based on the visualization in 4.1., what can you say about k? Would you try a larger k or a smaller k?

Based on the visualization in 4.1, there appear to be too many ks. There is a lot of overlap in each topic. Additionally, there doesn’t seem to be a discernible difference in the terms from each topic. I would try a smaller k.

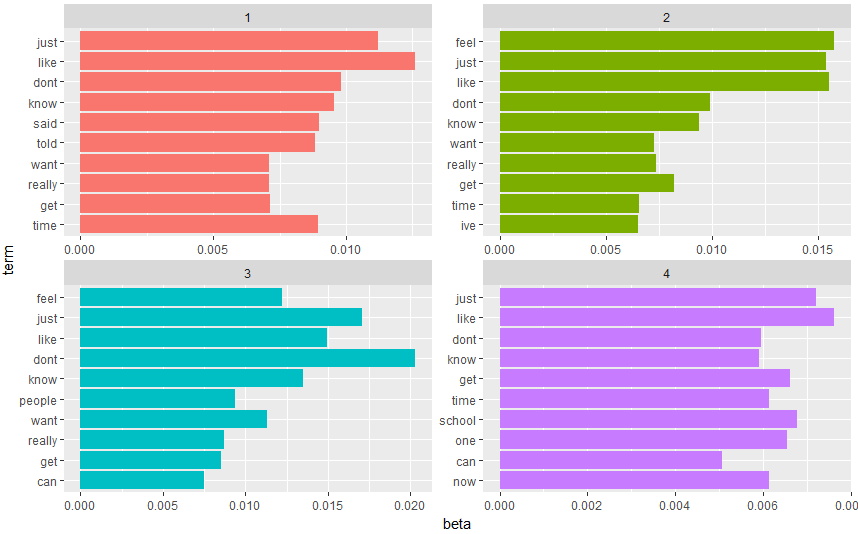
* 1. **(10 points)** Repeat 6.1. with the following ks:
     1. K = 2. Paste your visualization in the space below:



* + 1. K = 3. Paste your visualization in the space below:



* + 1. K = 4. Paste your visualization in the space below:



* + 1. K = 10. Paste your visualization in the space below:



* + 1. Based on the results recommend the number of topics that would be appropriate for this corpus.

Initially, I thought that less ks would be more helpful but as we move toward the larger ks we see more potential issues emerge. We see relationships, school, always, and mother emerge as we use >=5 ks.

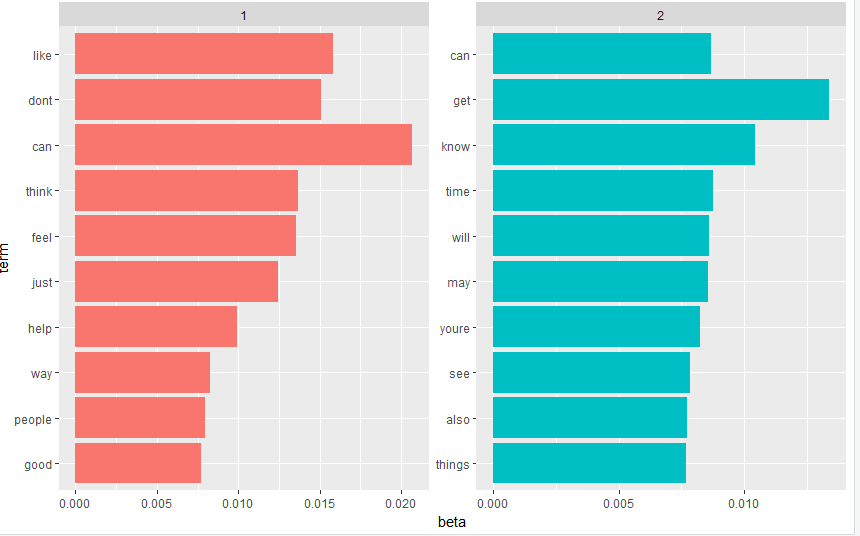
1. Use the following code to perform topic-modeling on answers:
   1. **(5 points)** The code above will create the beta scores for each document per topic (k = 10). Then create bar plots (similar to what we created in class) for each topic for 10 tokens (top\_n(10, beta)). Paste the visualization below.



* 1. **(5 points)** Based on the visualization in 6.1., are the tokens in all topics similar? Then what can you say about k? Would you try a larger k or a smaller k?

All the topics are similar. I would try a larger k to see if more words emerge with increased topics.

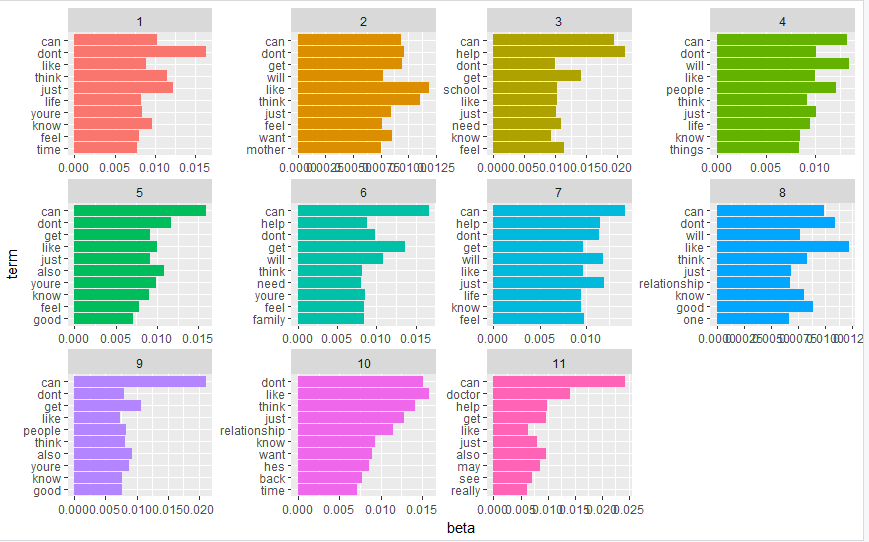
* 1. **(10 points)** Repeat 6.1. with the following ks:
     1. K = 2. Paste your visualization in the space below:



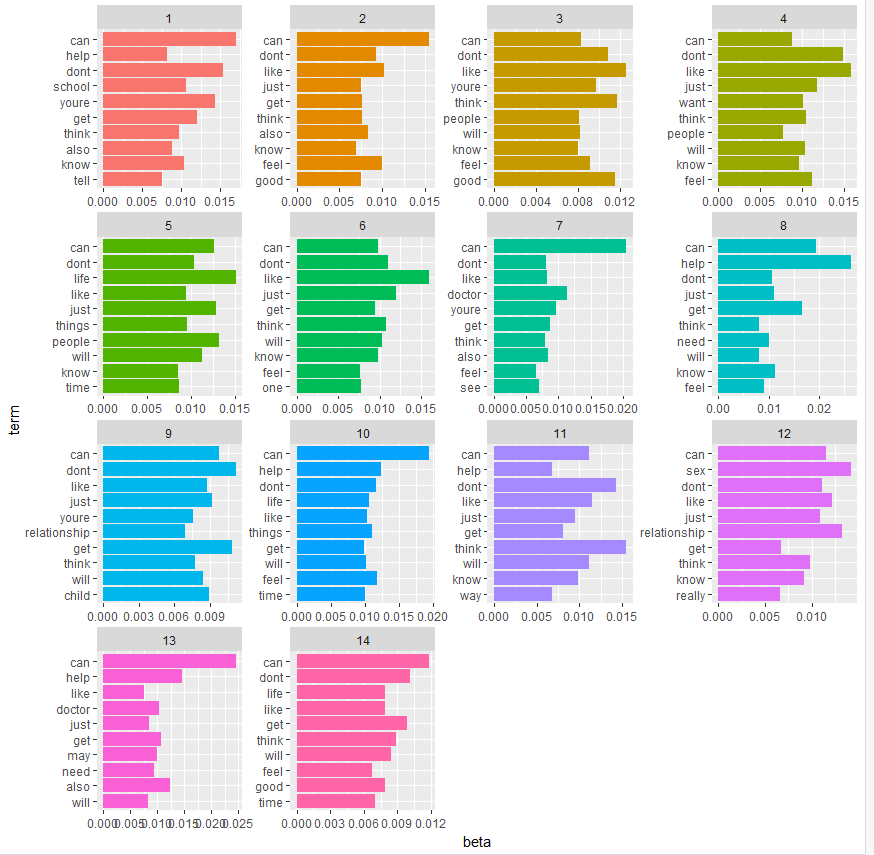
* + 1. K = 8. Paste your visualization in the space below:



* + 1. K = 11. Paste your visualization in the space below:



* + 1. K = 14. Paste your visualization in the space below:



* + 1. Based on the results recommend the number of topics that would be appropriate for this corpus.

I would recommend 11 topics. It really seemed as though certain trends started to emerge whereas 14 topics made the word groupings very unclear and kind of nonsensical.

1. **(20 points)** Suppose that you are a researcher who works for National Institutes of Health (NIH). You are working on a project that aims to identify the most important reasons for mental disorders. Based on your analysis above, can we propose any hypothesis about the reasons for mental disorders in the society? Please explain.

I would say that based on our analysis above, we can definitely hypothesize about mental health disorders and their reasons. When we analyze the text, we start to see certain trends emerge. School, parents, relationships, people, and family are common words found in the topics we created. Drilling deeper into the analysis, we see that uncertainty surrounding these topics are the cause of the some mental health disorders in society. It seems as though people are stressed about knowing whether or not their relationship is a good one. People are also stressed about the end of a relationship and wondering if their partner will come back. Uncertainty around school, parents, family, and mothers also seems to be trends that emerge.

Of course, we would like to be able to analyze more than 1,000 rows. There is a possibility that more concrete trends will emerge from this analysis. However, we need to ensure that we have the computer power to perform this analysis. It might also help to eliminate words that were found in all or most of the topics. This would allow less common words to emerge. Overall, I feel as though we can use this method to find reasons for mental health disorders in society.