Stats 141SL Final Project

Hayley Todd

```
setwd("/Users/Hayley/Downloads")
finalproj <- read.csv("Statistics Data  - Page 3.csv", na.strings = c("", "NA"))
#View(finalproj)
```

49 people responded to question 2

```
textmine <- finalproj[c(1,3)]
head(textmine)
```

```
##      ID
## 1 SEAN1
## 2 SEAN2
## 3 SEAN3
## 4 SEAN4
## 5 SEAN5
## 6 SEAN6
##                                                                       top_words
## 1                          insightful enjoyable interesting detailed beautiful
## 2                                                                         <NA>
## 3                           awareness groupwork positive disjointed applicable
## 4 easy simple positive fun relevant applicable crossdisciplinary grouporiented
## 5                      motivational interesting analytical informative helpful
## 6                                mindopening meaningful unique informative fun
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

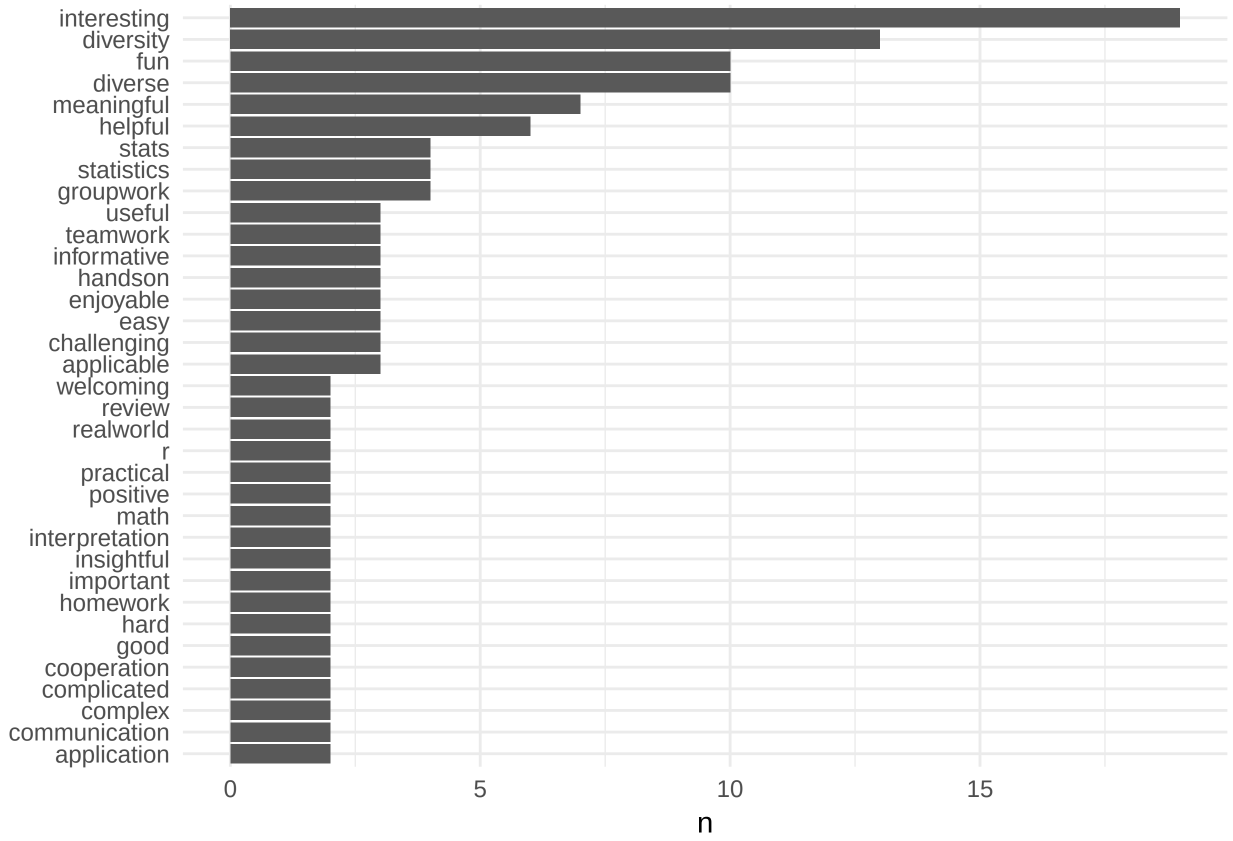
```
textmine <- textmine %>% na.omit()
textmine$top_words <- as.character(textmine$top_words)

text_df <- tibble(line = 1:length(textmine$top_words), text = textmine$top_words)
#install.packages("tidytext")
library(tidytext)

text_df <- text_df %>% unnest_tokens(word, text)
#words <- as.data.frame(text_df[,2])
#sort(unique(words[,1]))
```

```
q2words <- text_df %>% count(word, sort = T)
head(q2words)
```

```
## # A tibble: 6 x 2
##   word            n
##   <chr>       <int>
## 1 interesting    19
## 2 diversity      13
## 3 diverse        10
## 4 fun            10
## 5 meaningful      7
## 6 helpful         6
```

Plot of raw words (typos fixed and multiple words combined):

```
library(ggplot2)
library(dplyr)
library(tidyr)

#texttrial <- text_df %>%
#  count(word, sort = TRUE)

q2words %>% dplyr::filter(n > 1) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n)) +
  geom_col() +
  xlab(NULL) +
  theme_minimal() +
  coord_flip()
```



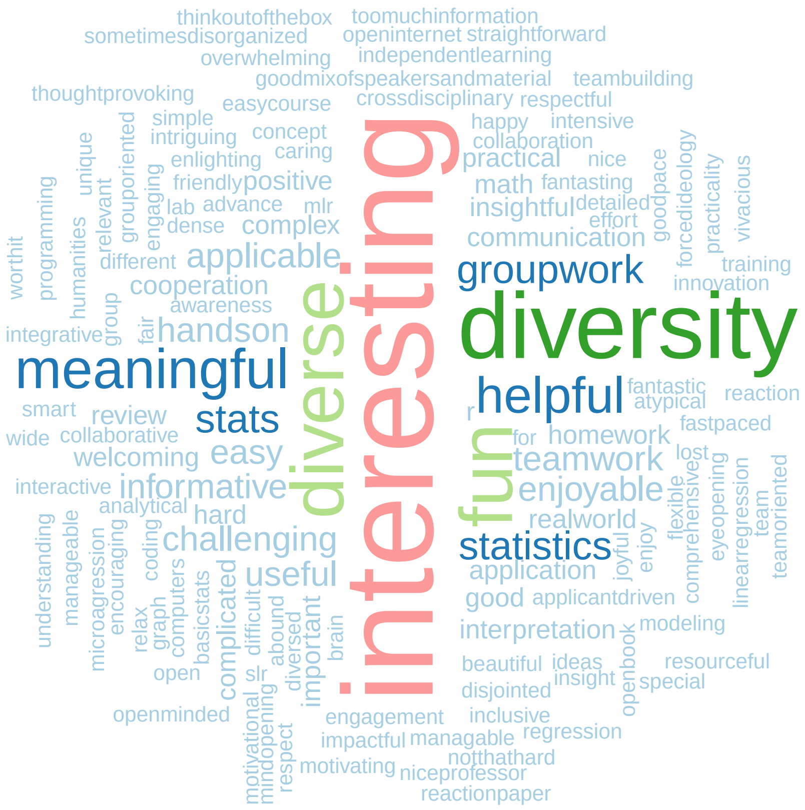
```
#text_df <- texttrial
```

Raw words word cloud:

```
#install.packages("wordcloud")
library("wordcloud")
```

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

```
set.seed(135719)
wordcloud(words = q2words$word, freq=q2words$n, min.freq = 1, max.words = 150, random.order = FALSE, ro
```



Many words had synonyms or similar themes (such as "diversity" and "diverse"), so we decided to combine those to figure out the major word themes.

Cleaning and combining synonyms/similar themes:

```
cleantext <- text_df

applicable <- c(which(cleantext[,2] == "application"), which(cleantext[,2] == "applicantdriven"), which
cleantext[applicable, 2] <- "applicable"

##########

difficult <- c(which(cleantext[,2] == "challenging"), which(cleantext[,2] == "hard"), which(cleantext[,
cleantext[difficult, 2] <- "difficult"

############
```

```r
complex <- c(which(cleantext[,2] == "comprehensive"), which(cleantext[,2] == "crossdisciplinary"))
cleantext[complex, 2] <- "complex"

###########

collaborative <- c(which(cleantext[,2] == "collaboration"), which(cleantext[,2] == "cooperation"))
cleantext[collaborative, 2] <- "collaborative"

###########

team <- c(which(cleantext[,2] == "group"), which(cleantext[,2] == "groupwork"), which(cleantext[,2] ==
cleantext[team, 2] <- "team"

###########

diversity <- c(which(cleantext[,2] == "diverse"))
cleantext[diversity, 2] <- "diversity"

###########

easy <- c(which(cleantext[,2] == "notthathard"), which(cleantext[,2] == "easycourse"), which(cleantext[
cleantext[easy, 2] <- "easy"

############

engaging <- c(which(cleantext[,2] == "engagement"))
cleantext[engaging, 2] <- "engaging"

###########

enjoyable <- c(which(cleantext[,2] == "enjoy"), which(cleantext[,2] == "fun"), which(cleantext[,2] == "
cleantext[enjoyable, 2] <- "enjoyable"

############

positive <- c(which(cleantext[,2] == "good"))
cleantext[positive, 2] <- "positive"

############

interactive <- c(which(cleantext[,2] == "handson"))
cleantext[interactive, 2] <- "interactive"

############

insightful <- c(which(cleantext[,2] == "insight"))
cleantext[insightful, 2] <- "insightful"

############

interesting <- c(which(cleantext[,2] == "intriguing"))
cleantext[interesting, 2] <- "interesting"
```

```r
###########

manageable <- c(which(cleantext[,2] == "managable"))
cleantext[manageable, 2] <- "manageable"

###########

enlightening <- c(which(cleantext[,2] == "mindopening"), which(cleantext[,2] == "thoughtprovoking"), whi
cleantext[enlightening, 2] <- "enlightening"

###########

motivational <- c(which(cleantext[,2] == "motivating"))
cleantext[motivational, 2] <- "motivational"

###########

practical <- c(which(cleantext[,2] == "practicality"))
cleantext[practical, 2] <- "practical"

###########

coding <- c(which(cleantext[,2] == "r"), which(cleantext[,2] == "programming"), which(cleantext[,2] ==
cleantext[coding, 2] <- "coding"

###########

reaction <- c(which(cleantext[,2] == "reactionpaper"))
cleantext[reaction, 2] <- "reaction"

###########

respectful <- c(which(cleantext[,2] == "respect"))
cleantext[respectful, 2] <- "respectful"

###########

special <- c(which(cleantext[,2] == "unique"))
cleantext[special, 2] <- "special"

###########

statistics <- c(which(cleantext[,2] == "stats"), which(cleantext[,2] == "basicstats"), which(cleantext[
cleantext[statistics, 2] <- "statistics"

###########

concepts <- c(which(cleantext[,2] == "linearregression"), which(cleantext[,2] == "mlr"), which(cleantex
cleantext[concepts, 2] <- "concepts"

###########
head(cleantext)
```

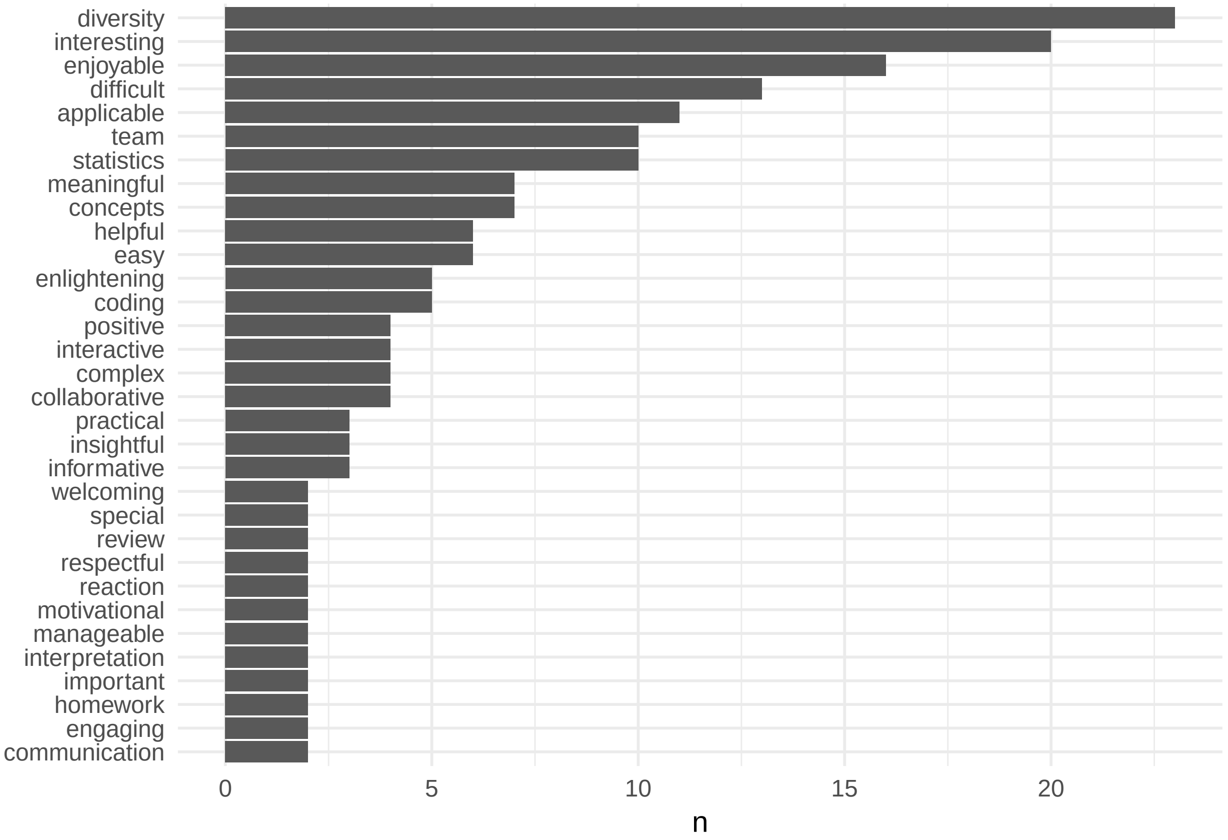
```
## # A tibble: 6 x 2
##    line word
##   <int> <chr>
## 1     1 insightful
## 2     1 enjoyable
## 3     1 interesting
## 4     1 detailed
## 5     1 beautiful
## 6     2 awareness
```

```
q2themewords <- cleantext %>% count(word, sort = TRUE)
head(q2themewords)
```

```
## # A tibble: 6 x 2
##   word            n
##   <chr>       <int>
## 1 diversity      23
## 2 interesting    20
## 3 enjoyable      16
## 4 difficult      13
## 5 applicable     11
## 6 statistics     10
```

Plot of cleaned words:

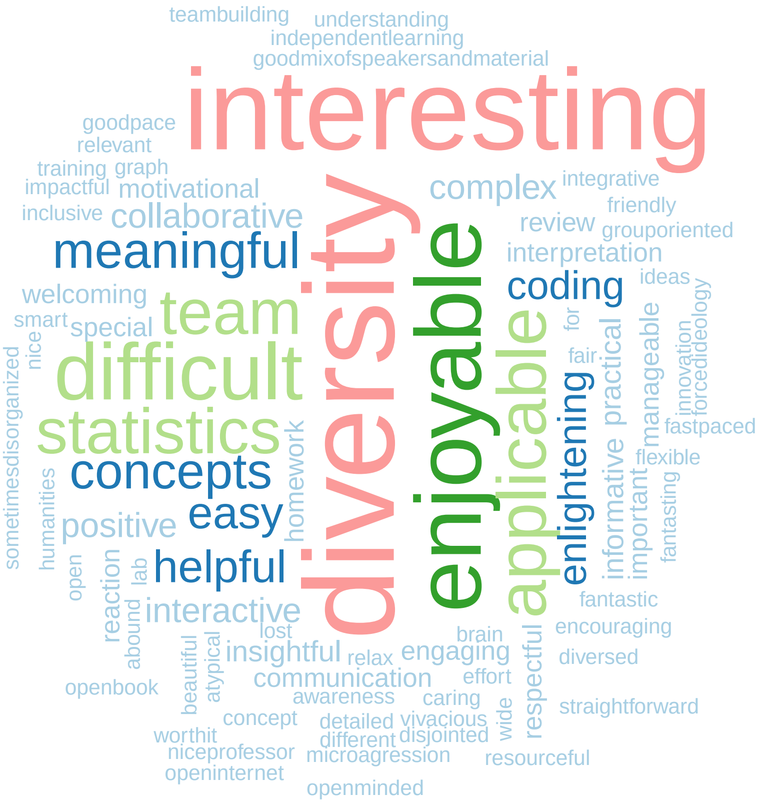
```
library(ggplot2)
library(dplyr)
library(tidyr)
#cleantext
#cleantext <- cleantext %>%
#  arrange(desc(n))

q2themewords %>% dplyr::filter(n > 1) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n)) +
  geom_col() +
  xlab(NULL) +
  theme_minimal() +
  coord_flip()
```



```
#cleantext %>%
#   count(word, sort = TRUE) %>%
#   filter(n > 1) %>%
#   mutate(word = reorder(word, n)) %>%
#   ggplot(aes(word, n)) +
#   geom_col() +
#   xlab(NULL) +
#   theme_minimal() +
#   coord_flip()
```

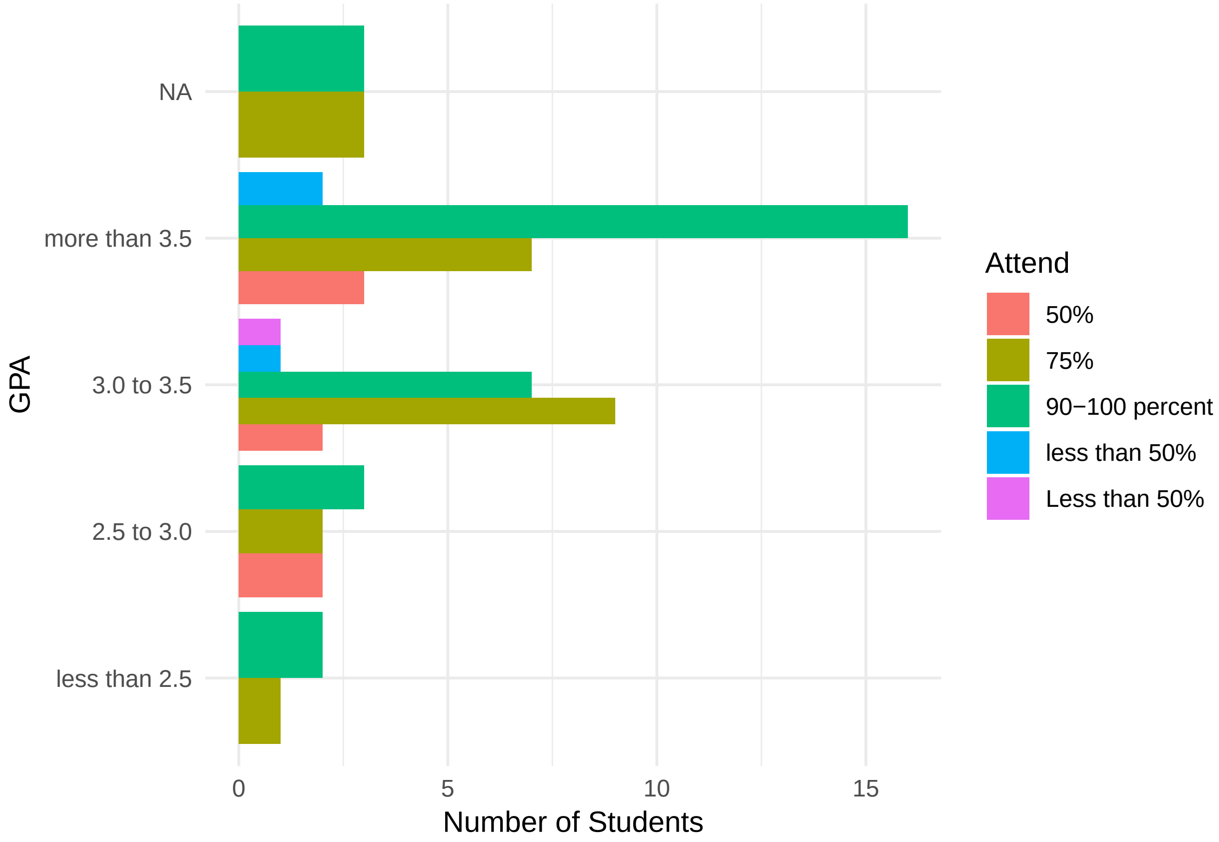
Word Cloud of Cleaned Words:

```
library("wordcloud")
set.seed(2019)
wordcloud(words = q2themewords$word, freq=q2themewords$n, min.freq = 1, max.words = 150, random.order =
```



Some exploratory analysis:

```
setwd("/Users/Hayley/Downloads")
page1 <- read.csv("Statistics Data  - Page 1.csv", na.strings = c("", "NA"))
#library(plyr)
#View(page1)

page1$GPA <- ordered(page1$GPA, levels = c("less than 2.5", "2.5 to 3.0", "3.0 to 3.5", "more than 3.5")

ggplot(page1, aes(x = GPA,fill = Attend)) + geom_bar(position = position_dodge()) + theme_minimal() + c
```



Large amount of students had a GPA of 3.5 or higher. Those who had 3.5 or higher mostly attended 90-100% of classes. People with 3.0-3.5 mostly attended 75% of class.

```
combo <- cbind(page1, finalproj[,3])
#View(combo)
```

```
cleantext <- cleantext %>% group_by(line) %>% summarise(words = paste(word, collapse=" "))
cleantext$ID <- textmine$ID
cleantext <- cleantext[,-1]
FINALDF <- full_join(combo, cleantext, by = "ID")
FINALDF <- FINALDF[,-21]
head(FINALDF)
```

```
##      ID              Major INT      Minor Trans Gender           GPA
## 1 SEAN1               Math  No       <NA>   Yes   Male    3.0 to 3.5
## 2 SEAN2                FAM Yes       <NA>   Yes Female    3.0 to 3.5
## 3 SEAN3         Statistics  No       <NA>   Yes   Male    2.5 to 3.0
## 4 SEAN4         Statistics  No       <NA>   Yes Female    3.0 to 3.5
## 5 SEAN5         Statistics Yes       <NA>   Yes Female more than 3.5
## 6 SEAN6 Business Economics  No Accounting    No Female more than 3.5
##           Attend English Recommend micro.agg understanding guest_speakers
## 1            75%      No       Yes         5             4              5
## 2 90-100 percent      No       Yes         4             4              4
## 3 90-100 percent     Yes       Yes         4             3              5
## 4  less than 50%     Yes       Yes         5             5              5
## 5            75%      No       Yes         4             4              4
## 6 90-100 percent      No       Yes         5             5              5
##   reaction_papers stats_lec stats_hw diverse_campus1 diverse_world1
## 1               4         5        5               5              5
## 2               4         5        5               5              5
## 3               5         2        2               4              4
## 4               5         5        5               5              5
## 5               4         4        4               4              4
## 6               5         5        5               5              5
##   diverse_campus2 diverse_world2
## 1               5              5
## 2               5              5
## 3               3              4
## 4               2              2
## 5               4              4
## 6               5              5
##                                                                      words
## 1                     insightful enjoyable interesting detailed beautiful
## 2                                                                     <NA>
## 3                           awareness team positive disjointed applicable
## 4 easy easy positive enjoyable relevant applicable complex grouporiented
## 5                 motivational interesting statistics informative helpful
## 6                   enlightening meaningful special informative enjoyable
```

```
summary(FINALDF)
```

```
##         ID                    Major     INT                Minor
##  BREANNA1: 1   Statistics        :25   No :31   Statistics    : 7
##  BREANNA2: 1   Sociology         : 6   Yes:33   Music Industry: 2
##  BREANNA3: 1   Math              : 5                          : 1
##  BREANNA4: 1   Applied Math      : 4            Accounting    : 1
##  BREANNA5: 1   Business Economics: 2            Accounting    : 1
##  BREANNA6: 1   FAM               : 2            (Other)       :11
##  (Other) :58   (Other)           :20            NA's          :41
##  Trans       Gender                 GPA                 Attend     English
##  No :26   Female:35   less than 2.5: 3   50%           : 7   No    :43
##  Yes:38   Male  :29   2.5 to 3.0   : 7   75%           :22   No    : 4
##                       3.0 to 3.5   :20   90-100 percent:31   Yes   :15
##                       more than 3.5:28   less than 50% : 3   Yes   : 1
##                       NA's         : 6   Less than 50% : 1   Yes/No: 1
##
##
##  Recommend   micro.agg     understanding   guest_speakers  reaction_papers
##  No : 4    Min.   :1.000   Min.   :2.000   Min.   :2.000   Min.   :2.000
##  No : 1    1st Qu.:4.000   1st Qu.:3.000   1st Qu.:4.000   1st Qu.:3.000
##  Yes:59    Median :4.000   Median :4.000   Median :4.000   Median :4.000
##            Mean   :4.047   Mean   :3.891   Mean   :4.078   Mean   :3.547
##            3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:4.000
##            Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
##
##    stats_lec        stats_hw     diverse_campus1 diverse_world1
##  Min.   :2.000   Min.   :2.000   Min.   :2.000   Min.   :2.000
##  1st Qu.:4.000   1st Qu.:3.000   1st Qu.:3.750   1st Qu.:3.250
##  Median :4.000   Median :4.000   Median :4.000   Median :4.000
##  Mean   :4.016   Mean   :3.969   Mean   :3.953   Mean   :3.871
##  3rd Qu.:4.250   3rd Qu.:5.000   3rd Qu.:4.250   3rd Qu.:4.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
##                                                 NA's   :2
##  diverse_campus2 diverse_world2     words
##  Min.   :1.0     Min.   :2.000   Length:64
##  1st Qu.:3.0     1st Qu.:3.000   Class :character
##  Median :4.0     Median :4.000   Mode  :character
##  Mean   :3.5     Mean   :3.594
##  3rd Qu.:4.0     3rd Qu.:4.000
##  Max.   :5.0     Max.   :5.000
##
```

We wanted to look into the various characteristics and see if different groups expressed different opinions or concerns about the class:

** INTERNATIONAL/NON INTERNATIONAL STUDENTS: **

24 International responses, 25 Non International responses –> even though fewer non-international people in the class, more responded to the question.

International Student Responses (raw):

```r
comboint <- combo[which(combo$INT == "Yes"),]
intwords <- comboint[,21]

combonotint <- combo[which(combo$INT == "No"),]
notintwords <- combonotint[,21]

sum(is.na(comboint[,21]) == F)
```

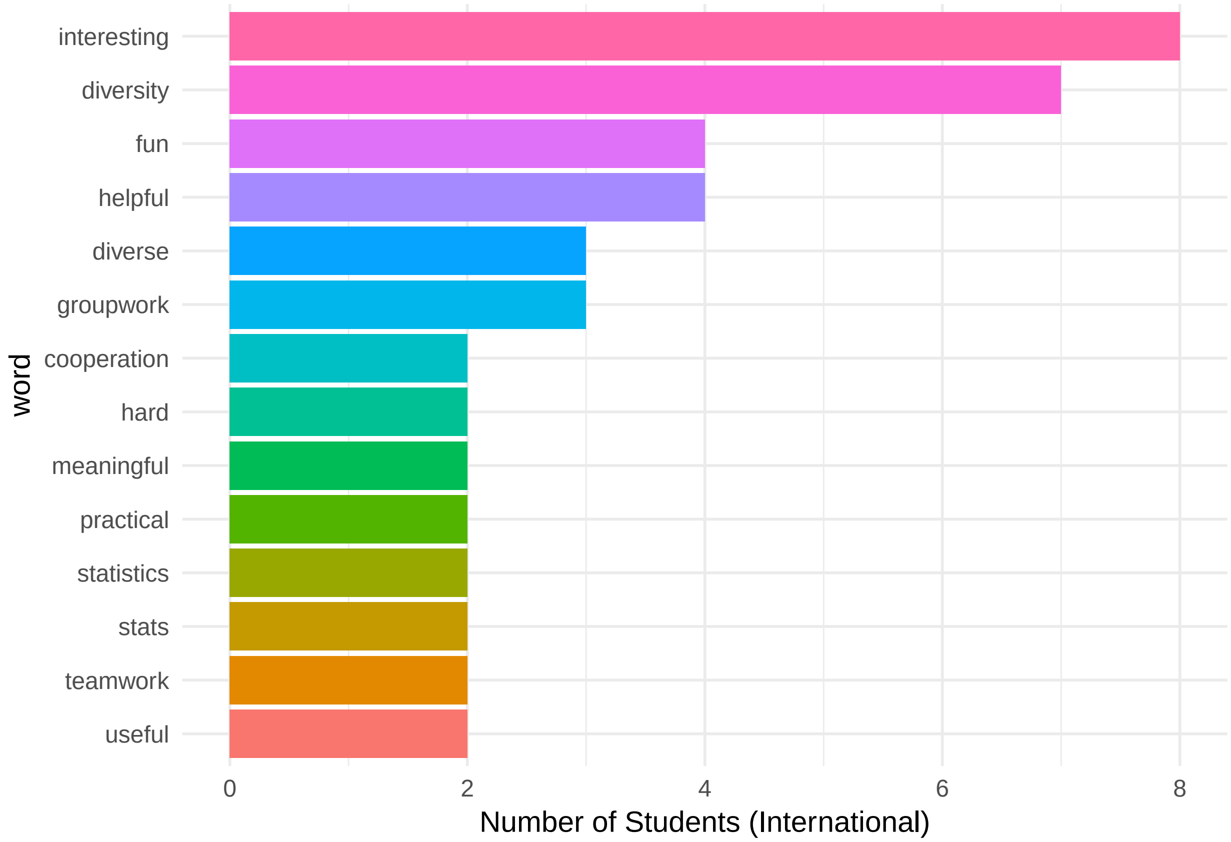
```
## [1] 22
```

```r
sum(is.na(combonotint[,21]) == F)
```

```
## [1] 27
```

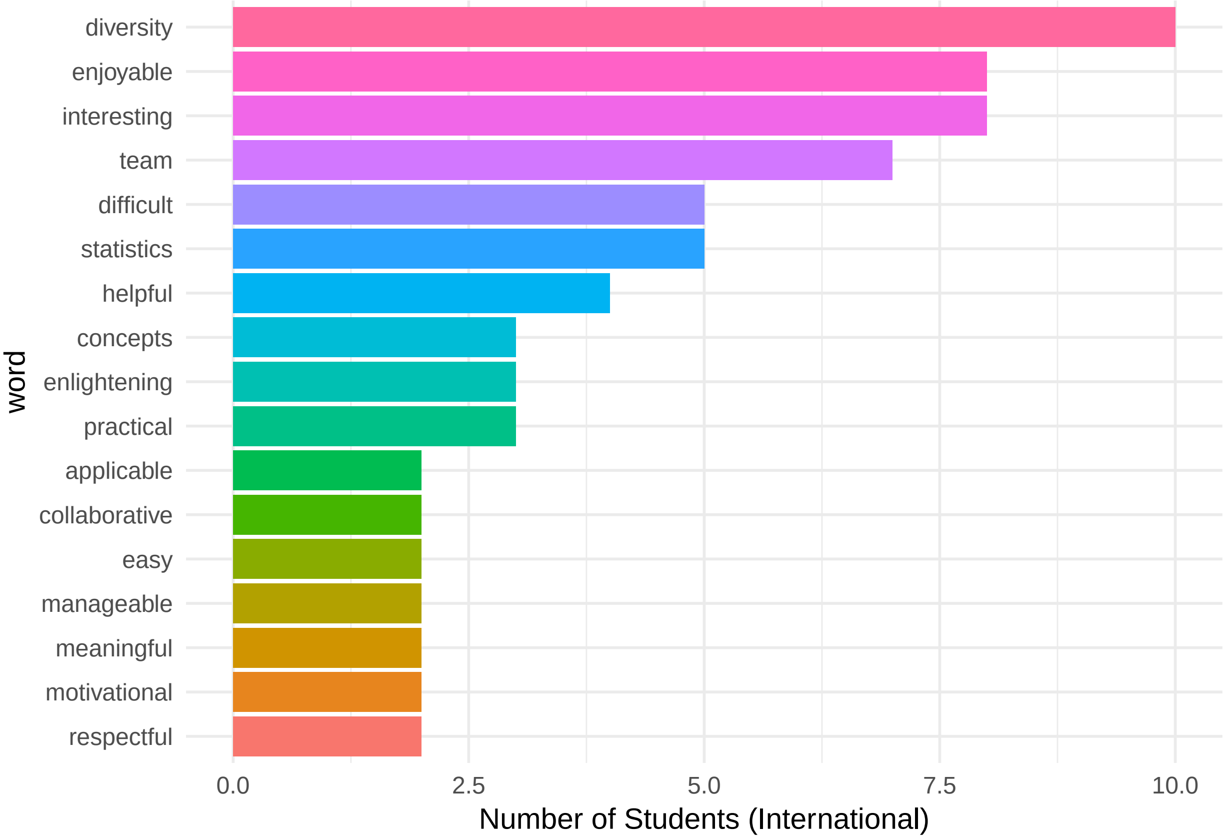
```r
library(dplyr)

intwords <- intwords %>% na.omit()
intwords <- as.character(intwords)

intwords <- tibble(line = 1:length(intwords), text = intwords)

library(tidytext)

intwords <- intwords %>% unnest_tokens(word, text)
intwordsdf <- as.data.frame(intwords[,2])
#intwordsdf
library(ggplot2)


q2intwords <- intwordsdf %>%
  count(word, sort = TRUE) %>%
  dplyr::filter(n > 1)

#q2intwords

q2intwords$word <- factor(q2intwords$word, levels = rev(factor(q2intwords$word)))
```

```
ggplot(q2intwords, aes(x = word, y = n, fill = word)) + geom_bar(stat = "identity") + coord_flip() + the
```



International Student Responses (clean):

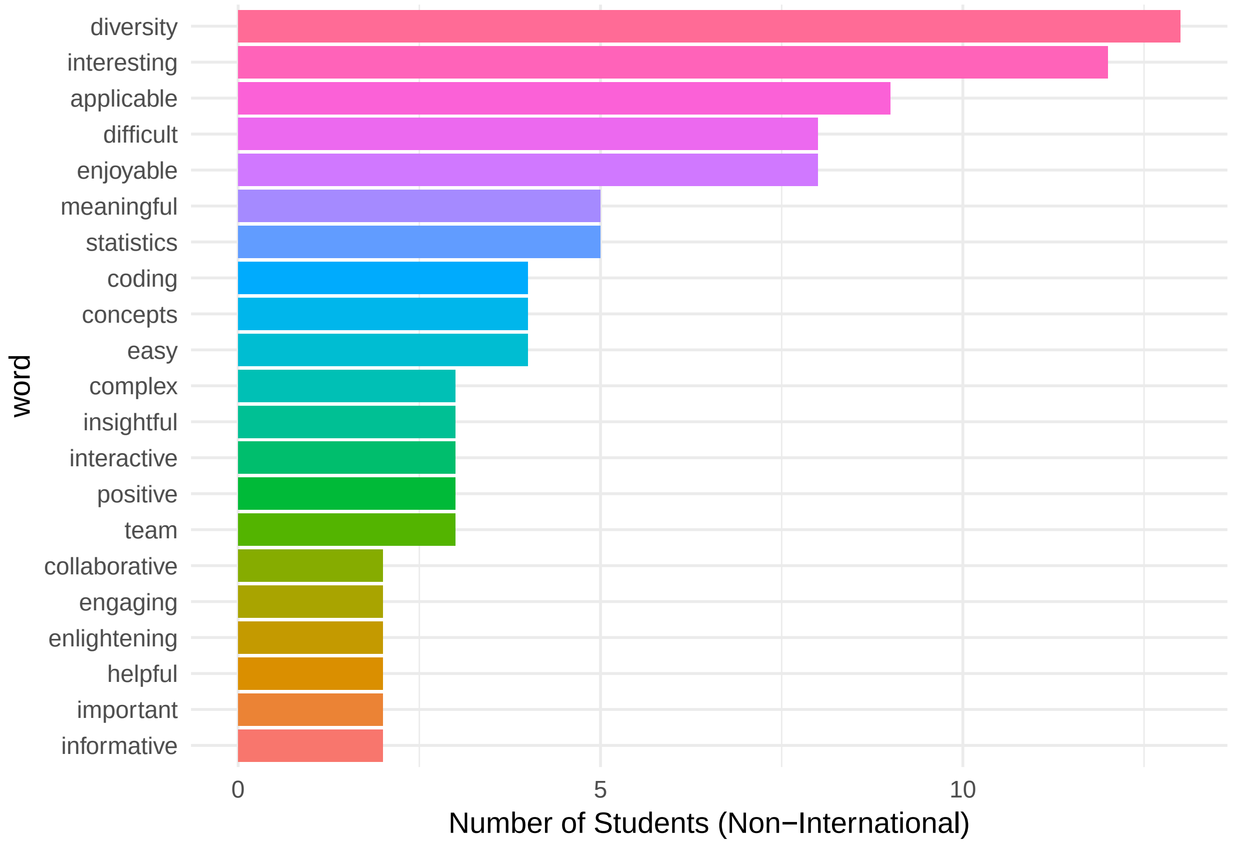
```
cleanint <- FINALDF[which(FINALDF$INT == "Yes"),]
cleanintwords <- cleanint[,21]

cleannotint <- FINALDF[which(FINALDF$INT == "No"),]
cleannotintwords <- cleannotint[,21]

library(dplyr)

cleanintwords <- cleanintwords %>% na.omit()
cleanintwords <- as.character(cleanintwords)

cleanintwords <- tibble(line = 1:length(cleanintwords), text = cleanintwords)

library(tidytext)

cleanintwords <- cleanintwords %>% unnest_tokens(word, text)
cleanintwordsdf <- as.data.frame(cleanintwords[,2])
#cleanintwordsdf
library(ggplot2)

q2cleanintwords <- cleanintwordsdf %>%
  count(word, sort = TRUE) %>%
  dplyr::filter(n > 1)
#q2cleanintwords

q2cleanintwords$word <- factor(q2cleanintwords$word, levels = rev(factor(q2cleanintwords$word)))

ggplot(q2cleanintwords, aes(x = word, y = n, fill = word)) + geom_bar(stat = "identity") + coord_flip()
```



Non-International Student Responses (clean):

```
library(dplyr)

cleannotintwords <- cleannotintwords %>% na.omit()
cleannotintwords <- as.character(cleannotintwords)

cleannotintwords <- tibble(line = 1:length(cleannotintwords), text = cleannotintwords)

library(tidytext)

cleannotintwords <- cleannotintwords %>% unnest_tokens(word, text)
cleannotintwordsdf <- as.data.frame(cleannotintwords[,2])

library(ggplot2)

q2cleannotintwords <- cleannotintwordsdf %>%
  count(word, sort = TRUE) %>%
  dplyr::filter(n > 1)
```

```
#q2cleanintwords

q2cleannotintwords$word <- factor(q2cleannotintwords$word, levels = rev(factor(q2cleannotintwords$word)

ggplot(q2cleannotintwords, aes(x = word, y = n, fill = word)) + geom_bar(stat = "identity") + coord_fli
```



Top 5 themes for international students: diversity, interesting, enjoyable, team, difficult.

Top 5 themes for non international students: diversity, interesting, applicable, enjoyable, difficult.

Not very different.

Separating by GPA:

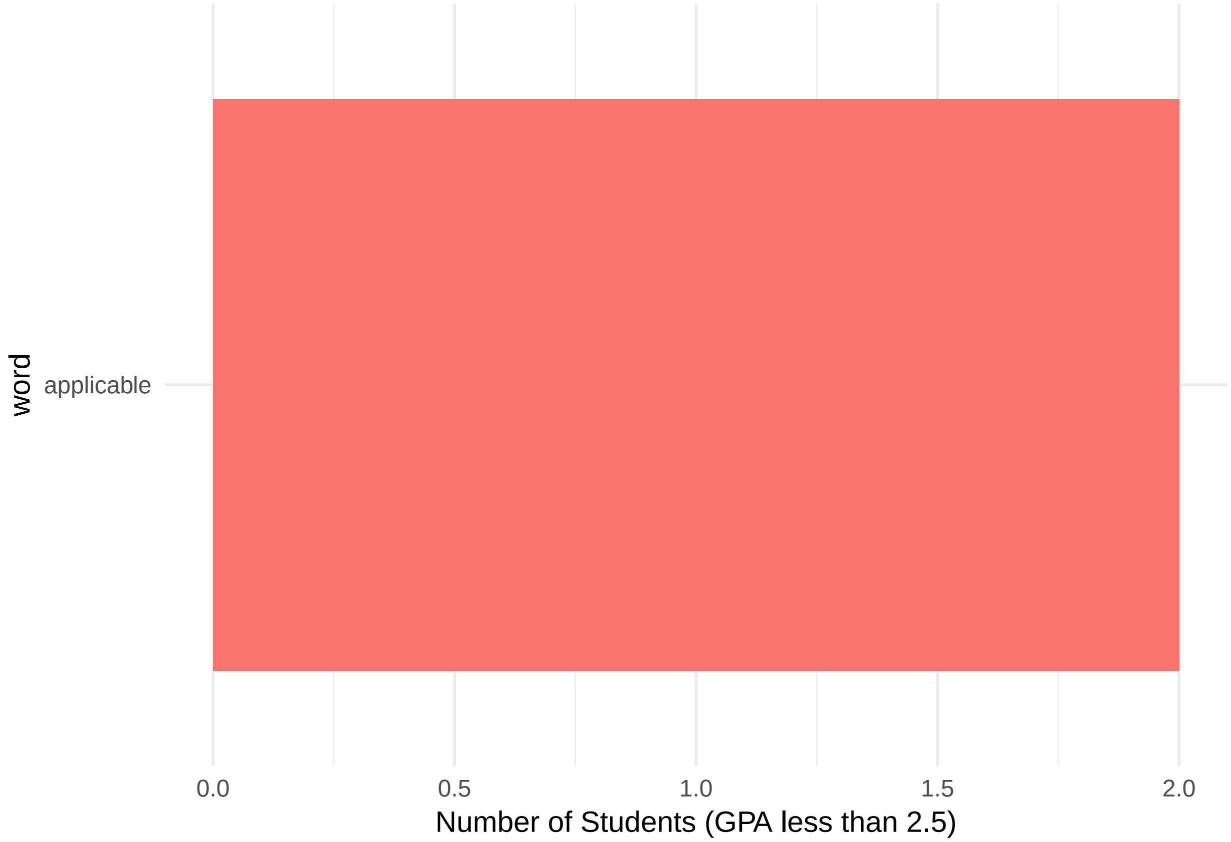
```
gpa1 <- FINALDF[which(FINALDF$GPA == "less than 2.5"),]
gpa1words <- gpa1[,21]

gpa2 <- FINALDF[which(FINALDF$GPA == "2.5 to 3.0"),]
gpa2words <- gpa2[,21]

gpa3 <- FINALDF[which(FINALDF$GPA == "3.0 to 3.5"),]
gpa3words <- gpa3[,21]

gpa4 <- FINALDF[which(FINALDF$GPA == "more than 3.5"),]
gpa4words <- gpa4[,21]
```

GPA below 2.5:

```
library(dplyr)

gpa1words <- gpa1words %>% na.omit()
gpa1words <- as.character(gpa1words)

gpa1words <- tibble(line = 1:length(gpa1words), text = gpa1words)

library(tidytext)

gpa1words <- gpa1words %>% unnest_tokens(word, text)
gpa1words
```

```
## # A tibble: 10 x 2
##     line word
##    <int> <chr>
##  1     1 applicable
##  2     1 applicable
##  3     1 interactive
##  4     1 training
##  5     1 forcedideology
##  6     2 diversity
##  7     2 enjoyable
##  8     2 interesting
##  9     2 helpful
## 10     2 collaborative
```

```
gpa1wordsdf <- as.data.frame(gpa1words[,2])
library(ggplot2)

q2gpa1words <- gpa1wordsdf %>%
  count(word, sort = TRUE) %>%
  dplyr::filter(n > 1)

q2gpa1words$word <- factor(q2gpa1words$word, levels = rev(factor(q2gpa1words$word)))

ggplot(q2gpa1words, aes(x = word, y = n, fill = word)) + geom_bar(stat = "identity") + coord_flip() + th
```



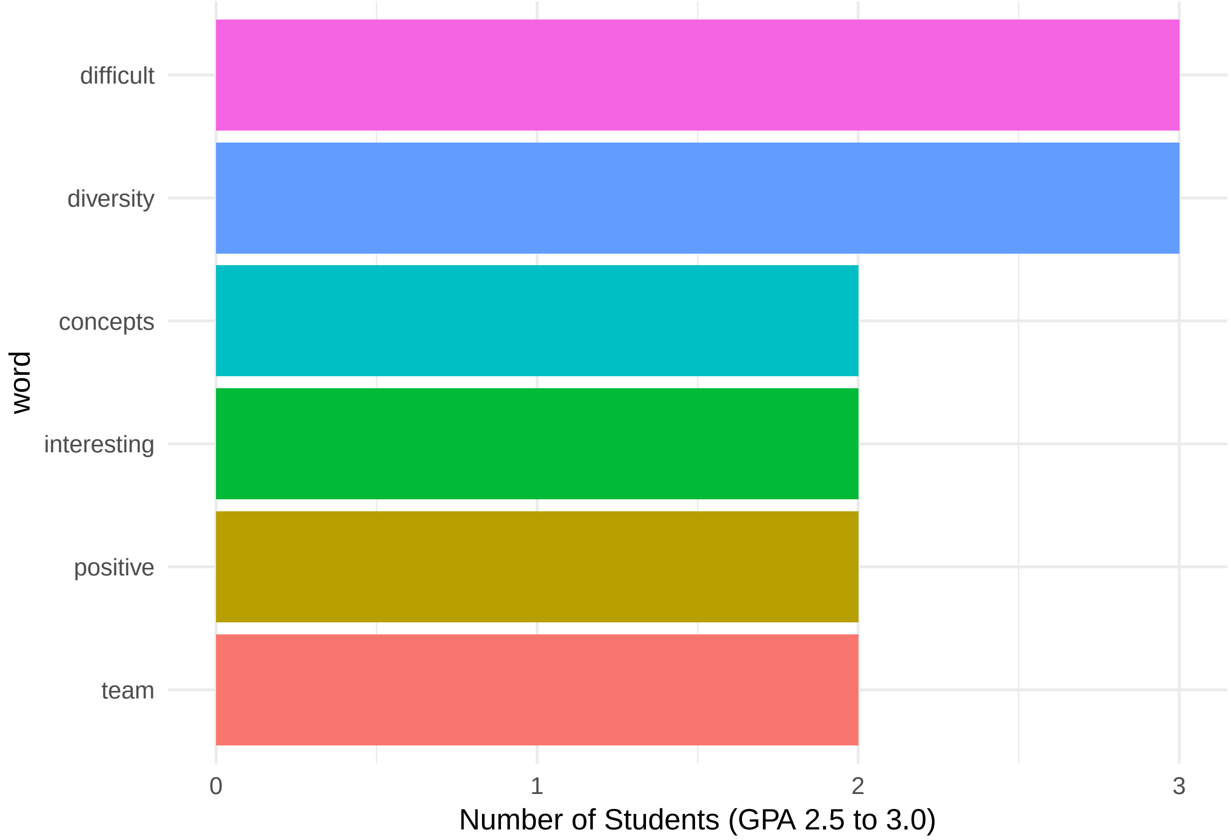
Not many people have GPA below 2.5.

GPA between 2.5 and 3.0:

```
library(dplyr)

gpa2words <- gpa2words %>% na.omit()
gpa2words <- as.character(gpa2words)

gpa2words <- tibble(line = 1:length(gpa2words), text = gpa2words)

library(tidytext)

gpa2words <- gpa2words %>% unnest_tokens(word, text)
gpa2words
```

```
## # A tibble: 25 x 2
##     line word
##    <int> <chr>
##  1     1 awareness
##  2     1 team
##  3     1 positive
##  4     1 disjointed
##  5     1 applicable
##  6     2 enjoyable
##  7     2 brain
##  8     2 concepts
##  9     2 difficult
## 10     2 positive
## # ... with 15 more rows
```

```
gpa2wordsdf <- as.data.frame(gpa2words[,2])
library(ggplot2)

q2gpa2words <- gpa2wordsdf %>%
  count(word, sort = TRUE) %>%
  dplyr::filter(n > 1)

q2gpa2words$word <- factor(q2gpa2words$word, levels = rev(factor(q2gpa2words$word)))

ggplot(q2gpa2words, aes(x = word, y = n, fill = word)) + geom_bar(stat = "identity") + coord_flip() + th
```



Also not many people have GPA between 2.5 to 3.0. However, out of those people, most people said “diversity”, “difficult”, and “interesting”.

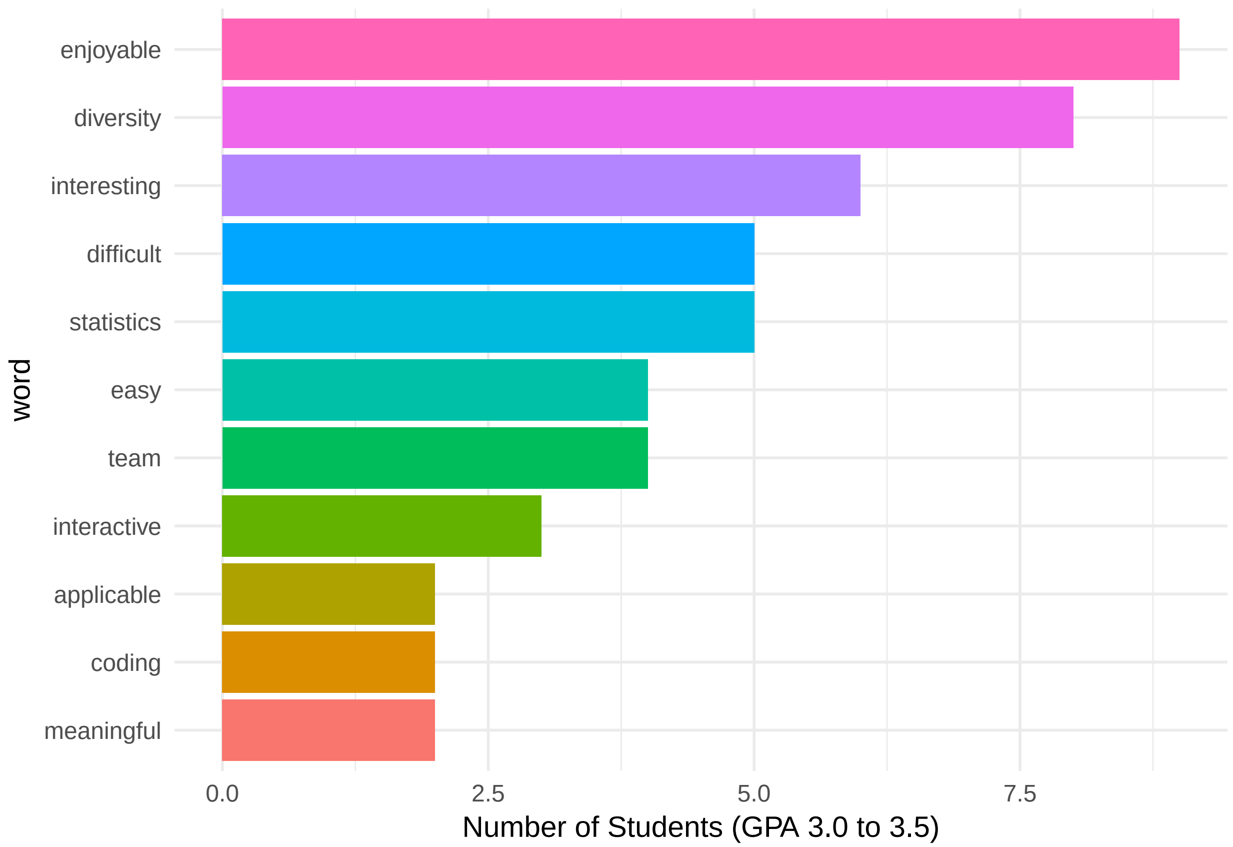
GPA between 3.0 and 3.5:

```
library(dplyr)

gpa3words <- gpa3words %>% na.omit()
gpa3words <- as.character(gpa3words)

gpa3words <- tibble(line = 1:length(gpa3words), text = gpa3words)

library(tidytext)
```

```
gpa3words <- gpa3words %>% unnest_tokens(word, text)
gpa3words
```

```
## # A tibble: 76 x 2
##      line word
##     <int> <chr>
##  1      1 insightful
##  2      1 enjoyable
##  3      1 interesting
##  4      1 detailed
##  5      1 beautiful
##  6      2 easy
##  7      2 easy
##  8      2 positive
##  9      2 enjoyable
## 10      2 relevant
## # ... with 66 more rows
```

```
gpa3wordsdf <- as.data.frame(gpa3words[,2])
library(ggplot2)

q2gpa3words <- gpa3wordsdf %>%
  count(word, sort = TRUE) %>%
  dplyr::filter(n > 1)

q2gpa3words$word <- factor(q2gpa3words$word, levels = rev(factor(q2gpa3words$word)))

ggplot(q2gpa3words, aes(x = word, y = n, fill = word)) + geom_bar(stat = "identity") + coord_flip() + t
```



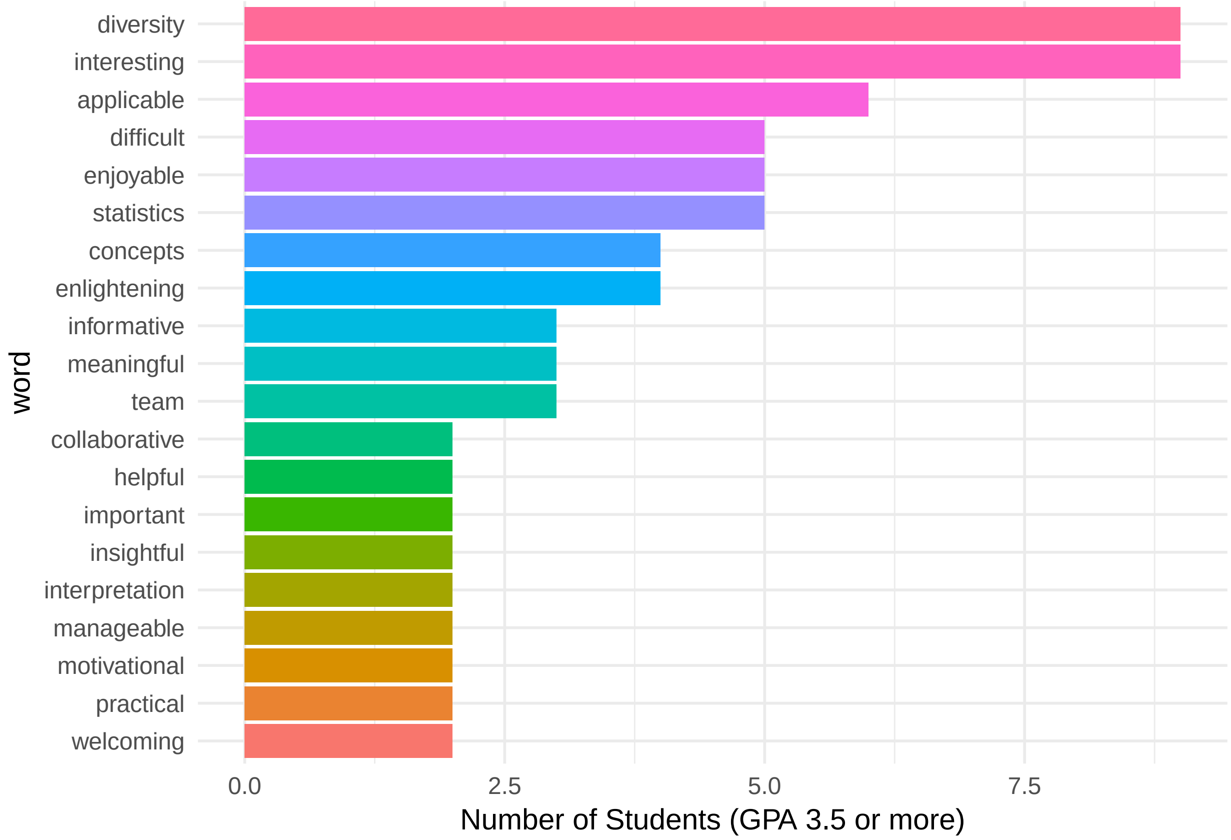
Most use “diversity”, “enjoyable”, and “interesting”.

GPA about 3.5:

```
library(dplyr)

gpa4words <- gpa4words %>% na.omit()
gpa4words <- as.character(gpa4words)

gpa4words <- tibble(line = 1:length(gpa4words), text = gpa4words)

library(tidytext)

gpa4words <- gpa4words %>% unnest_tokens(word, text)
gpa4words
```

```
## # A tibble: 105 x 2
##     line word
##    <int> <chr>
##  1     1 motivational
##  2     1 interesting
##  3     1 statistics
##  4     1 informative
##  5     1 helpful
##  6     2 enlightening
##  7     2 meaningful
##  8     2 special
## 9      2 informative
## 10     2 enjoyable
## # ... with 95 more rows
```

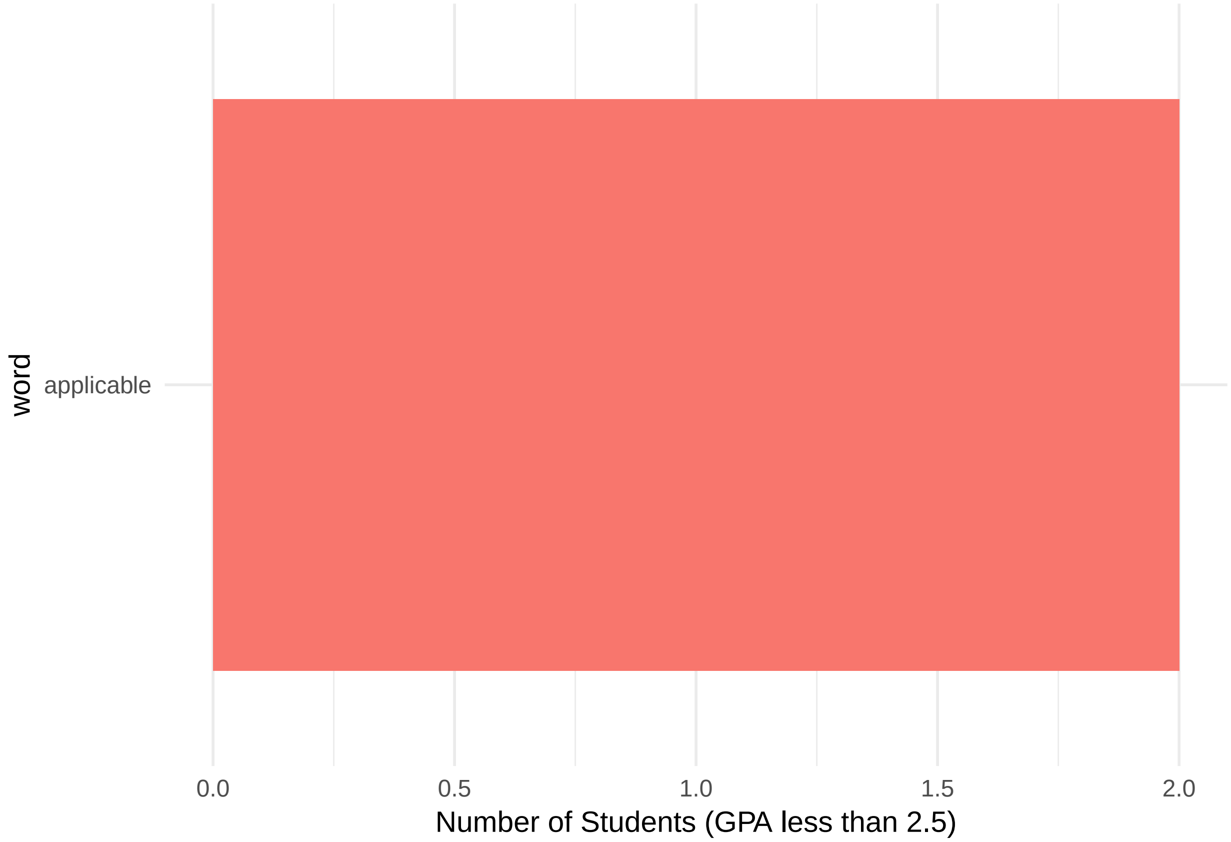
```
gpa4wordsdf <- as.data.frame(gpa4words[,2])
library(ggplot2)

q2gpa4words <- gpa4wordsdf %>%
  count(word, sort = TRUE) %>%
  dplyr::filter(n > 1)

q2gpa4words$word <- factor(q2gpa4words$word, levels = rev(factor(q2gpa4words$word)))

ggplot(q2gpa4words, aes(x = word, y = n, fill = word)) + geom_bar(stat = "identity") + coord_flip() + tl
```



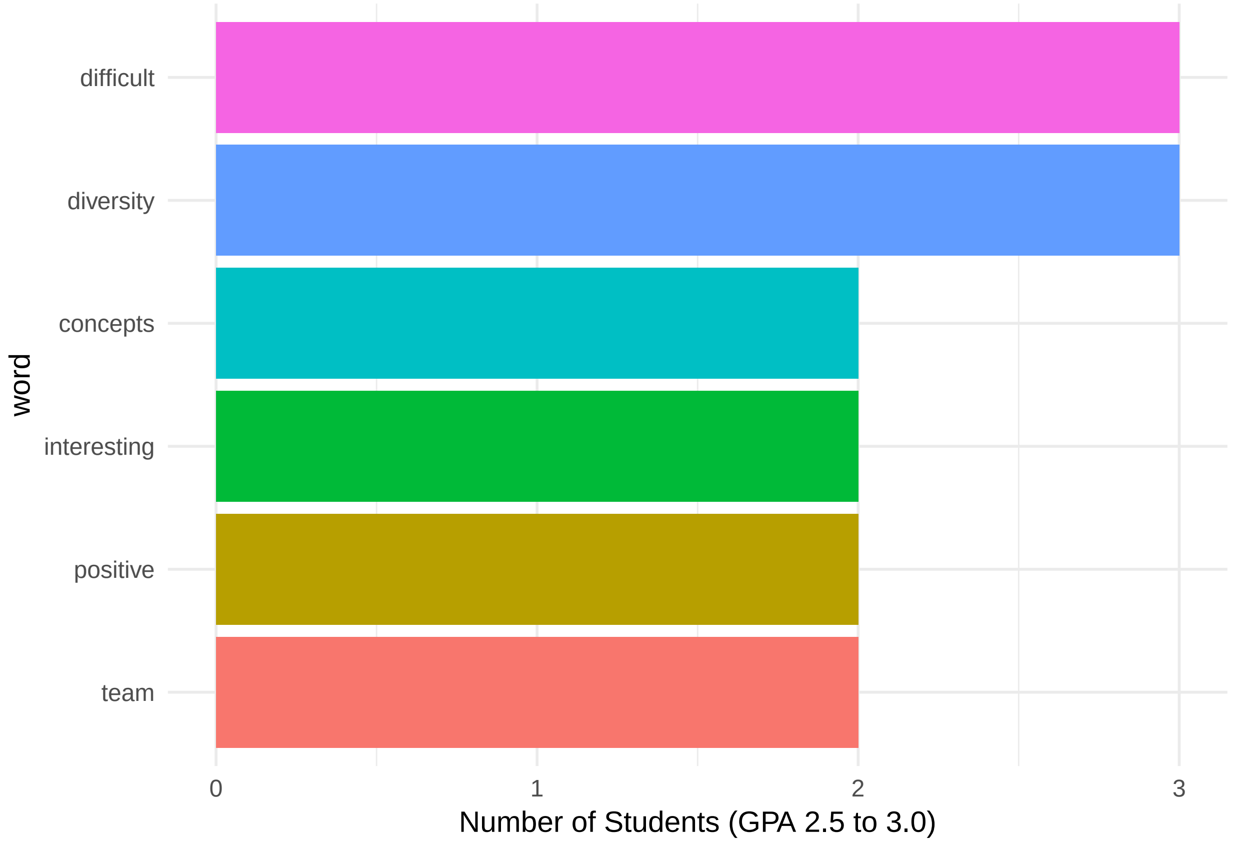
Most used “interesting”, “diversity”, and “applicable”.

Putting all plots together:

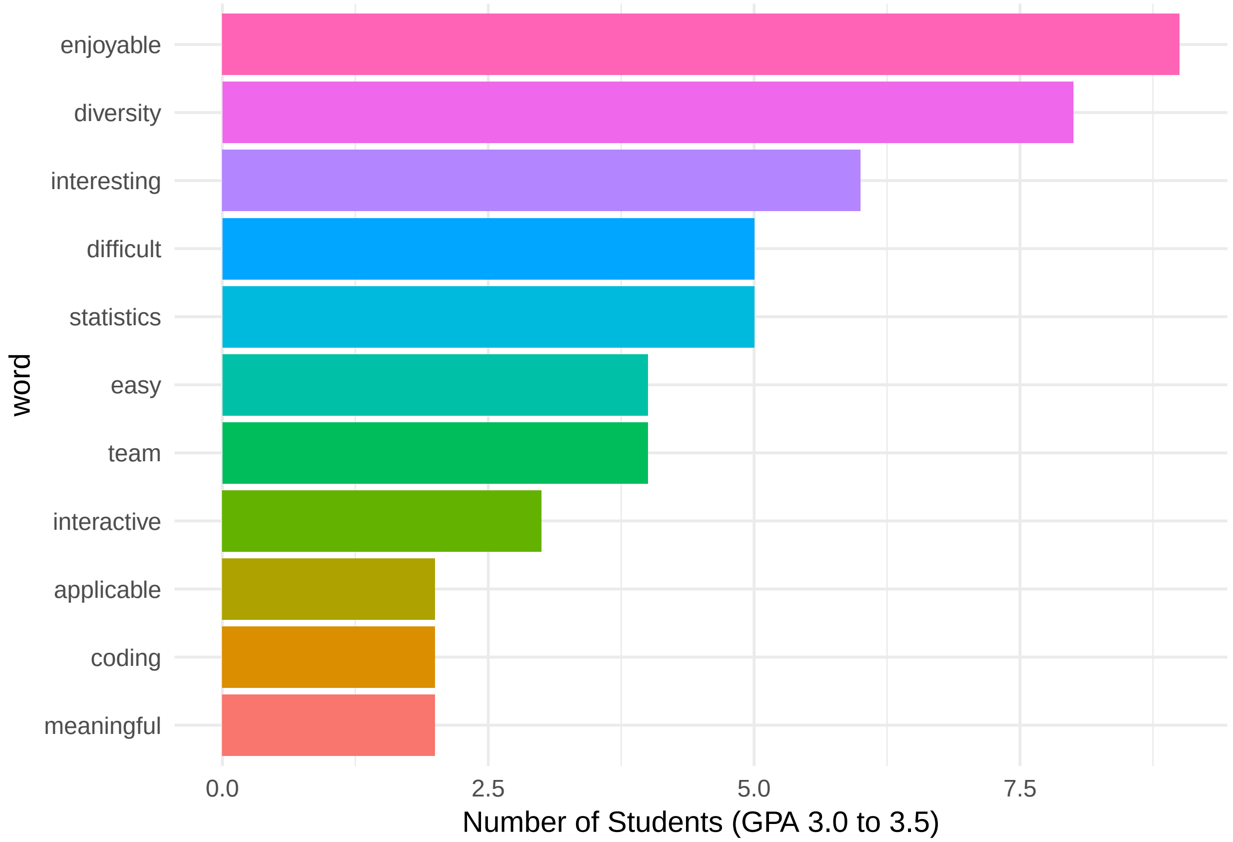
```
#summary(FINALDF)
par(mfrow = c(2,2))
ggplot(q2gpa1words, aes(x = word, y = n, fill = word)) + geom_bar(stat = "identity") + coord_flip() + tl
```



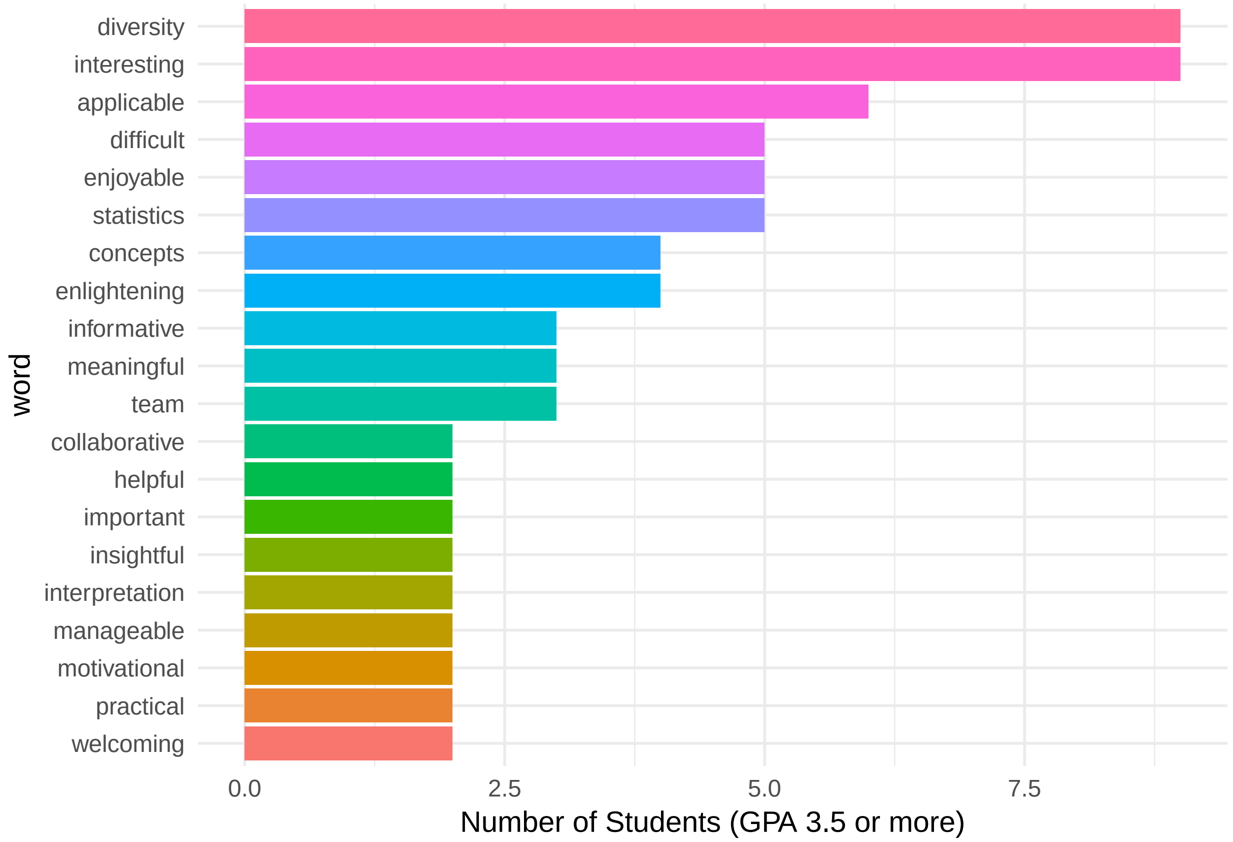
```
ggplot(q2gpa2words, aes(x = word, y = n, fill = word)) + geom_bar(stat = "identity") + coord_flip() + th
```



```
ggplot(q2gpa3words, aes(x = word, y = n, fill = word)) + geom_bar(stat = "identity") + coord_flip() + th
```



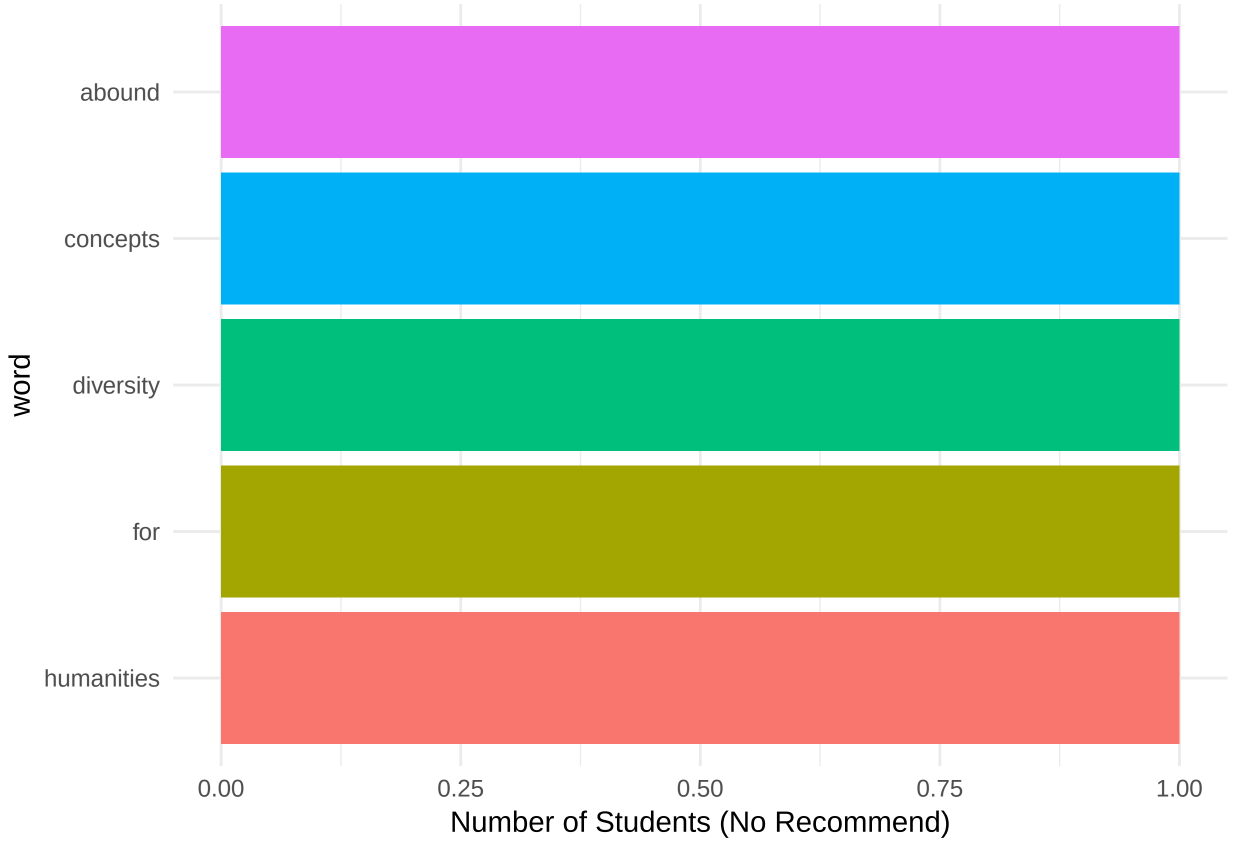
```
ggplot(q2gpa4words, aes(x = word, y = n, fill = word)) + geom_bar(stat = "identity") + coord_flip() + tl
```



Recommend:

3/5 did not respond to Page 3 - Question 2.

```
cleannotrec <- FINALDF[which(FINALDF$Recommend == "No"),]
cleannotrecwords <- cleannotrec[,21]

library(dplyr)

cleannotrecwords <- cleannotrecwords %>% na.omit()
cleannotrecwords <- as.character(cleannotrecwords)

cleannotrecwords <- tibble(line = 1:length(cleannotrecwords), text = cleannotrecwords)

library(tidytext)

cleannotrecwords <- cleannotrecwords %>% unnest_tokens(word, text)
cleannotrecwordsdf <- as.data.frame(cleannotrecwords[,2])
#cleanintwordsdf
library(ggplot2)

q2cleannotrecwords <- cleannotrecwordsdf %>%
  count(word, sort = TRUE) #%>%
#  filter(n > 1)

q2cleannotrecwords$word <- factor(q2cleannotrecwords$word, levels = rev(factor(q2cleannotrecwords$word))
ggplot(q2cleannotrecwords, aes(x = word, y = n, fill = word)) + geom_bar(stat = "identity") + coord_fli
```



Not many people did not recommend the class.

Statistics Majors vs Non-Statistics Majors:

```
finalpage1 <- read.csv("Copy of Statistics Data  - Page 1.csv", na.strings = c("", "NA"))
finalpage3 <- read.csv("Copy of Statistics Data  - Page 3.csv", na.strings = c("", "NA"))
finalpage1$words <- finalpage3[,3]

library(dplyr)
#textmine <- textmine %>% na.omit()

finalpage1$words <- as.character(finalpage1$words)

majortext_df <- tibble(line = 1:length(finalpage1$words), text = finalpage1$words)

library(tidytext)

majortext_df <- majortext_df %>% unnest_tokens(word, text) %>% na.omit()
majortext_df
```

```
## # A tibble: 271 x 2
##      line word
##     <int> <chr>
##  1      1 insightful
##  2      1 enjoyable
##  3      1 interesting
##  4      1 detailed
##  5      1 beautiful
##  6      3 awareness
##  7      3 groupwork
##  8      3 positive
##  9      3 disjointed
## 10      3 applicable
## # ... with 261 more rows
```

```
cleanmajortext <- majortext_df

applicable <- c(which(cleanmajortext[,2] == "application"), which(cleanmajortext[,2] == "applicantdrive
cleanmajortext[applicable, 2] <- "applicable"

##########

difficult <- c(which(cleanmajortext[,2] == "challenging"), which(cleanmajortext[,2] == "hard"), which(cl
cleanmajortext[difficult, 2] <- "difficult"

############

complex <- c(which(cleanmajortext[,2] == "comprehensive"), which(cleanmajortext[,2] == "crossdisciplina
cleanmajortext[complex, 2] <- "complex"

###########

collaborative <- c(which(cleanmajortext[,2] == "collaboration"), which(cleanmajortext[,2] == "cooperati
cleanmajortext[collaborative, 2] <- "collaborative"

###########

team <- c(which(cleanmajortext[,2] == "group"), which(cleanmajortext[,2] == "groupwork"), which(cleanma
cleanmajortext[team, 2] <- "team"

###########

diversity <- c(which(cleanmajortext[,2] == "diverse"))
cleanmajortext[diversity, 2] <- "diversity"

###########

easy <- c(which(cleanmajortext[,2] == "notthathard"), which(cleanmajortext[,2] == "easycourse"), which(
cleanmajortext[easy, 2] <- "easy"

############

engaging <- c(which(cleanmajortext[,2] == "engagement"))
cleanmajortext[engaging, 2] <- "engaging"

###########
```

```
enjoyable <- c(which(cleanmajortext[,2] == "enjoy"), which(cleanmajortext[,2] == "fun"), which(cleanmaj
cleanmajortext[enjoyable, 2] <- "enjoyable"

############

positive <- c(which(cleanmajortext[,2] == "good"))
cleanmajortext[positive, 2] <- "positive"

############

interactive <- c(which(cleantext[,2] == "handson"))
cleantext[interactive, 2] <- "interactive"
```

```
## Warning in `[<-.factor`(`*tmp*`, iseq, value = c("interactive",
## "interactive", : invalid factor level, NA generated
```

```
############

insightful <- c(which(cleanmajortext[,2] == "insight"))
cleanmajortext[insightful, 2] <- "insightful"

############

interesting <- c(which(cleanmajortext[,2] == "intriguing"))
cleanmajortext[interesting, 2] <- "interesting"

###########

manageable <- c(which(cleanmajortext[,2] == "managable"))
cleanmajortext[manageable, 2] <- "manageable"

###########

enlightening <- c(which(cleanmajortext[,2] == "mindopening"), which(cleanmajortext[,2] == "thoughtprovo
cleanmajortext[enlightening, 2] <- "enlightening"

###########

motivational <- c(which(cleanmajortext[,2] == "motivating"))
cleanmajortext[motivational, 2] <- "motivational"

###########

practical <- c(which(cleanmajortext[,2] == "practicality"))
cleanmajortext[practical, 2] <- "practical"

###########

coding <- c(which(cleanmajortext[,2] == "r"), which(cleanmajortext[,2] == "programming"), which(cleanma
cleanmajortext[coding, 2] <- "coding"

###########

reaction <- c(which(cleanmajortext[,2] == "reactionpaper"))
```

```
cleanmajortext[reaction, 2] <- "reaction"

###########

respectful <- c(which(cleanmajortext[,2] == "respect"))
cleanmajortext[respectful, 2] <- "respectful"

###########

special <- c(which(cleanmajortext[,2] == "unique"))
cleanmajortext[special, 2] <- "special"

###########

statistics <- c(which(cleanmajortext[,2] == "stats"), which(cleanmajortext[,2] == "basicstats"), which(c
cleanmajortext[statistics, 2] <- "statistics"

###########

concepts <- c(which(cleanmajortext[,2] == "linearregression"), which(cleanmajortext[,2] == "mlr"), which
cleanmajortext[concepts, 2] <- "concepts"

###########

cleanmajortext <- cleanmajortext %>% group_by(line) %>% summarise(words = paste(word, collapse=" "))
finalpage1$line <- c(1:70)
cleanmajortext <- dplyr::full_join(finalpage1, cleanmajortext, by = "line")
cleanmajortext <- cleanmajortext[,-c(21:22)]
```

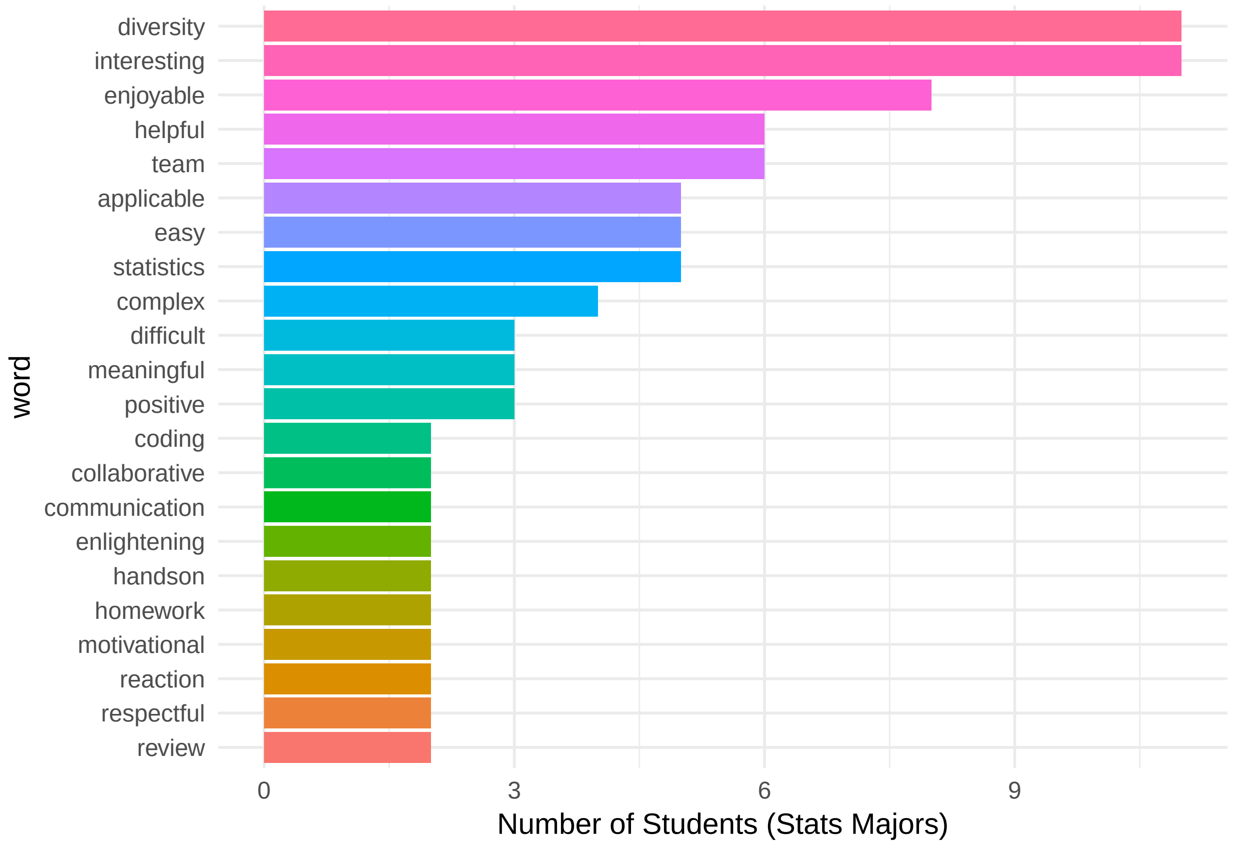
```
statsmajor <- cleanmajortext[which(cleanmajortext$Major == "Statistics"),]
statsmajorwords <- statsmajor[,21]

nonstatsmajor <- cleanmajortext[which(cleanmajortext$Major != "Statistics"),]
nonstatsmajorwords <- nonstatsmajor[,21]
```

Words Used by Statistics Majors (clean):

```
library(dplyr)

statsmajorwords <- statsmajorwords %>% na.omit()
statsmajorwords <- as.character(statsmajorwords)

statsmajorwords <- tibble(line = 1:length(statsmajorwords), text = statsmajorwords)

library(tidytext)

statsmajorwords <- statsmajorwords %>% unnest_tokens(word, text)
statsmajorwords
```

```
## # A tibble: 129 x 2
##     line word
##    <int> <chr>
##  1     1 awareness
##  2     1 team
##  3     1 positive
##  4      1 disjointed
##  5      1 applicable
##  6      2 easy
##  7      2 easy
##  8      2 positive
##  9      2 enjoyable
## 10      2 relevant
## # ... with 119 more rows
```

```
statswordsdf <- as.data.frame(statsmajorwords[,2])
library(ggplot2)

q2statswords <- statswordsdf %>%
  count(word, sort = TRUE) %>%
  dplyr::filter(n > 1)

q2statswords$word <- factor(q2statswords$word, levels = rev(factor(q2statswords$word)))

ggplot(q2statswords, aes(x = word, y = n, fill = word)) + geom_bar(stat = "identity") + coord_flip() +
```

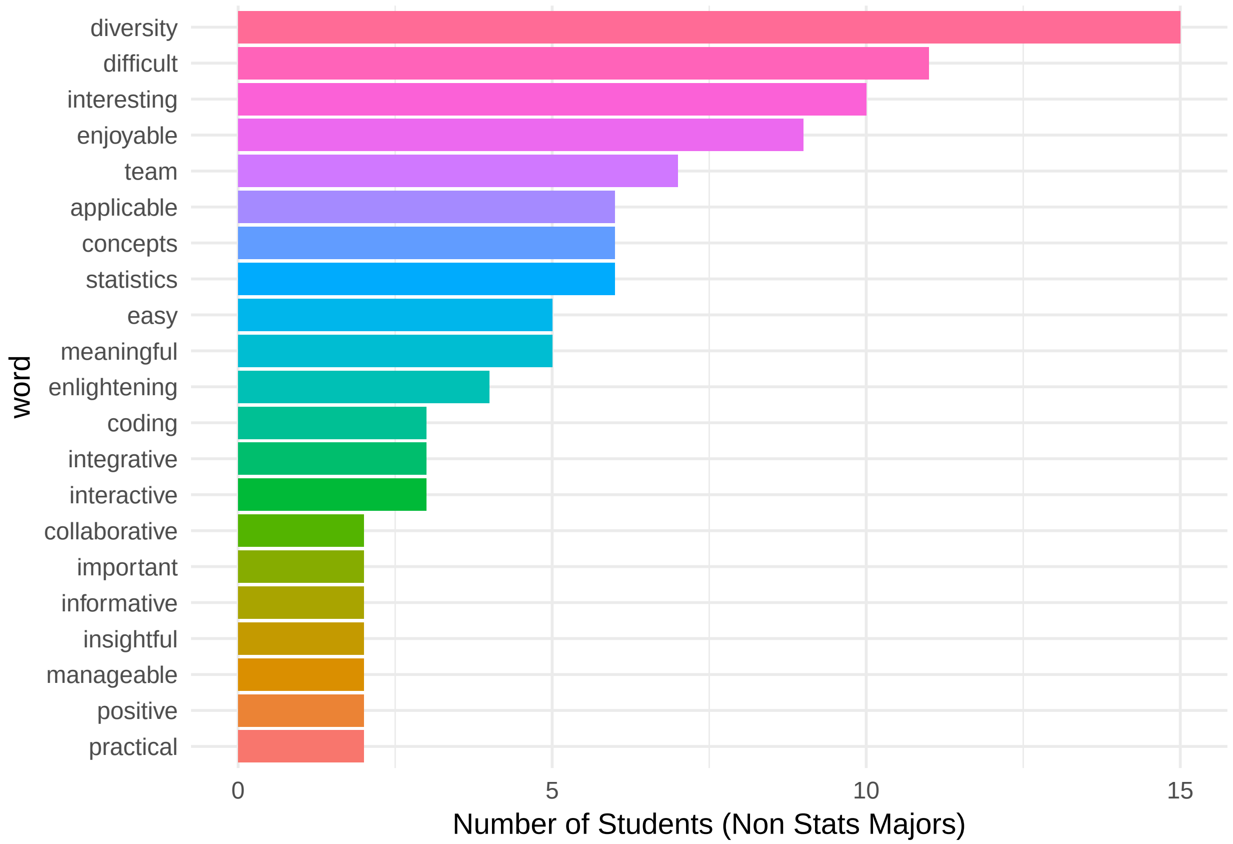


Words Used by Non-Statistics Majors:

```
library(dplyr)

nonstatsmajorwords <- nonstatsmajorwords %>% na.omit()
nonstatsmajorwords <- as.character(nonstatsmajorwords)
nonstatsmajorwords <- tibble(line = 1:length(nonstatsmajorwords), text = nonstatsmajorwords)

library(tidytext)

nonstatsmajorwords <- nonstatsmajorwords %>% unnest_tokens(word, text)
nonstatsmajorwords
```

```
## # A tibble: 142 x 2
##     line word
##    <int> <chr>
##  1     1 insightful
##  2     1 enjoyable
##  3     1 interesting
##  4     1 detailed
##  5     1 beautiful
##  6     2 enlightening
##  7     2 meaningful
##  8     2 special
##  9     2 informative
## 10     2 enjoyable
## # ... with 132 more rows
```

```
nonstatswordsdf <- as.data.frame(nonstatsmajorwords[,2])
library(ggplot2)

q2nonstatswords <- nonstatswordsdf %>%
  count(word, sort = TRUE) %>%
  dplyr::filter(n > 1)

q2nonstatswords$word <- factor(q2nonstatswords$word, levels = rev(factor(q2nonstatswords$word)))

ggplot(q2nonstatswords, aes(x = word, y = n, fill = word)) + geom_bar(stat = "identity") + coord_flip()
```



Graphically Depicting Words Used by Statistics Majors vs Non-Statistics Majors:

- 50/50 split between Statistics and Non-Statistics Majors (including double majors).

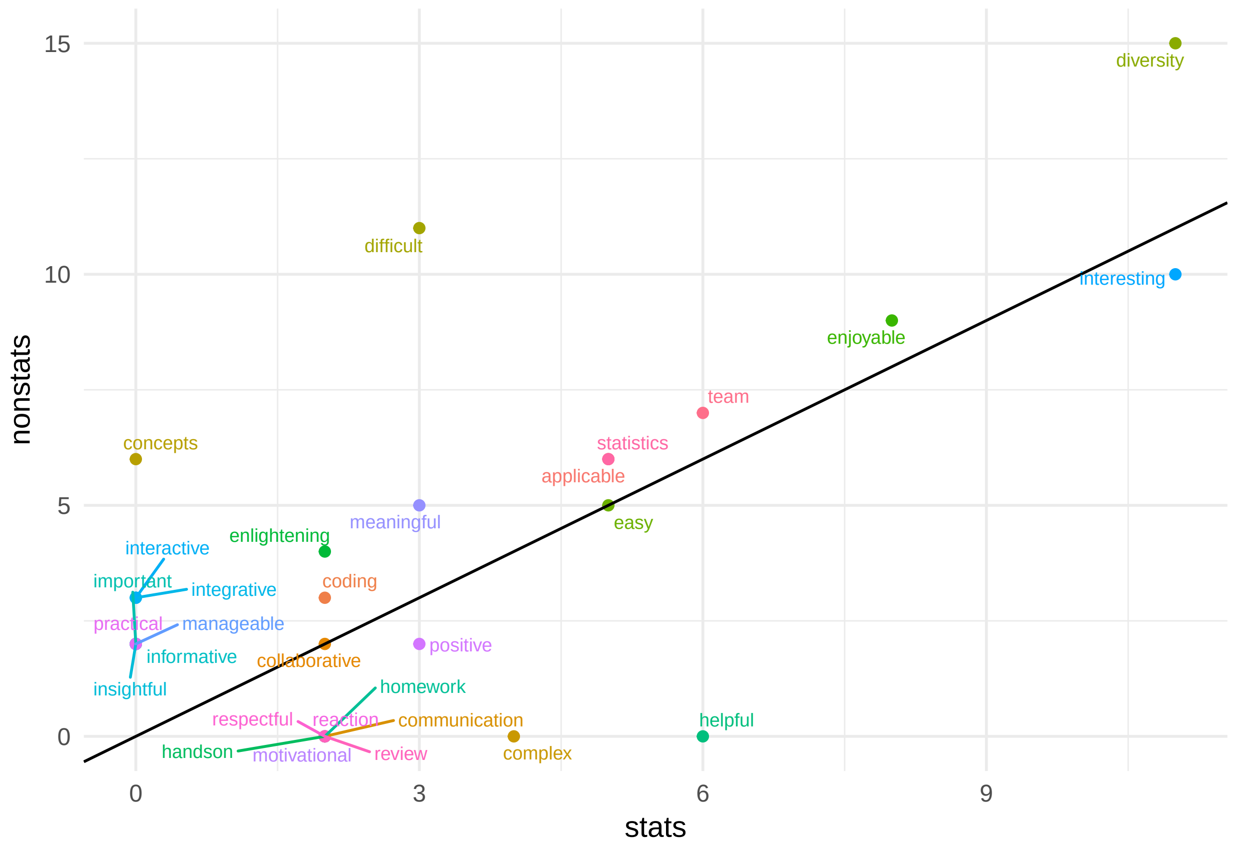
```
majorcombo <- dplyr::full_join(q2nonstatswords, q2statswords, by = "word")
```

```
## Warning: Column `word` joining factors with different levels, coercing to
## character vector
```

```
colnames(majorcombo)[c(2,3)] <- c("nonstats","stats")
majorcombo <- majorcombo %>% replace_na(list(nonstats = 0, stats = 0))

#install.packages("ggrepel")
library(ggrepel)
```

```
## Warning: package 'ggrepel' was built under R version 3.5.2
```

```
ggplot(majorcombo, aes(x = stats, y = nonstats, col = word, label = word)) + geom_point() + theme_minima
```


15
10
5
0
nonstats
0
3
6
9
stats
diversity
difficult
interesting
enjoyable
team
statistics
applicable
concepts
meaningful
easy
enlightening
interactive
important
integrative
coding
practical
manageable
informative
collaborative
positive
insightful
homework
respectful
reaction
communication
helpful
handson
motivational
review
complex