Tackling Total Withdrawals with Text Analytics

Christianna Brown

March 11, 2018

At the university I work for, there is an administrative process for students that want to withdraw from the university or take a leave of absence; this process requires them to submit an online Total Withdrawal/Leave of Absence (TWLOA) form. It is has been determined that the current process does not adequately inform students of the consequences that could result from this leave, both academically and financially. The goal of this project is to analyze the university's TWLOA data in order to improve the current process and find a way to assist students more successfully. This project will use exploratory data analysis and text analytics. The exploratory data analysis will be applied to data on students who have utilized the form; it will allow us to gain insight into what demographics of students are leaving, where these students are from, and if they are planning on transferring to a different school. From there text analytics will be applied to the drop-down text options and free-form text data that students submit via the form; this will include LSA, TF-IDF, and sentiment analysis.

Loading the Required Libraries

To begin the process in R, we will need to load the required libraries:

```
#Loading the required libraries
library(xlsx)
library(tm)
library(readr)
library(SnowballC)
library(ggplot2)
library(wordcloud)
library(RColorBrewer)
library(proxy)
library(dplyr)
library(lsa)
library(LSAfun)
library(stylo)
library(tidyverse)
library(tidytext)
library(glue)
library(stringr)
```

These libraries will allow for the data import of Excel files, exploratory and text analytics, as well as visual representations of the outcomes. As this project progresses, some libraries may be added or removed as needed. These many libraries are a combination of multiple prior projects using the same skills.

Loading and Reviewing the Data

The data for this project is approximately 334 KB of masked demographics and text responses saved as an Excel file. This Excel file is available for download in the git repository. Utilizing the **xlsx** library, we can easily import Excel files and review the data:

```
#Load the dataset
twloa_full <- read.xlsx("D:/Users/Christianna Brown/Desktop/MSDS 692/Data Fil
es/TWLOA All - Masked.xlsx", 1, header = T) #This imports the data into R
twloa_full <- twloa_full[, -c(12:25)] #This removes unnecessary, hidden rows
attach(twloa_full) #This allows the user to call specific rows easily

##   Reference..   Status          City         State Zip.Code       Class
## 1     9811295 Complete        Gurabo   Puerto Rico      778 Certificate
## 2    10802323 Complete     worcester Massachusetts     1605        Grad
## 3    10392813 Complete Jamaica Plain Massachusetts     2130        Grad
## 4    10914112 Complete    Somerville Massachusetts     2144          FR
## 5    10145568 Complete      winthrop Massachusetts     2152        Grad
## 6    11054893 Complete        Quincy Massachusetts     2169        Grad
##                                                 College
## 1          College of Contemporary Liberal Studies (CCLS)
## 2          College of Contemporary Liberal Studies (CCLS)
## 3 Rueckert-Hartman College for Health Professions (RHCHP)
## 4      College of Computer & Information Sciences (CC&IS)
## 5          College of Contemporary Liberal Studies (CCLS)
## 6                College of Business and Economics (CBE)
##                   Period Withdrawal       Start         End
## 1        Total Withdrawal        Yes
## 2        Total Withdrawal        Yes             Summer 2017
## 3        Total Withdrawal        Yes Spring 2018
## 4    Within Two Semesters        Yes               Fall 2018
## 5        Total Withdrawal        Yes
## 6 More Than Two Semesters        Yes Spring 2018
##
Text
## 1
Another Opportunity Change of Direction      Other Master's Forensic Science
Changed direction personally/ Regis no longer meets my needs  It's a very goo
d school with excellent programs, but I decided to go with Forensic Science t
aking courses on campus in a school here in Puerto Rico. In my opinion every
is fine, I just decided to take a different direction. Undecided Everything i
s already taken care of.
## 2 Unhappy/Institutional Fit Other The online professors were very irrespon
sible and showed no professionalism.       Other online student  An F had bee
n given to me for a final grade i tried to appeal the grade its been a month
now and i haven't heard of anything coming back. The online professors cant h
andle the amount of students they have it looks like they are working under p
ressure with so many students. i can say i will never recommend anyone to thi
s school. im am very disappointing.  I've been charged 60% of the second cour
se when the professor never acknowledged me. Work Undecided
```

```
## 3
Financial  My financial situation has changed and my aid package is no longer
sufficient          Work
## 4                                                 Another Opportunity Scheduli
ng Conflicts      Other International arbitration   I work as a lawyer and we
are in the middle of a large international arbitration which is taking a disp
roportionate amount of my time. It will be done in the summer at the latest.
I fell behind by a small amount in a computer science class and explained the
situation to my professor, who was not sympathetic.  Work I would simply like
to have the option to study here again as I truly enjoyed the experience and
the work that I did resulted in good grades (A-s and As).
## 5
Financial Personal Scheduling Conflicts  Not affordable/ too expensive My fin
ancial situation has changed and my aid package is no longer sufficient  Need
ed to work        Work
## 6
Financial  Other employer discount          Work
##   Duration Browser      OS
## 1      422  Safari     Mac
## 2      748  Chrome Windows
## 3      239 Firefox     Mac
## 4      305  Safari     Mac
## 5      246      IE Windows
## 6      136      IE Windows

## 'data.frame':    905 obs. of  15 variables:
##  $ Reference..: Factor w/ 905 levels "10003399","10004246",..: 851 462 236
549 118 665 469 832 541 241 ...
##  $ Status     : Factor w/ 1 level "Complete": 1 1 1 1 1 1 1 1 1 1 ...
##  $ City       : Factor w/ 382 levels "Alameda","Alamosa",..: 129 380 152 3
32 379 292 180 209 196 108 ...
##  $ State      : Factor w/ 48 levels "Alabama","Alaska",..: 38 20 20 20 20
20 20 20 28 28 ...
##  $ Zip.Code   : num  778 1605 2130 2144 2152 ...
##  $ Class      : Factor w/ 7 levels "Certificate",..: 1 4 4 2 4 4 2 6 2 4 .
..
##  $ College    : Factor w/ 5 levels "College of Business and Economics (CBE
)",..: 3 3 5 2 3 1 4 4 2 3 ...
##  $ Period     : Factor w/ 3 levels "More Than Two Semesters",..: 2 2 2 3 2
1 2 2 2 2 ...
##  $ Withdrawal : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 1 2 2 ...
##  $ Start      : Factor w/ 4 levels "","Fall 2018",..: 1 1 3 1 1 3 1 1 1 3
...
##  $ End        : Factor w/ 7 levels "","Fall 2017",..: 1 6 1 3 1 1 1 1 1 1
...
##  $ Text       : Factor w/ 838 levels "Academic             ",..: 112 836 1
66 118 334 216 589 260 802 661 ...
##  $ Duration   : Factor w/ 541 levels "100680","1014",..: 363 489 188 259 1
97 65 55 180 285 102 ...
##  $ Browser    : Factor w/ 5 levels "Chrome","Firefox",..: 5 1 2 5 3 3 1 5
```

```
1 1 ...
##  $ OS         : Factor w/ 5 levels "Linux","Mac",..: 2 5 2 2 5 5 5 2 5 5 .
..
```

From this we can see that this dataset contains 905 observations across 15 variables. These variables include:

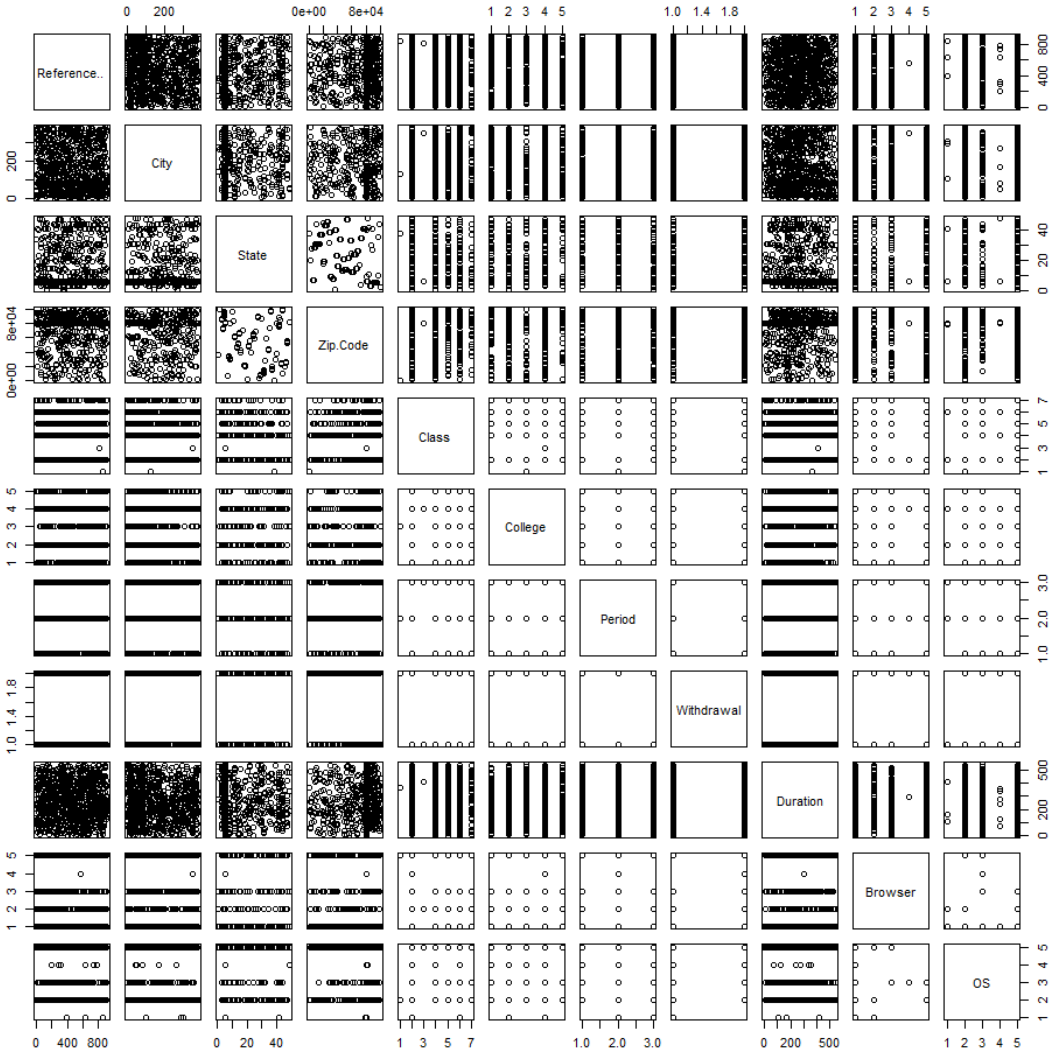
1. **Reference** - The unique identifier for each student submission
2. **Status** - The status of the form: Complete/Incomplete (This project only evaluates Complete forms)
3. **City** - The student's city of residence
4. **State** - The student's state of residence
5. **Zip Code** - The student's residential zip code
6. **Class** - The student's class level (e.g. Freshman, Sophomore, etc.)
7. **College** - The student's college of enrollment (e.g. Regis College, College of Computer & Information Systems, etc.)
8. **Period** - The length of the withdrawal/leave period
9. **Withdrawal** - Whether the student wants to be withdrawn from their current registration
10. **Start** - The start of leave
11. **End** - The end of leave
12. **Text** - The text answers submitted by the student
13. **Duration** - The time the student took to complete the form
14. **Browser** - The web browser used to complete the form
15. **OS** - The operating system used to complete the form

With these different variables we are able to complete both the exploratory and text analytics. The full dataset will be broken down into two smaller sets, one for the EDA and another for the text analytics.

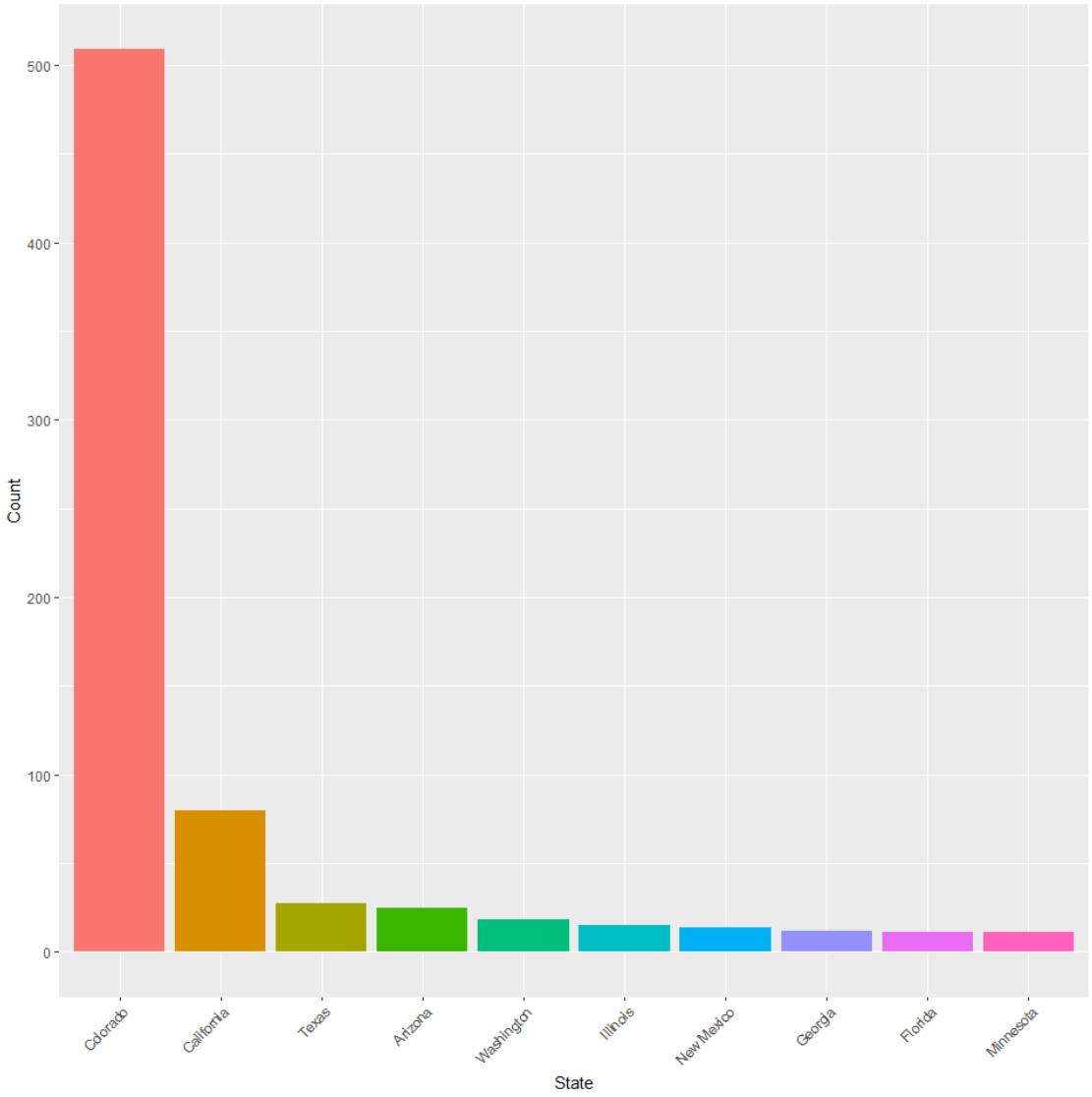
Exploratory Data Analysis

The exploratory analysis will review the student demographic and educational information from the submitted forms. First, we will create our EDA dataset and review the pairs data to determine if there are any interesting relationships present:

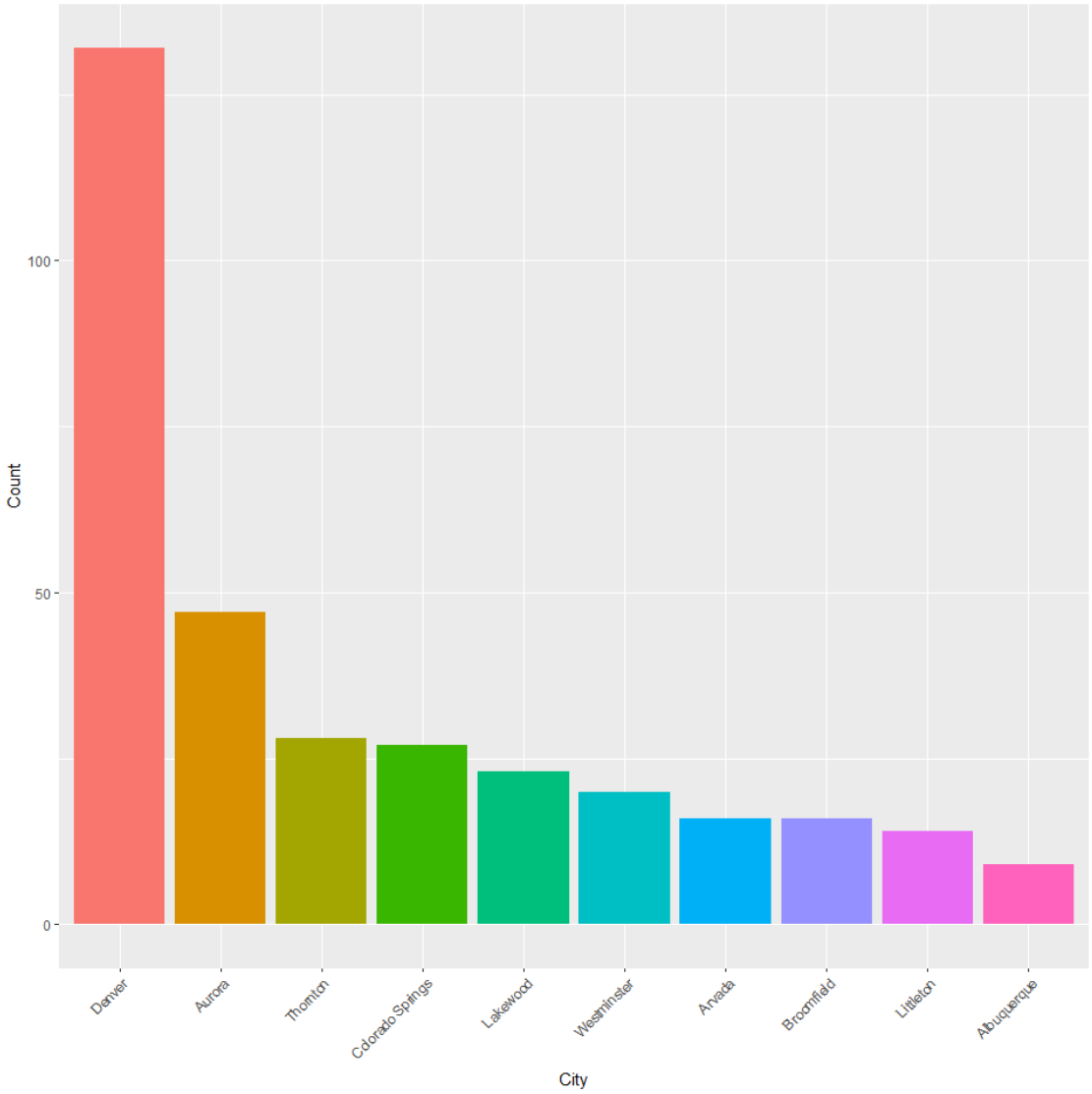
```
#EDA
twloa_EDA <- twloa_full[, c(1, 3, 4, 5, 6, 7, 8, 9, 13, 14, 15)]
pairs(twloa_EDA)
```



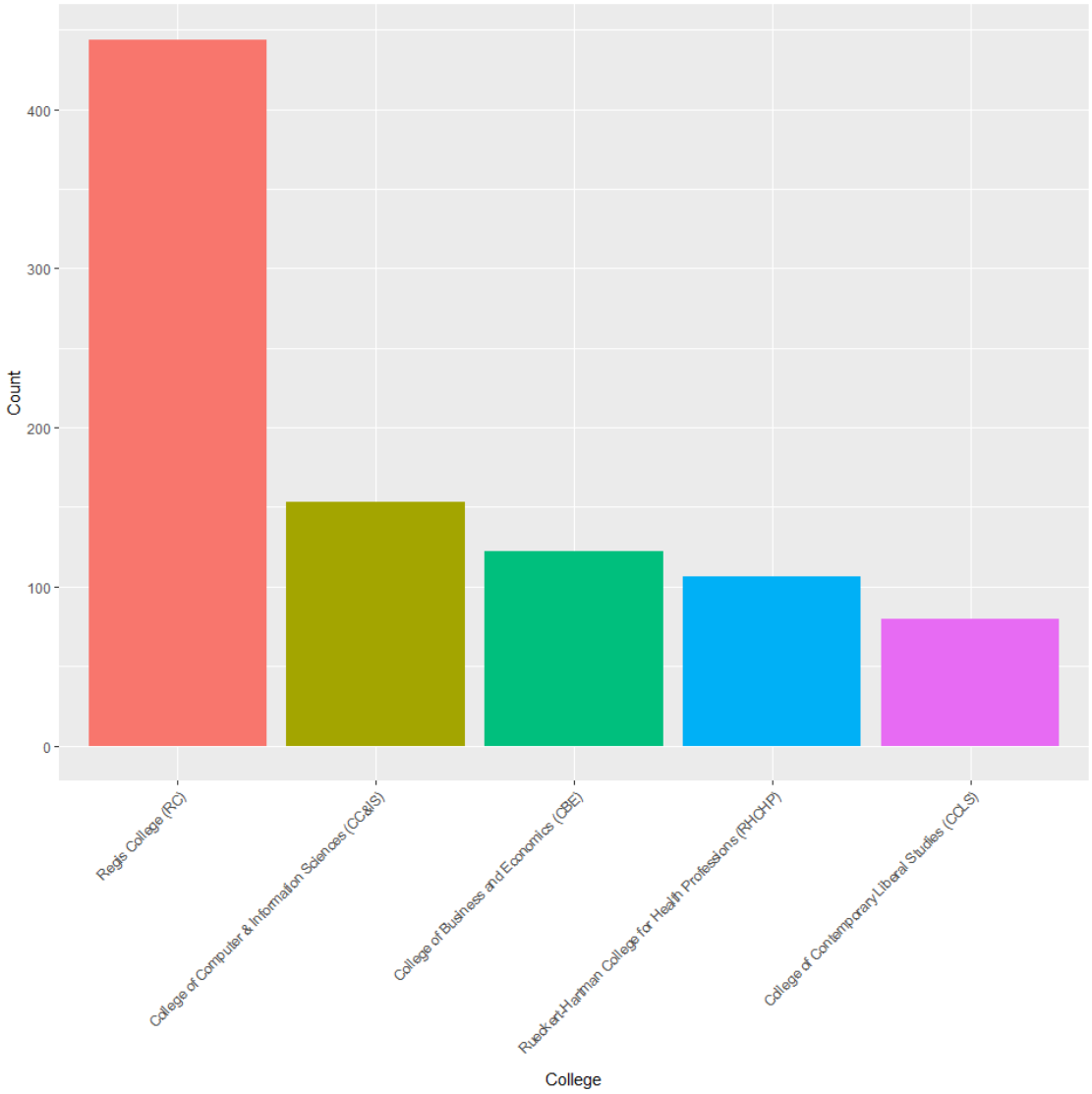
This shows us that certain variables may have some relationships; such as certain class levels may be from certain colleges or the browser used may be dependent on the operating system. However, it appears that most of these items may need to be reviewed individually. We will now review plots of the state and city information to determine where students that leave are from; this will help us to better understand if we are losing more local students or more students from out-of-state:



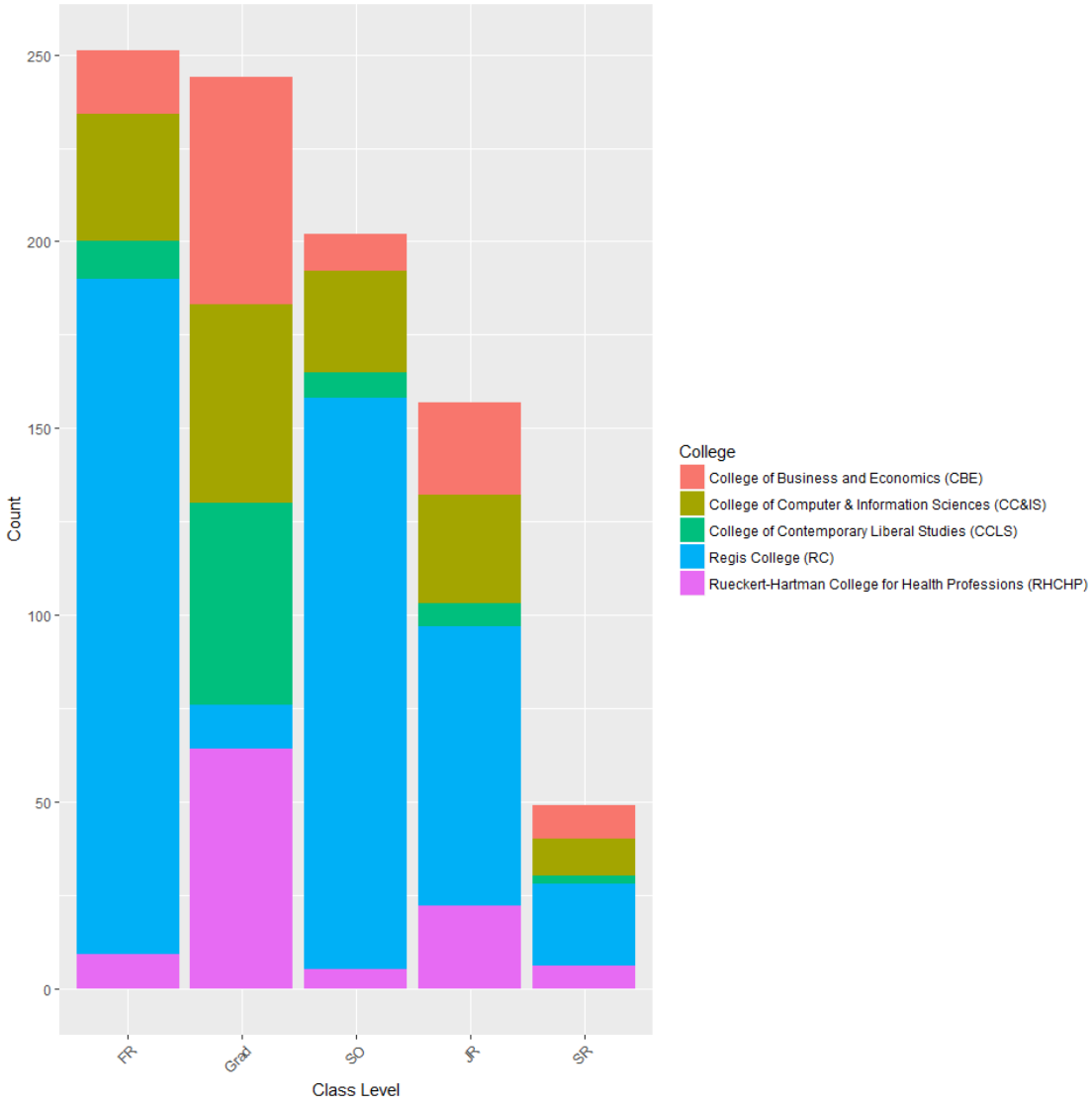
This clearly shows that the majority of students leaving are from Colorado, with California students leaving second. This is disheartening to see, since we definitely need our local student population to thrive. Perhaps looking into the city data with provide us with greater insight on where they are from within Colorado:



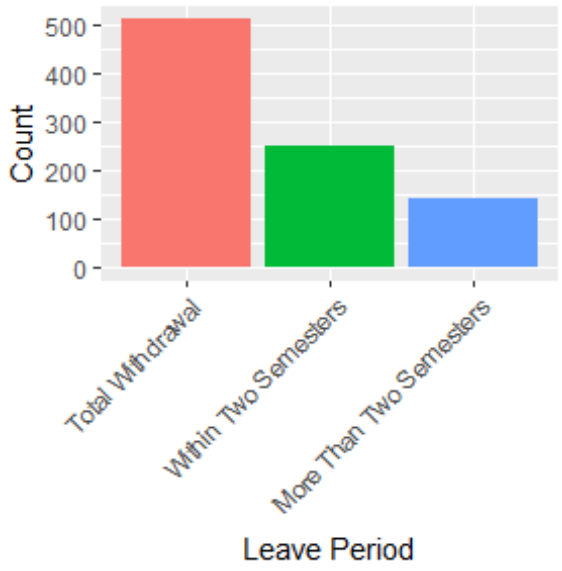
The city data provides us with further detail emphasizing that the majority of students leaving are from the Metro-Denver area or a suburb. This is something that could be analyzed deeper and perhaps we could pinpoint what student populations are leaving from specific cities, but this will be reviewed later. We will now review the college and class level of students leaving. It is important for us to understand which colleges within the university are at the greatest risk of losing students:

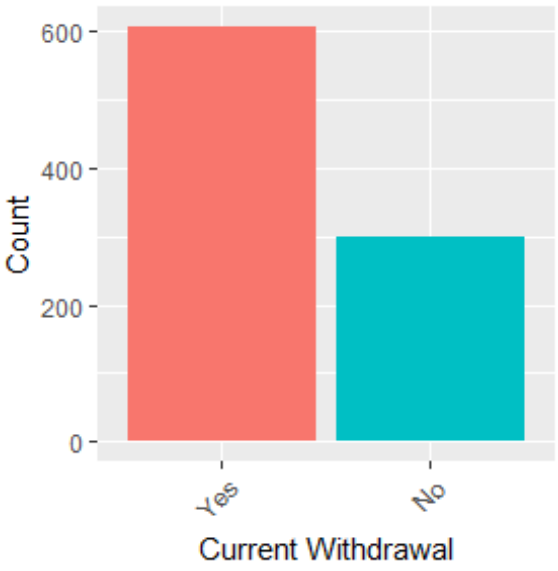


This shows us that the majority of students who have left and completed this form are from Regis College (RC). This is not particularly surprising at this time, since the form was originally intended for only the traditional undergraduate students. However, now that the form has been opened up to all programs and colleges it would be beneficial to track which students are leaving from which colleges. This may be accomplished by combining class level and filling with the college data:

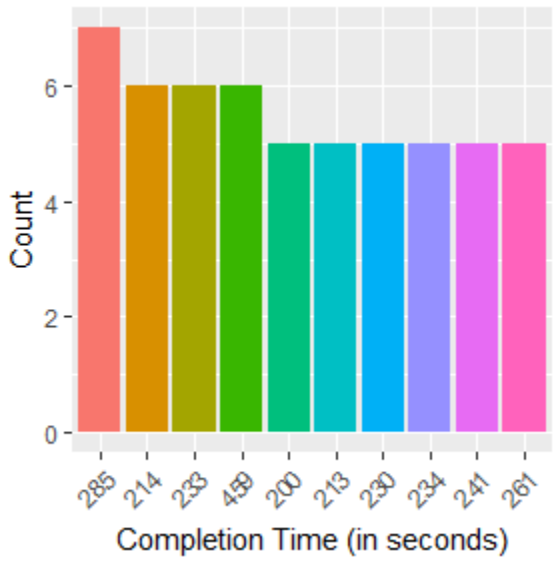


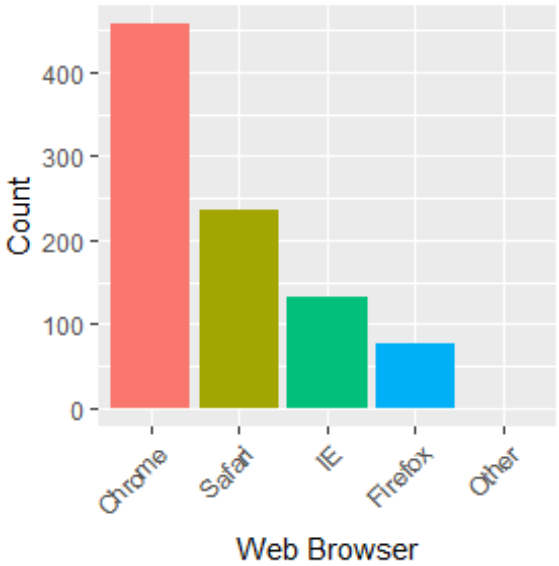
This provides us with further insight into which students are leaving from which college. The majority of undergraduates that leave are from Regis College; this is not surprising since, once again, this form was originally intended for that population of student. CCLS has the fewest number of undergraduate withdrawals, but they are losing a large number of graduate students. When examining graduate programs, all colleges are evenly split, with the exception of Regis College, since they only have four, highly competitive graduate programs available. There is in interesting spike in RHCHP withdrawals during the junior year, this could be due to the make-or-break clinicals required during that year. There are definitely some very apparent issues that need to be addressed in both undergraduate and graduate programs in order to prevent future withdrawals. We will now examine whether students have elected to totally withdrawal or simply take a break and whether they want to be withdrawn from current registration:

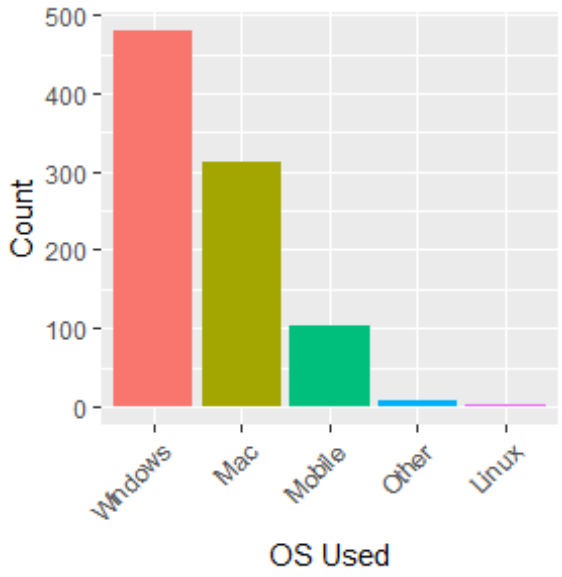




This shows that most students submitting this form have determined they will not be returning to Regis and they want to be withdrawn immediately. This indicates that when students are unhappy, they are extremely unhappy and are completely done with the university. The final bit of EDA is on technical items; including the duration, browser, and operating system. This will examine how long it takes students to complete the form and their method of submission:







The average student takes between three and five minutes to complete the form and the majority of them use Chrome and on a Windows machine. This further emphasizes the fact that when these students complete the form, they are really ready to be done with Regis and do not need to take much time to complete the form.

Text Analytics

The text analytics portion will utilize TF-IDF, LSA, and Cosine Similarity. It will also review the sentiment of the words used within the form. We will start by creating the text analytics dataset and creating the corpus that will be used for the TF-IDF, LSA, and Cosine Similarity:

```
#Text analytics
twloa_text <- twloa_full[, c(1, 12)]
twloa_corpus <- VCorpus(VectorSource(twloa_text$Text))
twloa_corpus

## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 905

twloa_corpus <- tm_map(twloa_corpus, removePunctuation)
twloa_corpus <- tm_map(twloa_corpus, removeNumbers)
twloa_corpus <- tm_map(twloa_corpus, content_transformer(tolower))
twloa_corpus <- tm_map(twloa_corpus, removeWords, stopwords("english"))
twloa_corpus <- tm_map(twloa_corpus, removeWords, c("regis", "need", "change"
, "person", "time", "direct", "longer", "fit", "feel", "student", "didnt", "f
elt"))
twloa_corpus <- tm_map(twloa_corpus, stripWhitespace)
twloa_corpus <- tm_map(twloa_corpus, stemDocument)
```

Once the corpus is created we can review the most frequent terms used:

```r
#Inspect words used at least 100 times and at least 300 times
#Creating a Document-Term Matrix
twloa_dtm <- DocumentTermMatrix(twloa_corpus)
inspect(twloa_dtm)
```

```
## <<DocumentTermMatrix (documents: 905, terms: 2220)>>
## Non-/sparse entries: 17757/1991343
## Sparsity           : 99%
## Maximal term length: 18
## Weighting          : term frequency (tf)
## Sample             :
##      Terms
## Docs  academ chang direct famili financi health need person transfer work
##   502      0     0      1      4       0      1    0      1        0    1
##   575      0     1      0      0       2      0    2      1        1    2
##   597      0     0      0      0       1      0    0      0        1    0
##   611      2     3      3      1       3      0    4      2        0    2
##   641      0     1      0      1       4      0    1      3        0    3
##   67       3     0      1      0       0      0    1      1        0    1
##   763      3     0      0      0       0      0    1      0        0    0
##   870      2     2      3      1       0      0    2      2        1    2
##   871      3     1      1      1       2      0    1      2        1    0
##   872      0     0      0      0       1      0    1      1        1    1
```

```r
#Finding terms used at least 100 times and 300 times
findFreqTerms(twloa_dtm, 100)
```

```
##  [1] "academ"          "afford"          "aid"
##  [4] "anoth"           "burn"            "chang"
##  [7] "class"           "colleg"          "conflict"
## [10] "cours"           "direct"          "expens"
## [13] "famili"          "financi"         "general"
## [16] "happi"           "health"          "help"
## [19] "home"            "just"            "life"
## [22] "love"            "medicalhealth"   "meet"
## [25] "mental"          "need"            "opportun"
## [28] "packag"          "person"          "program"
## [31] "school"          "situat"          "take"
## [34] "transfer"        "undecid"         "unhappyinstitut"
## [37] "univers"         "wasnt"           "will"
## [40] "work"
```

```r
findFreqTerms(twloa_dtm, 300)
```

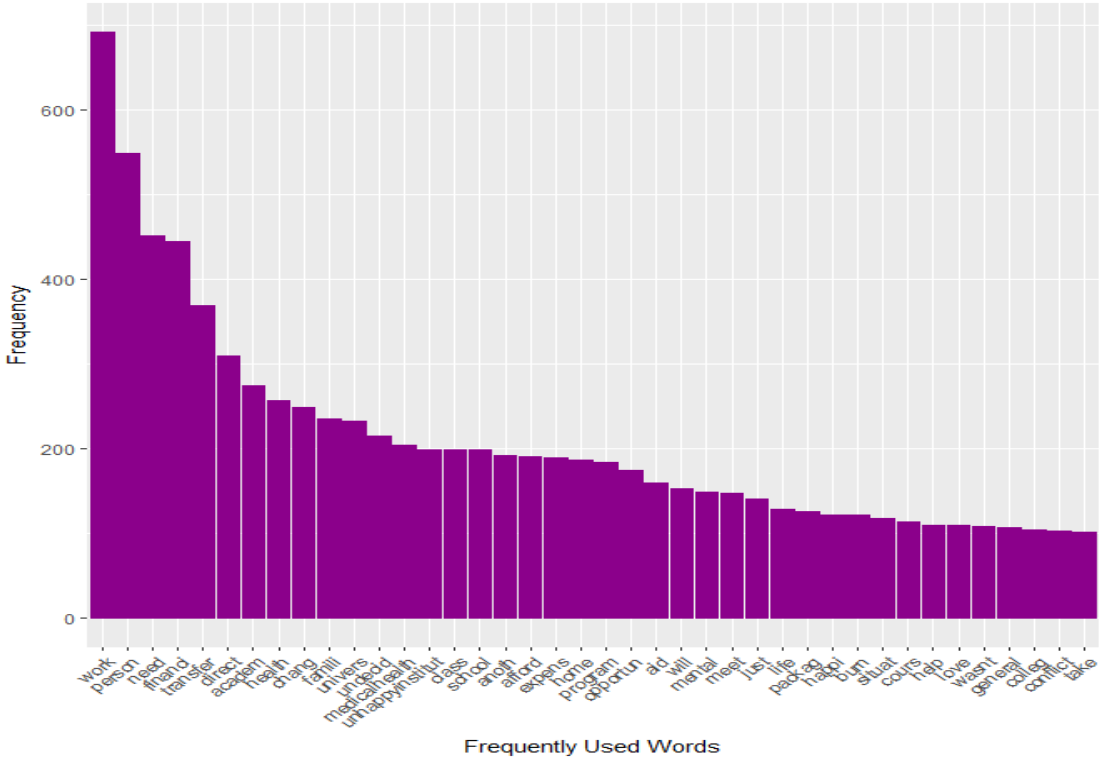
```
## [1] "direct"   "financi"  "need"     "person"   "transfer" "work"
```

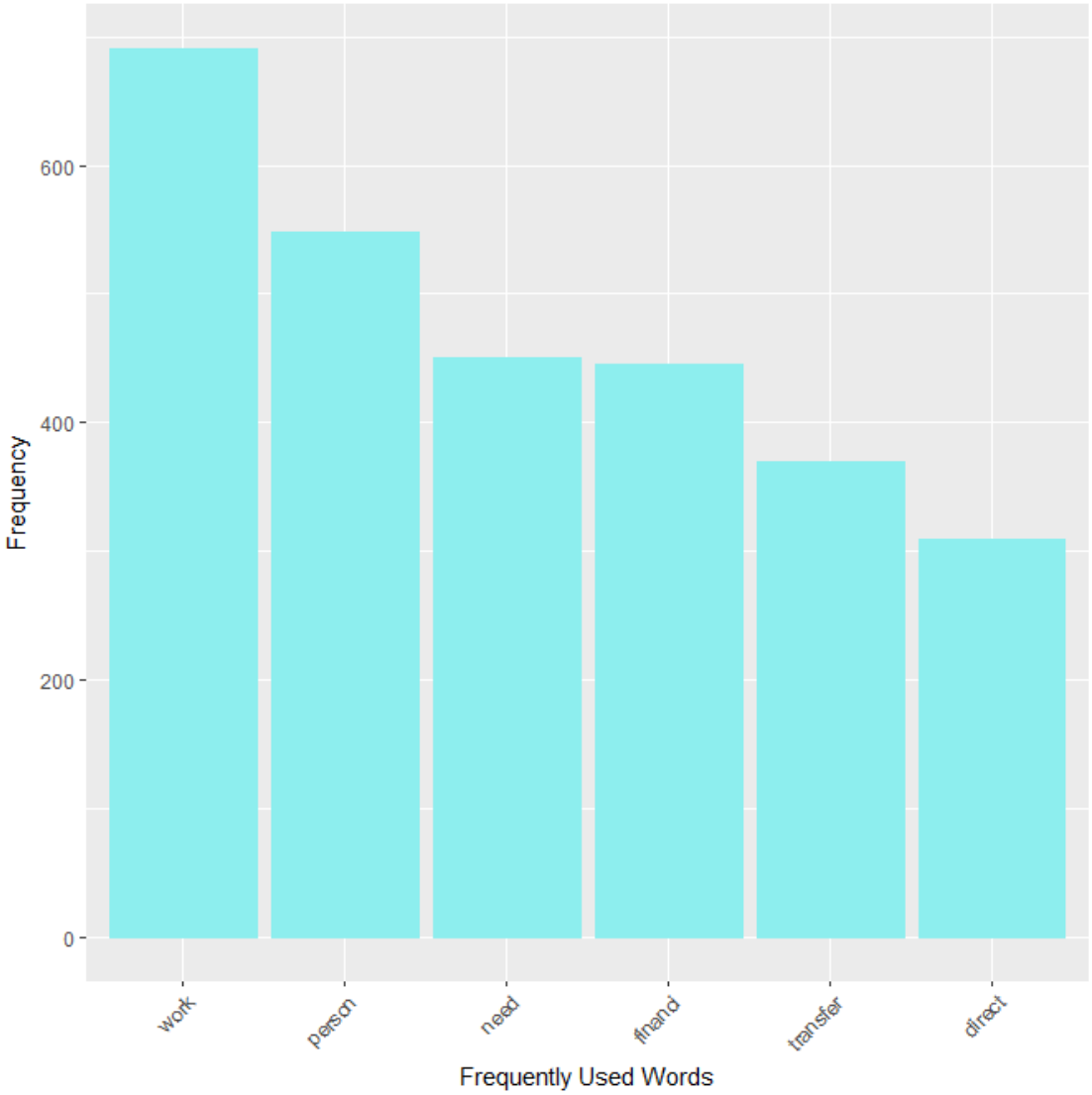
```r
#Creating a frequency table for the wordclouds
twloa_freq <- colSums(as.matrix(twloa_dtm))
ord <- order(twloa_freq, decreasing = T)
```

```
#Listing the most and least frequent terms
twloa_freq[head(ord)]

##     work   person     need  financi transfer   direct
##      691      548      451      445      369      309

twloa_freq[tail(ord)]

## younger   youth  youtub    youv  yunnan    ywam
##       1       1       1       1       1       1
```

This shows that there are 17757 non-sparse terms within the documents, which is a 99% sparsity rate. The most frequently used terms are *work*, *person*, *need*, and *financi*. We can now review the plots for words used at least 100 times and at least 300 times:

```
#Plotting the words used at least 100 and 300 times
twloa_wf <- data.frame(word = names(twloa_freq), freq = twloa_freq)
head(twloa_wf)

##          word freq
## aacsb   aacsb    2
## abil     abil    9
## abl       abl   53
## abnorm abnorm    1
## abroad abroad   26
## abrupt abrupt    1
```





There are 40 words that are used at least 100 times; these including *work*, *health*, *program*, and *conflict*. There are only six words that are used at least 300 times; one of the most troubling for any university is *transfer*, because no institution wants to lose students to another institution. This can also be reviewed using word clouds:

```
## Warning in wordcloud(names(twloa_freq), twloa_freq, min.freq = 100, scale
=
## c(5, : person could not be fit on page. It will not be plotted.
```

This clearly shows that when I student leaves the university, their main focus is to work. This could be due to personal reasons, financial reason, or institutional unhappiness. We can now review specific terms with TF-IDF and LSA word associations and the Cosine Similarity of two frequent terms. The first term will be *financi*:

```
#Applying TF-IDF, LSA, and cosine distance to specific terms in the dataset
#Reviewing the term 'financi'
#TF-IDF
twloa_dtm <- DocumentTermMatrix(twloa_corpus, control = list(weighting = weig
htTfIdf(twloa_dtm, normalize = F)))
findAssocs(twloa_dtm, "financi", 0.5)

## $financi
##     aid  packag suffici  situat  afford  expens
##    0.83    0.80    0.72    0.68    0.62    0.61

inspect(DocumentTermMatrix(twloa_corpus, control = list(weighting = weightTfI
df(twloa_dtm, normalize = F), dictionary = c("financi"))))

## <<DocumentTermMatrix (documents: 905, terms: 1)>>
## Non-/sparse entries: 275/630
## Sparsity           : 70%
## Maximal term length: 7
## Sample             :
##     Terms
## Docs financi
##   12       1
##   14       1
##   26       1
##   29       1
##   3        1
##   31       1
##   49       1
```

```
##   5        1
##   6        1
##   8        1

#LSA
twloa_tdm <- TermDocumentMatrix(twloa_corpus)
twloa_lsa <- lsa(twloa_tdm, dims = dimcalc_share())

## Warning in lsa(twloa_tdm, dims = dimcalc_share()): [lsa] - there are
## singular values which are zero.

twloa_matrix <- as.textmatrix(twloa_lsa)
as.textmatrix(twloa_lsa)

## $matrix
##                   D1    D2    D3    D4    D5    D6    D7    D8    D9   D10
## 1. aacsb       -0.01  0.00  0.00 -0.01  0.00  0.00  0.00  0.01  0.02  0.00
## 2. abil         0.01 -0.12  0.01  0.06 -0.02  0.00  0.00  0.08  0.02  0.01
## 3. abl         -0.06  0.26 -0.03 -0.15  0.03  0.02  0.00 -0.02 -0.09  0.02
## 4. abnorm       0.03  0.00  0.00  0.08  0.00  0.00  0.00 -0.02 -0.01  0.00
## 5. abroad      -0.09 -0.32  0.01  0.06 -0.01  0.01  0.00  0.08  0.08  0.00
## 6. abrupt      -0.01  0.01  0.00  0.00  0.00  0.00  0.00 -0.01 -0.01  0.00
## 7. absenc      -0.01 -0.12  0.01 -0.06 -0.02 -0.02 -0.01  0.03 -0.01 -0.02
## 8. absolut     -0.05  0.10  0.00 -0.05 -0.03  0.02  0.03 -0.05  0.05 -0.02
## 9. absurd       0.00  0.01  0.00  0.01  0.00  0.00  0.03 -0.02  0.01  0.00
## 10. abund       0.05 -0.02  0.00  0.00  0.00 -0.01  0.00 -0.01 -0.03  0.00
## 11. abus        0.02  0.05 -0.01 -0.17 -0.01  0.00  0.00 -0.04  0.10 -0.01
## 12. academ     -0.15 -0.03 -0.01 -0.01  0.00  0.01 -0.01  0.91  0.09  0.01
## 1110. legist    0.01 -0.03  0.00 -0.03  0.00  0.01  0.00  0.01  0.01  0.00
## 1111. leo       0.00 -0.01  0.00 -0.01  0.01  0.00  0.00 -0.01  0.00  0.00
## 1112. leonard   0.02  0.05  0.00  0.00  0.00  0.00  0.00  0.01  0.00  0.00
## 1113. les       0.01  0.01  0.00  0.04  0.00  0.00  0.00  0.00 -0.01  0.00
## 1114. lesbian  -0.01 -0.02  0.00  0.00  0.00  0.00  0.00 -0.01 -0.02  0.00
## 1115. less     -0.02  0.07  0.01 -0.01  0.01 -0.01  0.00 -0.02  0.05  0.00
## 1116. lesson    0.03 -0.04  0.00  0.07  0.00  0.00  0.00  0.00 -0.01  0.00
## 1117. let      -0.06  0.05 -0.01 -0.05 -0.01 -0.01  0.00  0.01  0.00  0.01
## 1118. letter    0.01  0.04  0.01  0.01 -0.01  0.01  0.00  0.01 -0.02  0.00
## 1119. level    -0.06  0.08 -0.01 -0.03  0.00 -0.01 -0.01 -0.03  0.09  0.00
## 1120. lewi      0.03 -0.15 -0.01 -0.08  0.00  0.00  0.00  0.00  0.02  0.00
## 1121. lgbtq     0.00  0.04  0.00 -0.01  0.00  0.00  0.00  0.02 -0.02  0.00
## 2209. yellow   -0.01  0.00  0.00 -0.01  0.00  0.00  0.00  0.02 -0.01  0.00
## 2210. yes       0.01 -0.05  0.01 -0.07  0.02  0.00  0.00 -0.03  0.08  0.01
## 2211. yet      -0.01  0.00  0.01 -0.03  0.02  0.01  0.00  0.06  0.07  0.01
## 2212. youll     0.00 -0.01  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00
## 2213. young     0.01  0.02  0.00  0.03  0.00  0.00  0.00  0.00  0.00  0.01
## 2214. younger   0.01 -0.02  0.02 -0.01  0.01  0.01  0.00  0.00  0.00  0.00
## 2215. youth    -0.01 -0.02  0.00 -0.01  0.00  0.00  0.00 -0.01  0.01  0.00
## 2216. youtub    0.02  0.01  0.00  0.00 -0.01  0.00  0.00 -0.01  0.05  0.00
## 2217. youv      0.01  0.00  0.00  0.03 -0.01  0.00  0.00  0.06  0.00  0.00
## 2218. yunnan   -0.01  0.00  0.00  0.01  0.00  0.00  0.00  0.00  0.00  0.00
## 2219. ywam     -0.01 -0.02  0.00 -0.01  0.00  0.00  0.00 -0.01  0.01  0.00
```

```
## 2220. zero    -0.03 -0.04  0.01 -0.04  0.00  0.00  0.00 -0.03 -0.03 -0.01
##                D452  D453  D454  D455  D456  D457  D458  D459  D460  D461
## 1. aacsb       0.00  0.01  0.01  0.00  0.00  0.03  0.00  0.00  0.01  0.00
## 2. abil       -0.01  0.00  0.00 -0.02 -0.01  0.02 -0.02 -0.07  0.01 -0.02
## 3. abl        -0.12 -0.01  0.00 -0.05 -0.04 -0.01  0.02  0.12  0.04  0.13
## 4. abnorm      0.00  0.01  0.00  0.00 -0.01  0.01  0.01 -0.02 -0.01  0.02
## 5. abroad     -0.05  0.01  0.08  0.03  0.02 -0.05  0.04 -0.06  0.03  0.01
## 6. abrupt     -0.01  0.00  0.00  0.00  0.00  0.01  0.00  0.01 -0.02  0.00
## 7. absenc      0.03  0.01  0.02  0.00 -0.03  0.02  0.00 -0.01 -0.02 -0.07
## 8. absolut     0.04 -0.01  0.02  0.00 -0.02  0.13  0.03  0.05  0.00 -0.04
## 9. absurd      0.00 -0.01  0.01  0.00  0.01  0.03  0.00  0.00  0.01 -0.01
## 10. abund      0.01  0.00 -0.01 -0.01  0.00 -0.02 -0.01 -0.01  0.02  0.02
## 11. abus      -0.03  0.01  0.00  0.02 -0.01 -0.05  0.05  0.13  0.13 -0.04
## 12. academ    -0.02 -0.03  0.01  0.01  1.97  0.92  0.02 -0.02  0.93  1.04
## 1110. legist   0.00  0.00  0.00  0.00  0.02  0.01  0.00  0.04  0.03 -0.02
## 1111. leo     -0.01  0.00 -0.01  0.00  0.00  0.00  0.01  0.00  0.00  0.00
## 1112. leonard -0.01  0.01  0.04  0.01  0.00  0.01  0.00  0.05  0.02 -0.01
## 1113. les      0.00  0.00  0.00  0.00  0.01  0.00  0.00 -0.01  0.00  0.00
## 1114. lesbian  0.01 -0.01  0.01  0.01  0.01  0.04  0.00  0.00  0.00 -0.01
## 1115. less    -0.01  0.00 -0.04  0.00  0.00  0.02  0.00  0.00  0.00 -0.05
## 1116. lesson  -0.01  0.01  0.00  0.01  0.00 -0.03  0.00 -0.01 -0.01  0.02
## 1117. let      0.03  0.01  0.05 -0.04 -0.02  0.08  0.00 -0.01 -0.05 -0.01
## 1118. letter   0.03  0.00  0.01  0.00  0.01  0.00  0.00  0.01 -0.01 -0.01
## 1119. level   -0.02  0.00  0.00  0.01 -0.04 -0.03 -0.02  0.14  0.15  0.03
## 1120. lewi     0.03 -0.02  0.00  0.04  0.04 -0.06 -0.01  0.02  0.02  0.14
## 1121. lgbtq    0.00  0.00 -0.02 -0.01  0.00  0.03 -0.01  0.00 -0.01  0.01
## 2209. yellow  -0.01  0.00  0.01 -0.01  0.00  0.01  0.01  0.03 -0.01  0.01
## 2210. yes     -0.02  0.00  0.00  0.02 -0.02  0.00  0.00  0.10  0.00  0.01
## 2211. yet     -0.01  0.02  0.00 -0.01 -0.02  0.06 -0.01 -0.02  0.01  0.02
## 2212. youll    0.00  0.00  0.00  0.00 -0.01  0.00  0.00  0.01  0.01  0.01
## 2213. young    0.00  0.00  0.00  0.00  0.00 -0.01  0.00 -0.01  0.01  0.02
## 2214. younger  0.01  0.00  0.00 -0.01  0.00  0.00  0.00  0.05 -0.01  0.00
## 2215. youth    0.00  0.00 -0.01  0.00  0.00  0.01  0.00 -0.02 -0.01  0.01
## 2216. youtub   0.02  0.01  0.00  0.01  0.00 -0.04  0.00 -0.01 -0.01  0.00
## 2217. youv     0.03  0.00  0.03  0.00 -0.01  0.01  0.00  0.01  0.00  0.01
## 2218. yunnan   0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.01  0.00  0.00
## 2219. ywam     0.00  0.00 -0.01  0.00  0.00  0.01  0.00 -0.02 -0.01  0.01
## 2220. zero     0.03 -0.01  0.00  0.02  0.01  0.04  0.01  0.09 -0.01  0.00
##                D896  D897  D898  D899  D900  D901  D902  D903  D904  D905
## 1. aacsb       0.01  0.00 -0.02  0.00  0.00  0.00  0.02 -0.01 -0.02  0.00
## 2. abil       -0.04 -0.01  0.10  0.03 -0.02  0.00 -0.01 -0.05  0.14  0.11
## 3. abl        -0.01  0.00 -0.01  0.02  0.03 -0.03 -0.03  0.08 -0.05  0.77
## 4. abnorm      0.00  0.00 -0.01 -0.01  0.03  0.00 -0.01 -0.01 -0.01 -0.02
## 5. abroad     -0.01  0.00 -0.02 -0.01 -0.02  0.01  0.02  0.01  0.06 -0.03
## 6. abrupt      0.00  0.00 -0.01  0.00  0.02  0.00 -0.01  0.01 -0.02 -0.02
## 7. absenc      0.00  0.01  0.33  0.00  0.07  0.00 -0.02 -0.03  0.01  0.02
## 8. absolut     0.04  0.00  0.14  0.01 -0.01 -0.01  0.04 -0.06 -0.02 -0.01
## 9. absurd     -0.01  0.00  0.01  0.00  0.01  0.00  0.00  0.01 -0.03  0.00
## 10. abund     -0.01  0.00  0.04 -0.01  0.01  0.00  0.00  0.03  0.00 -0.02
## 11. abus       0.02 -0.01 -0.11 -0.04 -0.07 -0.03 -0.01 -0.20  0.10 -0.06
```

```
## 12. academ     1.86 -0.02  0.01 -0.01  0.02  0.00  0.03  0.97  0.05  0.00
## 1110. legist   0.00  0.00  0.02  0.01 -0.02 -0.02 -0.01 -0.02 -0.01 -0.01
## 1111. leo      0.00  0.01  0.00  0.00  0.00  0.00  0.00  0.00 -0.01  0.00
## 1112. leonard  0.00  0.01 -0.01  0.00 -0.01 -0.01  0.00 -0.01  0.02  0.01
## 1113. les      0.00  0.00  0.01  0.00  0.00  0.00 -0.01  0.00  0.00 -0.01
## 1114. lesbian  0.00  0.00 -0.03  0.00  0.00  0.00 -0.01  0.01  0.00  0.02
## 1115. less     0.00  0.02  0.15 -0.02  0.03 -0.01 -0.01 -0.03 -0.06 -0.02
## 1116. lesson  -0.01  0.00  0.01  0.00  0.00  0.00  0.00 -0.01 -0.02 -0.01
## 1117. let      0.00  0.00  0.05 -0.02  0.03  0.00  0.07  0.01  0.03  0.01
## 1118. letter  -0.01  0.00  0.02  0.00  0.01  0.00  0.01  0.00 -0.03  0.00
## 1119. level    0.01 -0.01  0.04  0.01  0.00 -0.01 -0.04 -0.03  0.83  0.06
## 1120. lewi     0.01 -0.01  0.11 -0.01  0.03  0.00  0.01  0.04 -0.01 -0.06
## 1121. lgbtq    0.00  0.00 -0.06  0.00 -0.01  0.00  0.01  0.00  0.02  0.01
## 2209. yellow  -0.01  0.00  0.02  0.00  0.00  0.00  0.01  0.01  0.00 -0.02
## 2210. yes      0.01  0.02  0.00 -0.01 -0.03  0.01 -0.04  0.05 -0.05  0.00
## 2211. yet      0.00  0.01 -0.10  0.01  0.02  0.02 -0.06  0.07  0.10  0.04
## 2212. youll    0.00  0.00  0.02  0.00  0.02  0.00  0.01  0.01 -0.01  0.00
## 2213. young    0.00  0.00  0.00  0.00  0.01  0.00  0.01 -0.02  0.02 -0.01
## 2214. younger -0.01  0.01  0.00  0.01 -0.01  0.00 -0.01 -0.02 -0.02  0.00
## 2215. youth   -0.01  0.00  0.00  0.00 -0.01  0.00  0.00  0.00  0.00  0.01
## 2216. youtub   0.00  0.00  0.03  0.01  0.06  0.00  0.01 -0.01  0.05  0.00
## 2217. youv     0.00  0.00  0.00  0.00  0.02  0.00  0.04 -0.01  0.00 -0.02
## 2218. yunnan   0.00  0.00 -0.01  0.00  0.00  0.00  0.00  0.02  0.00  0.00
## 2219. ywam    -0.01  0.00  0.00  0.00 -0.01  0.00  0.00  0.00  0.00  0.01
## 2220. zero     0.00  0.01  0.03  0.00 -0.01  0.00  0.02  0.00  0.03 -0.02
##
## $legend
##  [1] "D1 = 1"     "D2 = 2"     "D3 = 3"     "D4 = 4"     "D5 = 5"
##  [6] "D6 = 6"     "D7 = 7"     "D8 = 8"     "D9 = 9"     "D10 = 10"
## [11] "D452 = 452" "D453 = 453" "D454 = 454" "D455 = 455" "D456 = 456"
## [16] "D457 = 457" "D458 = 458" "D459 = 459" "D460 = 460" "D461 = 461"
## [21] "D896 = 896" "D897 = 897" "D898 = 898" "D899 = 899" "D900 = 900"
## [26] "D901 = 901" "D902 = 902" "D903 = 903" "D904 = 904" "D905 = 905"
```

```
associate(twloa_matrix, "financi", threshold = 0.5)
```

```
##       aid    packag   suffici    situat    afford    expens     chang
## 0.8798720 0.8405553 0.7702425 0.7458396 0.7185556 0.7081518 0.5721407
```

Both TF-IDF and LSA produce the same word associations, but their association percentages differ. The most highly associated term is *aid*, this makes sense since many students rely on financial aid to get the education and with Regis being an expensive institution, it would make it difficult to complete without enough financial aid. Now we will review the associations between the term *health*:

```
#Reviewing the term 'health'
#TF-IDF
twloa_dtm <- DocumentTermMatrix(twloa_corpus, control = list(weighting = weig
htTfIdf(twloa_dtm, normalize = F)))
findAssocs(twloa_dtm, "health", 0.5)
```

```
## $health
## medicalhealth        mental        physic
##          0.77          0.77          0.66
```

```
inspect(DocumentTermMatrix(twloa_corpus, control = list(weighting = weightTfI
df(twloa_dtm, normalize = F), dictionary = c("health"))))
```

```
## <<DocumentTermMatrix (documents: 905, terms: 1)>>
## Non-/sparse entries: 181/724
## Sparsity           : 80%
## Maximal term length: 6
## Sample             :
##     Terms
## Docs health
##   11      1
##   23      1
##   27      1
##   33      1
##   42      1
##   43      1
##   44      1
##   56      1
##   58      1
##   73      1
```

```
#LSA
twloa_tdm <- TermDocumentMatrix(twloa_corpus)
twloa_lsa <- lsa(twloa_tdm, dims = dimcalc_share())
```

```
## Warning in lsa(twloa_tdm, dims = dimcalc_share()): [lsa] - there are
## singular values which are zero.
```

```
twloa_matrix <- as.textmatrix(twloa_lsa)
as.textmatrix(twloa_lsa)
```

```
## $matrix
##                  D1    D2    D3    D4    D5    D6    D7    D8    D9   D10
## 1. aacsb      -0.01  0.00  0.00 -0.01  0.00  0.00  0.00  0.01  0.02  0.00
## 2. abil        0.01 -0.12  0.01  0.06 -0.02  0.00  0.00  0.08  0.02  0.01
## 3. abl        -0.06  0.26 -0.03 -0.15  0.03  0.02  0.00 -0.02 -0.09  0.02
## 4. abnorm      0.03  0.00  0.00  0.08  0.00  0.00  0.00 -0.02 -0.01  0.00
## 5. abroad     -0.09 -0.32  0.01  0.06 -0.01  0.01  0.00  0.08  0.08  0.00
## 6. abrupt     -0.01  0.01  0.00  0.00  0.00  0.00  0.00 -0.01 -0.01  0.00
## 7. absenc     -0.01 -0.12  0.01 -0.06 -0.02 -0.02 -0.01  0.03 -0.01 -0.02
## 8. absolut    -0.05  0.10  0.00 -0.05 -0.03  0.02  0.03 -0.05  0.05 -0.02
## 9. absurd      0.00  0.01  0.00  0.01  0.00  0.00  0.03 -0.02  0.01  0.00
## 10. abund      0.05 -0.02  0.00  0.00  0.00 -0.01  0.00 -0.01 -0.03  0.00
## 11. abus       0.02  0.05 -0.01 -0.17 -0.01  0.00  0.00 -0.04  0.10 -0.01
## 12. academ    -0.15 -0.03 -0.01 -0.01  0.00  0.01 -0.01  0.91  0.09  0.01
## 1110. legist   0.01 -0.03  0.00 -0.03  0.00  0.01  0.00  0.01  0.01  0.00
## 1111. leo      0.00 -0.01  0.00 -0.01  0.01  0.00  0.00 -0.01  0.00  0.00
```

```
## 1112. leonard  0.02  0.05  0.00  0.00  0.00  0.00  0.00  0.01  0.00  0.00
## 1113. les      0.01  0.01  0.00  0.04  0.00  0.00  0.00  0.00 -0.01  0.00
## 1114. lesbian -0.01 -0.02  0.00  0.00  0.00  0.00  0.00 -0.01 -0.02  0.00
## 1115. less    -0.02  0.07  0.01 -0.01  0.01 -0.01  0.00 -0.02  0.05  0.00
## 1116. lesson   0.03 -0.04  0.00  0.07  0.00  0.00  0.00  0.00 -0.01  0.00
## 1117. let     -0.06  0.05 -0.01 -0.05 -0.01 -0.01  0.00  0.01  0.00  0.01
## 1118. letter   0.01  0.04  0.01  0.01 -0.01  0.01  0.00  0.01 -0.02  0.00
## 1119. level   -0.06  0.08 -0.01 -0.03  0.00 -0.01 -0.01 -0.03  0.09  0.00
## 1120. lewi     0.03 -0.15 -0.01 -0.08  0.00  0.00  0.00  0.00  0.02  0.00
## 1121. lgbtq    0.00  0.04  0.00 -0.01  0.00  0.00  0.00  0.02 -0.02  0.00
## 2209. yellow  -0.01  0.00  0.00 -0.01  0.00  0.00  0.00  0.02 -0.01  0.00
## 2210. yes      0.01 -0.05  0.01 -0.07  0.02  0.00  0.00 -0.03  0.08  0.01
## 2211. yet     -0.01  0.00  0.01 -0.03  0.02  0.01  0.00  0.06  0.07  0.01
## 2212. youll    0.00 -0.01  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00
## 2213. young    0.01  0.02  0.00  0.03  0.00  0.00  0.00  0.00  0.00  0.01
## 2214. younger  0.01 -0.02  0.02 -0.01  0.01  0.01  0.00  0.00  0.00  0.00
## 2215. youth   -0.01 -0.02  0.00 -0.01  0.00  0.00  0.00 -0.01  0.01  0.00
## 2216. youtub   0.02  0.01  0.00  0.00 -0.01  0.00  0.00 -0.01  0.05  0.00
## 2217. youv     0.01  0.00  0.00  0.03 -0.01  0.00  0.00  0.06  0.00  0.00
## 2218. yunnan  -0.01  0.00  0.00  0.01  0.00  0.00  0.00  0.00  0.00  0.00
## 2219. ywam    -0.01 -0.02  0.00 -0.01  0.00  0.00  0.00 -0.01  0.01  0.00
## 2220. zero    -0.03 -0.04  0.01 -0.04  0.00  0.00  0.00 -0.03 -0.03 -0.01
##                D452  D453  D454  D455  D456  D457  D458  D459  D460  D461
## 1. aacsb       0.00  0.01  0.01  0.00  0.00  0.03  0.00  0.00  0.01  0.00
## 2. abil       -0.01  0.00  0.00 -0.02 -0.01  0.02 -0.02 -0.07  0.01 -0.02
## 3. abl        -0.12 -0.01  0.00 -0.05 -0.04 -0.01  0.02  0.12  0.04  0.13
## 4. abnorm      0.00  0.01  0.00  0.00 -0.01  0.01  0.01 -0.02 -0.01  0.02
## 5. abroad     -0.05  0.01  0.08  0.03  0.02 -0.05  0.04 -0.06  0.03  0.01
## 6. abrupt     -0.01  0.00  0.00  0.00  0.00  0.01  0.00  0.01 -0.02  0.00
## 7. absenc      0.03  0.01  0.02  0.00 -0.03  0.02  0.00 -0.01 -0.02 -0.07
## 8. absolut     0.04 -0.01  0.02  0.00 -0.02  0.13  0.03  0.05  0.00 -0.04
## 9. absurd      0.00 -0.01  0.01  0.00  0.01  0.03  0.00  0.00  0.01 -0.01
## 10. abund      0.01  0.00 -0.01 -0.01  0.00 -0.02 -0.01 -0.01  0.02  0.02
## 11. abus      -0.03  0.01  0.00  0.02 -0.01 -0.05  0.05  0.13  0.13 -0.04
## 12. academ    -0.02 -0.03  0.01  0.01  1.97  0.92  0.02 -0.02  0.93  1.04
## 1110. legist   0.00  0.00  0.00  0.00  0.02  0.01  0.00  0.04  0.03 -0.02
## 1111. leo     -0.01  0.00 -0.01  0.00  0.00  0.00  0.01  0.00  0.00  0.00
## 1112. leonard -0.01  0.01  0.04  0.01  0.00  0.01  0.00  0.05  0.02 -0.01
## 1113. les      0.00  0.00  0.00  0.00  0.01  0.00  0.00 -0.01  0.00  0.00
## 1114. lesbian  0.01 -0.01  0.01  0.01  0.01  0.04  0.00  0.00  0.00 -0.01
## 1115. less    -0.01  0.00 -0.04  0.00  0.00  0.02  0.00  0.00  0.00 -0.05
## 1116. lesson  -0.01  0.01  0.00  0.01  0.00 -0.03  0.00 -0.01 -0.01  0.02
## 1117. let      0.03  0.01  0.05 -0.04 -0.02  0.08  0.00 -0.01 -0.05 -0.01
## 1118. letter   0.03  0.00  0.01  0.00  0.01  0.00  0.00  0.01 -0.01 -0.01
## 1119. level   -0.02  0.00  0.00  0.01 -0.04 -0.03 -0.02  0.14  0.15  0.03
## 1120. lewi     0.03 -0.02  0.00  0.04  0.04 -0.06 -0.01  0.02  0.02  0.14
## 1121. lgbtq    0.00  0.00 -0.02 -0.01  0.00  0.03 -0.01  0.00 -0.01  0.01
## 2209. yellow  -0.01  0.00  0.01 -0.01  0.00  0.01  0.01  0.03 -0.01  0.01
## 2210. yes     -0.02  0.00  0.00  0.02 -0.02  0.00  0.00  0.10  0.00  0.01
## 2211. yet     -0.01  0.02  0.00 -0.01 -0.02  0.06 -0.01 -0.02  0.01  0.02
```

```
## 2212. youll    0.00  0.00  0.00  0.00 -0.01  0.00  0.00  0.01  0.01  0.01
## 2213. young    0.00  0.00  0.00  0.00  0.00 -0.01  0.00 -0.01  0.01  0.02
## 2214. younger  0.01  0.00  0.00 -0.01  0.00  0.00  0.00  0.05 -0.01  0.00
## 2215. youth    0.00  0.00 -0.01  0.00  0.00  0.01  0.00 -0.02 -0.01  0.01
## 2216. youtub   0.02  0.01  0.00  0.01  0.00 -0.04  0.00 -0.01 -0.01  0.00
## 2217. youv     0.03  0.00  0.03  0.00 -0.01  0.01  0.00  0.01  0.00  0.01
## 2218. yunnan   0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.01  0.00  0.00
## 2219. ywam     0.00  0.00 -0.01  0.00  0.00  0.01  0.00 -0.02 -0.01  0.01
## 2220. zero     0.03 -0.01  0.00  0.02  0.01  0.04  0.01  0.09 -0.01  0.00
##                D896  D897  D898  D899  D900  D901  D902  D903  D904  D905
## 1. aacsb       0.01  0.00 -0.02  0.00  0.00  0.00  0.02 -0.01 -0.02  0.00
## 2. abil       -0.04 -0.01  0.10  0.03 -0.02  0.00 -0.01 -0.05  0.14  0.11
## 3. abl        -0.01  0.00 -0.01  0.02  0.03 -0.03 -0.03  0.08 -0.05  0.77
## 4. abnorm      0.00  0.00 -0.01 -0.01  0.03  0.00 -0.01 -0.01 -0.01 -0.02
## 5. abroad     -0.01  0.00 -0.02 -0.01 -0.02  0.01  0.02  0.01  0.06 -0.03
## 6. abrupt      0.00  0.00 -0.01  0.00  0.02  0.00 -0.01  0.01 -0.02 -0.02
## 7. absenc      0.00  0.01  0.33  0.00  0.07  0.00 -0.02 -0.03  0.01  0.02
## 8. absolut     0.04  0.00  0.14  0.01 -0.01 -0.01  0.04 -0.06 -0.02 -0.01
## 9. absurd     -0.01  0.00  0.01  0.00  0.01  0.00  0.00  0.01 -0.03  0.00
## 10. abund     -0.01  0.00  0.04 -0.01  0.01  0.00  0.00  0.03  0.00 -0.02
## 11. abus       0.02 -0.01 -0.11 -0.04 -0.07 -0.03 -0.01 -0.20  0.10 -0.06
## 12. academ     1.86 -0.02  0.01 -0.01  0.02  0.00  0.03  0.97  0.05  0.00
## 1110. legist   0.00  0.00  0.02  0.01 -0.02 -0.02 -0.01 -0.02 -0.01 -0.01
## 1111. leo      0.00  0.01  0.00  0.00  0.00  0.00  0.00  0.00 -0.01  0.00
## 1112. leonard  0.00  0.01 -0.01  0.00 -0.01 -0.01  0.00 -0.01  0.02  0.01
## 1113. les      0.00  0.00  0.01  0.00  0.00  0.00 -0.01  0.00  0.00 -0.01
## 1114. lesbian  0.00  0.00 -0.03  0.00  0.00  0.00 -0.01  0.01  0.00  0.02
## 1115. less     0.00  0.02  0.15 -0.02  0.03 -0.01 -0.01 -0.03 -0.06 -0.02
## 1116. lesson  -0.01  0.00  0.01  0.00  0.00  0.00  0.00 -0.01 -0.02 -0.01
## 1117. let      0.00  0.00  0.05 -0.02  0.03  0.00  0.07  0.01  0.03  0.01
## 1118. letter  -0.01  0.00  0.02  0.00  0.01  0.00  0.01  0.00 -0.03  0.00
## 1119. level    0.01 -0.01  0.04  0.01  0.00 -0.01 -0.04 -0.03  0.83  0.06
## 1120. lewi     0.01 -0.01  0.11 -0.01  0.03  0.00  0.01  0.04 -0.01 -0.06
## 1121. lgbtq    0.00  0.00 -0.06  0.00 -0.01  0.00  0.01  0.00  0.02  0.01
## 2209. yellow  -0.01  0.00  0.02  0.00  0.00  0.00  0.01  0.01  0.00 -0.02
## 2210. yes      0.01  0.02  0.00 -0.01 -0.03  0.01 -0.04  0.05 -0.05  0.00
## 2211. yet      0.00  0.01 -0.10  0.01  0.02  0.02 -0.06  0.07  0.10  0.04
## 2212. youll    0.00  0.00  0.02  0.00  0.02  0.00  0.01  0.01 -0.01  0.00
## 2213. young    0.00  0.00  0.00  0.00  0.01  0.00  0.01 -0.02  0.02 -0.01
## 2214. younger -0.01  0.01  0.00  0.01 -0.01  0.00 -0.01 -0.02 -0.02  0.00
## 2215. youth   -0.01  0.00  0.00  0.00 -0.01  0.00  0.00  0.00  0.00  0.01
## 2216. youtub   0.00  0.00  0.03  0.01  0.06  0.00  0.01 -0.01  0.05  0.00
## 2217. youv     0.00  0.00  0.00  0.00  0.02  0.00  0.04 -0.01  0.00 -0.02
## 2218. yunnan   0.00  0.00 -0.01  0.00  0.00  0.00  0.00  0.02  0.00  0.00
## 2219. ywam    -0.01  0.00  0.00  0.00 -0.01  0.00  0.00  0.00  0.00  0.01
## 2220. zero     0.00  0.01  0.03  0.00 -0.01  0.00  0.02  0.00  0.03 -0.02
##
## $legend
##  [1] "D1 = 1"     "D2 = 2"     "D3 = 3"     "D4 = 4"     "D5 = 5"
##  [6] "D6 = 6"     "D7 = 7"     "D8 = 8"     "D9 = 9"     "D10 = 10"
```

```
## [11] "D452 = 452" "D453 = 453" "D454 = 454" "D455 = 455" "D456 = 456"
## [16] "D457 = 457" "D458 = 458" "D459 = 459" "D460 = 460" "D461 = 461"
## [21] "D896 = 896" "D897 = 897" "D898 = 898" "D899 = 899" "D900 = 900"
## [26] "D901 = 901" "D902 = 902" "D903 = 903" "D904 = 904" "D905 = 905"
```

```
associate(twloa_matrix, "health", threshold = 0.5)
```

```
## medicalhealth        mental        physic       chronic campusunivers
##     0.8388170     0.8222919     0.7285057     0.5381152     0.5249631
```

The TF-IDF analysis shows only three associations, while the LSA shows five; once again the LSA analysis appears to be more thorough and shows higher rates of association. The main *health* associated term is *medicalhealth*, which implies that students leave the university due to physical health issues. This is heartbreaking and I wonder if there is something we could do to help assist them through this process. Finally we can review the Cosine Similarity between *academ* and *financi*:

```
#Looking at how 'academ' and 'financi' are related with Cosine
Cosine("academ", "financi", tvectors = twloa_matrix)
```

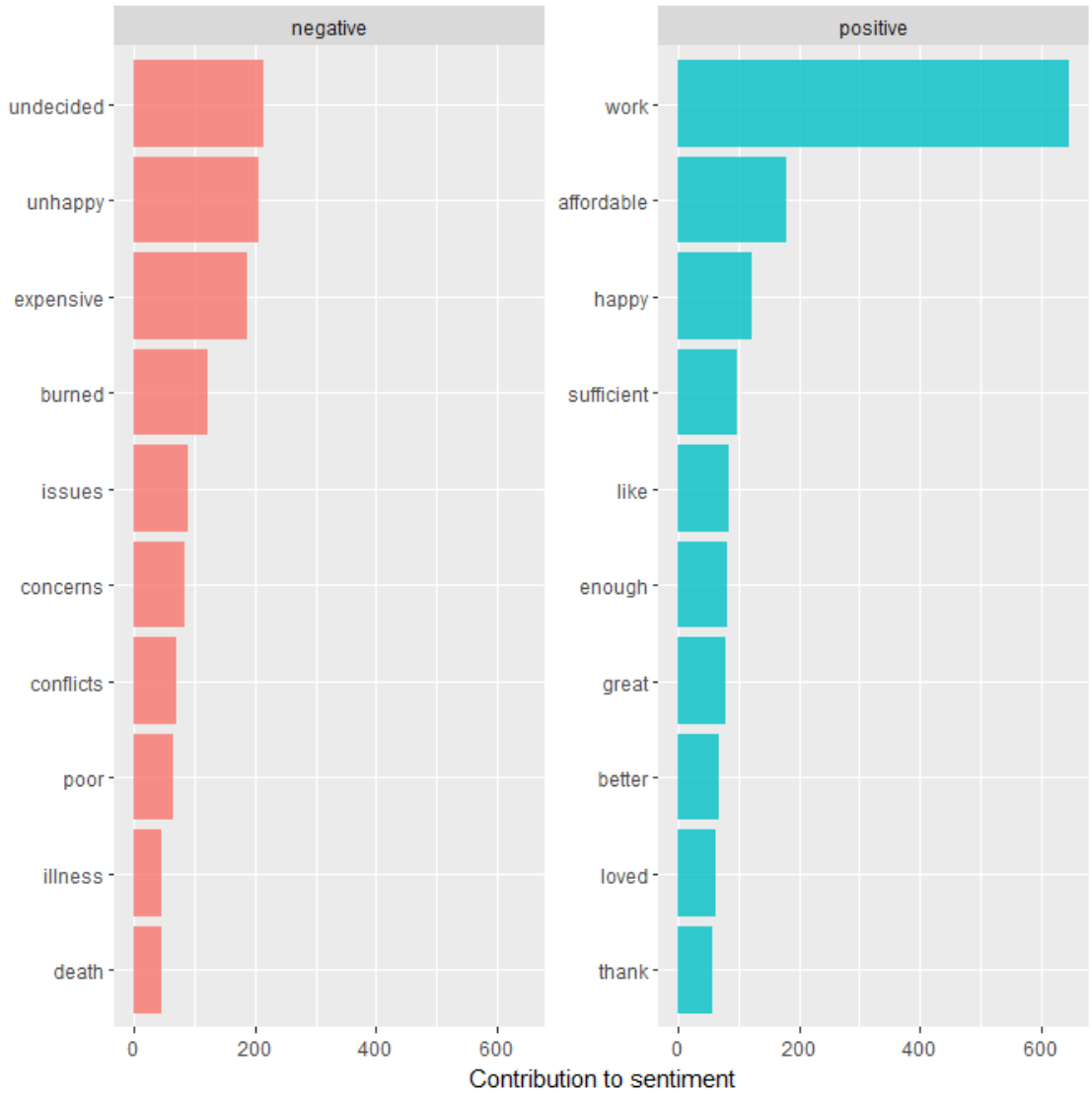
```
## [1] 0.2481606
```

These results are rather surprising, since I expected there to be a closer relationship between the two. However, since this is a form that students submit on their own, there is still a possibility that the two are related. It would be interesting to dive deeper and determine if student who leave the university for financial reasons are also experience academic hardship. The last bit of text analytics utilizes bing and nrc sentiment analysis. We will be able to interpret the sentiment behind the words student are using on the form. We will first review the bing sentiment:

```
#Sentiment analysis
twloa_txt <- twloa_full[, c(12:12)]
twloa_txt <- trimws(twloa_txt)
twloa_txt <- gsub("\\$", "",twloa_txt)
twloa_tokens <- data_frame(text = twloa_txt) %>% unnest_tokens(word, text)

bing_sent <- twloa_tokens %>%
  inner_join(get_sentiments("bing")) %>% # pull out only sentiment words
  count(sentiment) %>% # count the # of positive & negative words
  spread(sentiment, n, fill = 0) %>% # made data wide rather than narrow
  mutate(sentiment = positive - negative) # # of positive words - # of negati
ve words
bing_sent
```

```
## # A tibble: 1 x 3
##   negative positive sentiment
##      <dbl>    <dbl>     <dbl>
## 1     1853     2265       412
```

Within the dataset there are 1853 negative words and 2265 positive words. We can now review a plot of the bing sentiments:



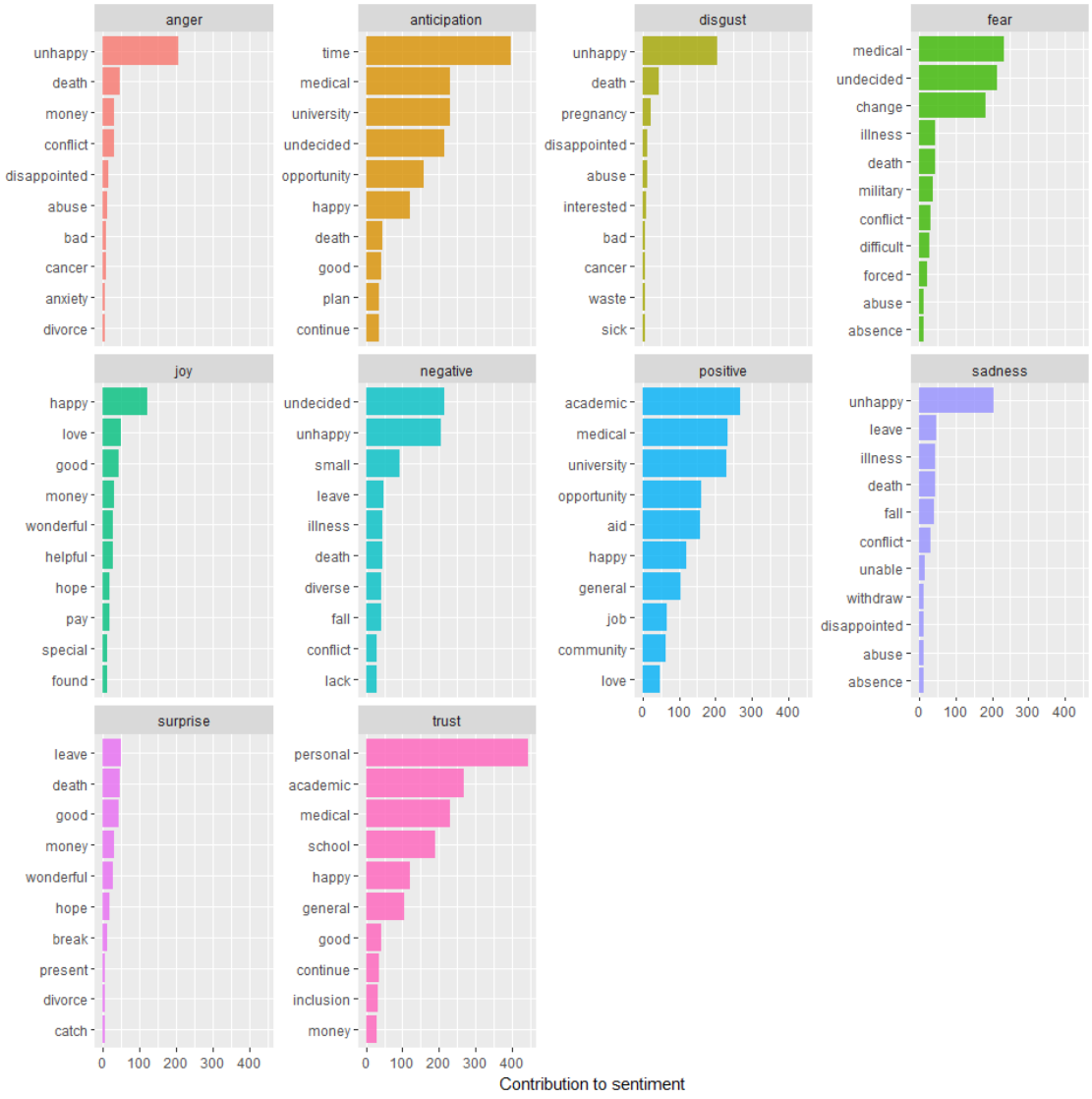
Looking at these two bar charts, the negative words clearly fit, however, I believe there may be some false-positives present. Leaving school to go *work* without completing your degree does not sound positive. Being that *sufficient* was also highly associated with *financi*, this is actual a negative term since many students indicate that their aid package is not sufficient enough for them to complete their program. Let's review the nrc sentiments to see if the same false-positives are present:

```
nrc_sent <- twloa_tokens %>%
  right_join(get_sentiments("nrc")) %>%
  filter(!is.na(sentiment)) %>%
  count(sentiment, sort = TRUE)
nrc_sent
```

```
## # A tibble: 10 x 2
##       sentiment     n
##           <chr> <int>
```

```
##  1     positive  4992
##  2     negative  4457
##  3        trust  3203
##  4 anticipation  2619
##  5         fear  2415
##  6      sadness  1835
##  7        anger  1639
##  8      disgust  1411
##  9          joy  1186
## 10     surprise   796
```

Here we can see there are 4992 positive terms, 4457 negative, and 3203 trust terms. Looking at the plot could further enhance our understanding of the terms used:



It appears once again there are terms that are being placed in false categories. The term *personal* is in the trust category, which is pink, which implies a positive sense of trust, but in all actuality, many students state personal reasons for leaving, without elaborating,

which shows a lack of trust in the university. Similarly, with *academic*, when I student leaves for academic reasons this is rarely a positive thing; it could either mean they are leaving because they are unhappy with the academics or because the university was unhappy with their academics. In the future, it would be helpful to perform sentiment analysis on *n*-grams from the forms in order to greater understand the full sentiment of the statements being made.

Conclusion

Working on this project has been extremely insightful and has aided me in gaining a greater understanding of the TWLOA process for students. We now understand which students are leaving which colleges, where these students are from, and what words they are using. Ultimately, this project is barely scratching the surface of the TWLOA analysis possible and I am looking forward to diving deeper in. It would be extremely helpful to understand in more detail about the students and find ways to relate the text analytics to the EDA. In regards to the text analytics, it is clear that certain terms are generating false-positives and a more involved sentiment analysis will need to be completed in order to fully understand the sentiments of withdrawn students. I would also like to perform predictive analytics to help prevent student withdrawals in the future, but that is another project for another time.

References

ilir. (2014, April 15). *Only displaying the top 3 bars in a ggplot2 chart*. Retrieved from https://stackoverflow.com/questions/23095129/only-displaying-top-3-bars-in-a-ggplot2-chart

Machlis, S. (2015, June 17). *My ggplot2 cheat sheet: Search by task*. Retrieved from https://www.computerworld.com/article/2935394/business-intelligence/my-ggplot2-cheat-sheet-search-by-task.html

rafa.pereira. (2016, April 21). *Eliminating NAs from a ggplot*. Retrieved from https://stackoverflow.com/questions/17216358/eliminating-nas-from-a-ggplot

Tatman, R. (2017, October 5). *Data science 101: Sentiment analysis in R tutorial*. Retrieved from http://blog.kaggle.com/2017/10/05/data-science-101-sentiment-analysis-in-r-tutorial/

Ulrich, J. (2012, April 10). *Extracting specific columns from a data frame*. Retrieved from https://stackoverflow.com/questions/10085806/extracting-specific-columns-from-a-data-frame