Text Analysis

Business Insight Report

Submitted To: Thomas Kurnicki

Submitted By: Neil Parekh

Student Id: 4647284

MSBA- 5

Analysis for Masters in Business Analytics

This text analysis is based on the course structure designed by Hult International business school. The course structure designed is aligning towards the current industry needs and what the market requires from New graduates. The course enriches students with the perfect combination of hard skills and soft skills. This report is designed to derive insights that will benefit student and give a better direction to their career path.

Business Insights:

Hult has catered all its subjects to industry standards. Each industry requires data to give consumer insight and Hult helps its students to hone their skills by giving them a lot of data driven cases and how to analyse them effectively.

One of the most compelling insights that I came across was the design of the course structure that Hult is created for its students.

Hult has curated the course structure of the MBAN program proving it beneficial for the Technical as well as non-technical background students.

Subjects like Data Strategy, Digital Visualization, Critical analysis, Digital marketing strategy may not be technical subjects but they focus more on the customers and stakeholders' side of the organisation but they are equally important as much as any technical subject like Python, R, Machine learning which focuses on the problems that the customers are facing and also to ensure that the customer needs are satisfied.

Another business insight that amazed me was the fact that this course bridges the gap between IT and business, students at hult are trained to do the job of a data scientist as well as a business professional which can solve the organization issue as well as understand the customer problem and solve it accordingly with the help of data. This will help the organisation in cutting cost by hiring someone who can thrive in different fields with just a single degree.

Observation:

While doing the frequency histogram subject wise, there were a lot of words which were matching the other subjects key words like strategy, analysis, learn, data that give a student a holistic idea of how the business works in an organisation and also the importance of these skills in the market.

While doing NRC sentiment analysis , Most of the words were inclined towards positive , trust, surprise, anticipation and joy and less on the words like negative, disgust. The word “Data” kept appearing over and over again proving the importance of it in any field.

Framework Used:

This analysis is useful for the students coming to Hult International Business School to do their master’s in Business analytics and to make them understand which subjects are being useful to choose a particular stream based on the course structure designed by the school.

The data was being extracted from the Hult International Business School website in R and was processed using the following framework:

- Stop Words
- Tokenization
- Frequency Histogram
- NRC Sentiment
- Bing sentiment
- Word Cloud
- Bigrams
- Quadrograms

The above framework gives an exact estimate as to how a particular student can take up these courses based on their goal. The main aim of the organisation is to give a holistic approach to its students by giving them enough exposure to live case studies and hands on experience on the industry.

I also used biagram and quadrogram out of which quadrogram was more beneficial for this particular analysis, as it gave better insights with deeper

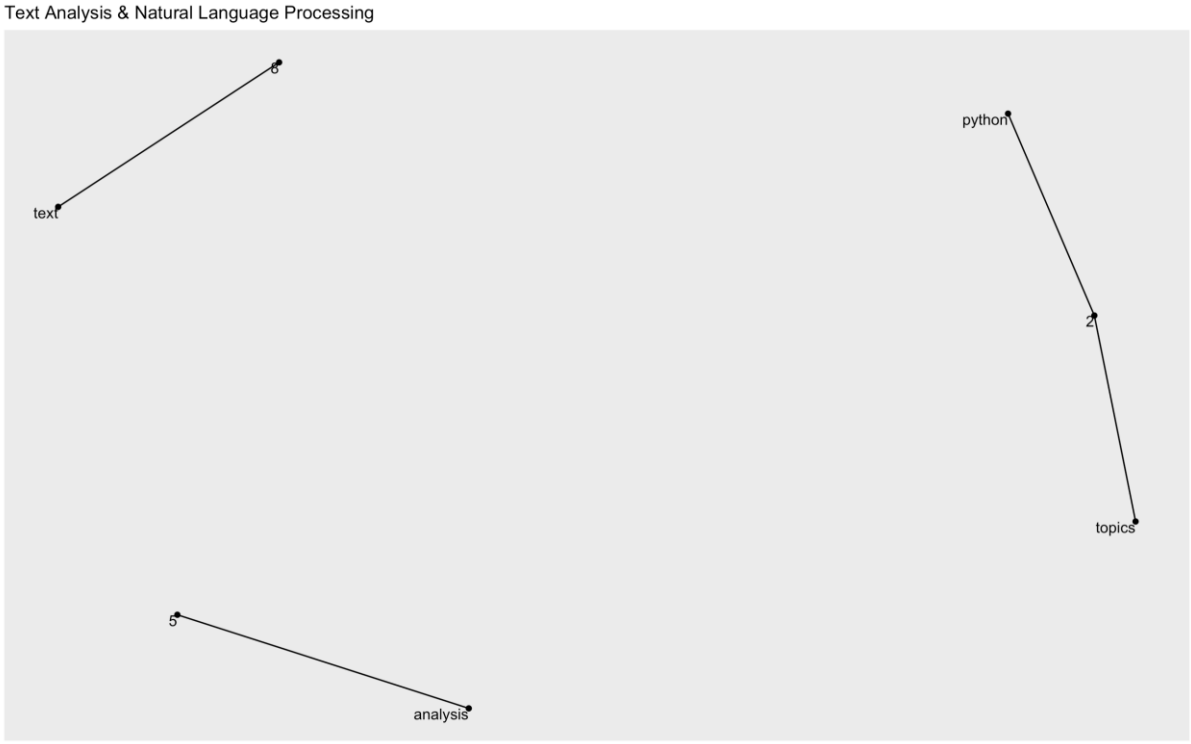
analysis. For example, In Machine learning when I did NRC sentiment, Words like "Intelligence" is derived as "Joy , Fear and Positive".

I used the AFINN to derive the mean value of every subject in order to understand the numerical side of each subject. Most of the subjects are in positive except for Data Science Python and R subject which had a negative mean of (-1) and (-0.3) respectively

I designed a word cloud using sentiment analysis("nrc") which helped me classify words into different sentiments that in return helped devlop my insight.

Overall , from this analysis I believe that every course designed by hult is equally important to thrive in the market . A combination of hard and soft skills is the best way to grow in the market .

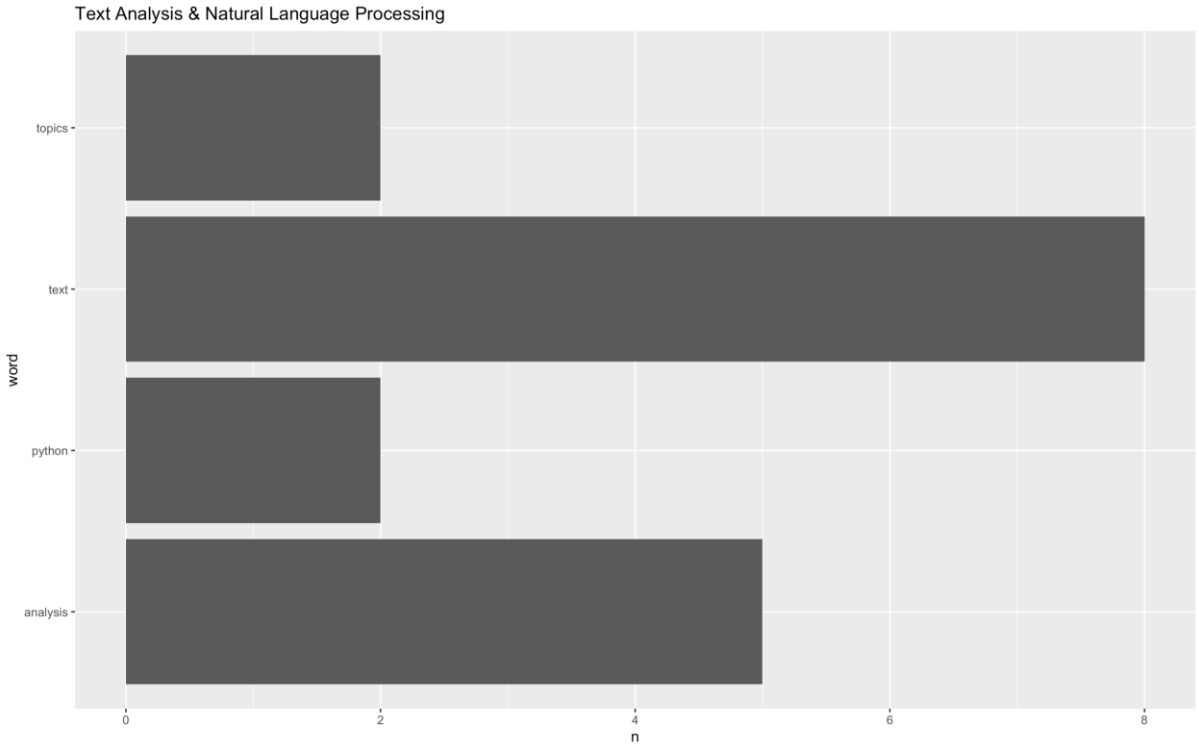
Appendix:



Text Analysis & Natural Language Processing

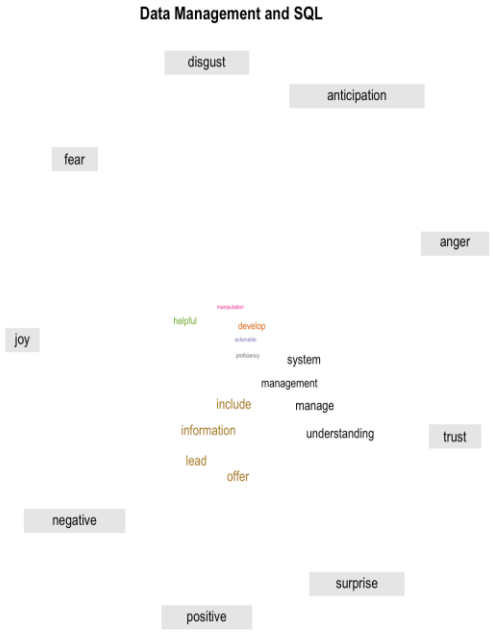
positive

trust

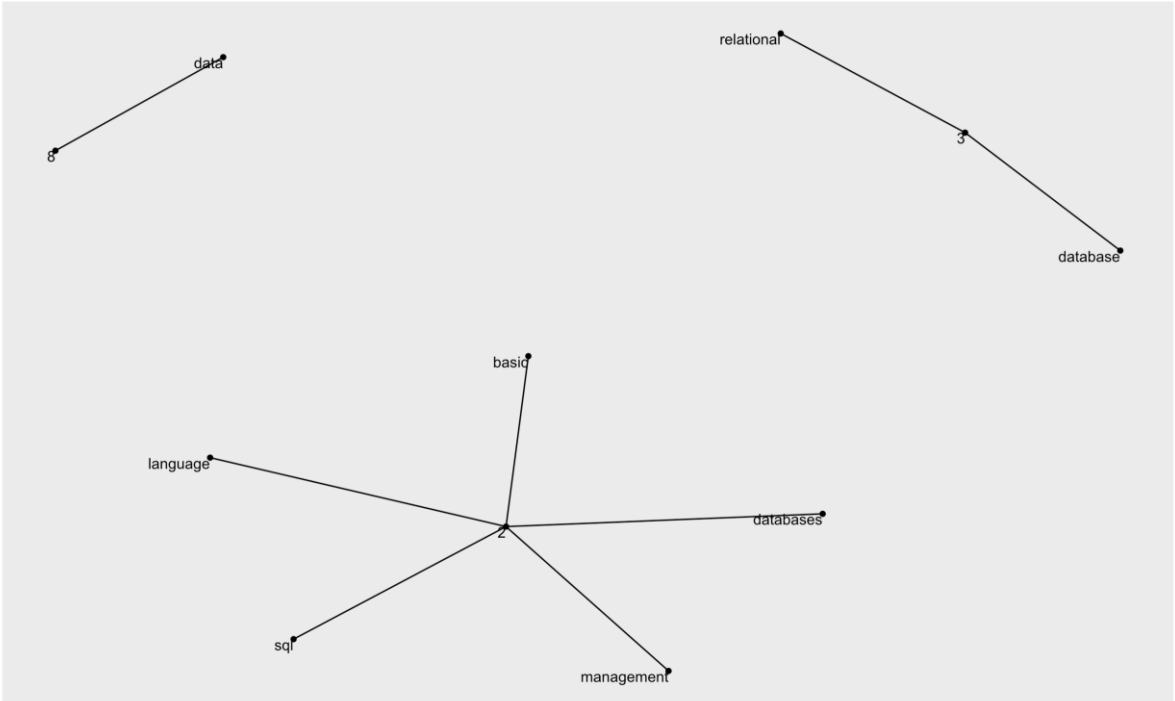


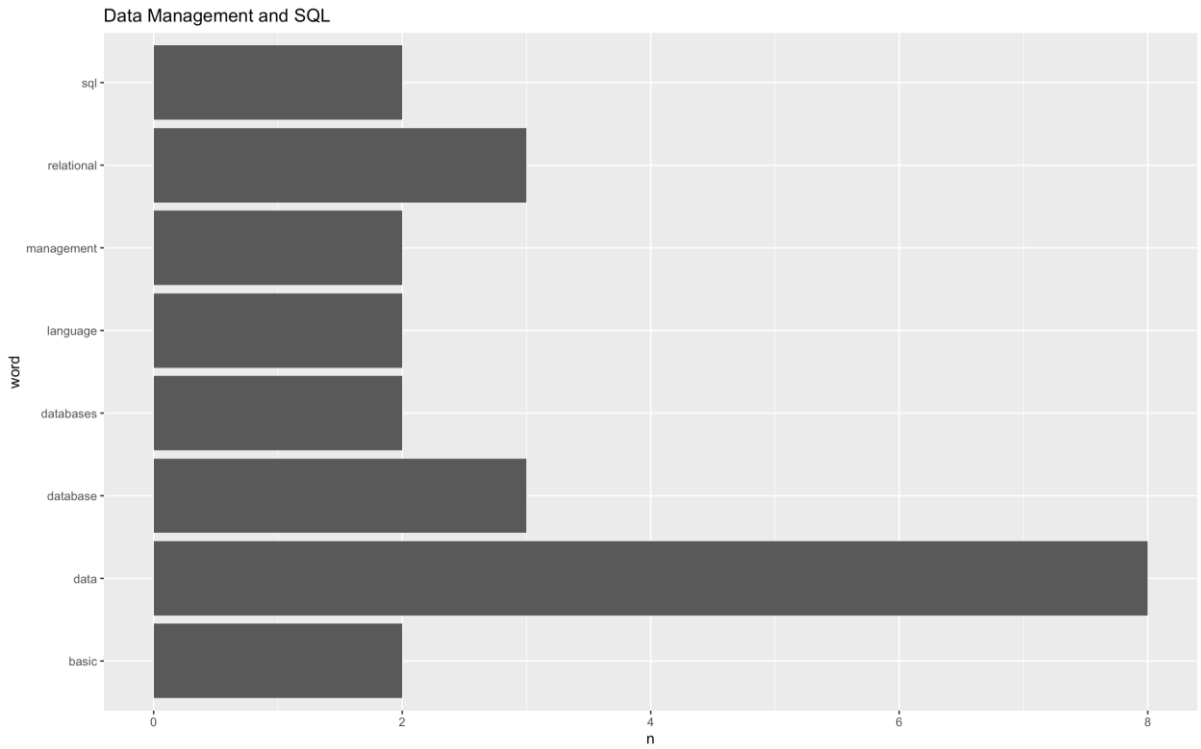
	line	subject	word	sentiment
1	1	Text Analysis & Natural Language Processing	include	positive
2	1	Text Analysis & Natural Language Processing	reading	positive
3	1	Text Analysis & Natural Language Processing	learn	positive
4	1	Text Analysis & Natural Language Processing	effective	positive
5	1	Text Analysis & Natural Language Processing	effective	trust
6	1	Text Analysis & Natural Language Processing	statistical	trust
7	1	Text Analysis & Natural Language Processing	provide	positive
8	1	Text Analysis & Natural Language Processing	provide	trust

Showing 1 to 8 of 8 entries, 4 total columns



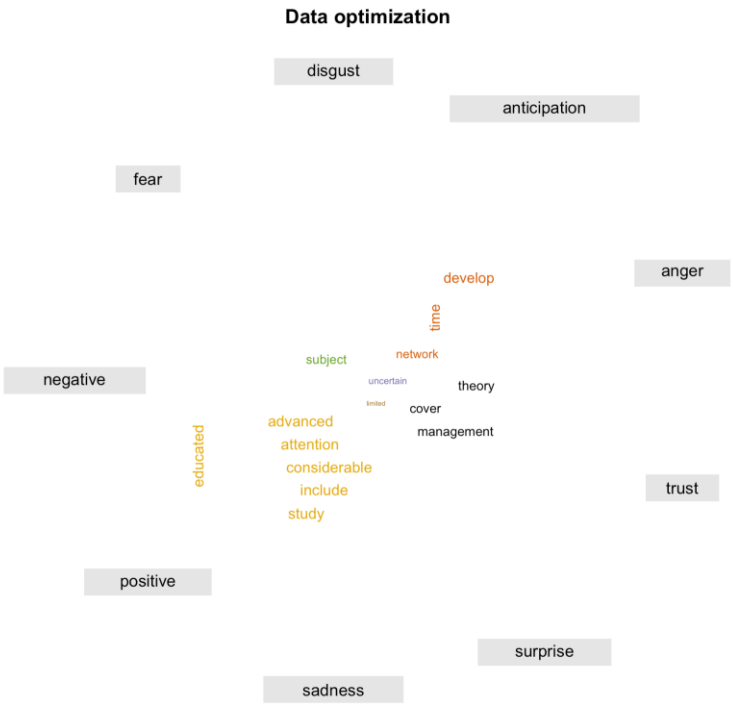
Data Management and SQL



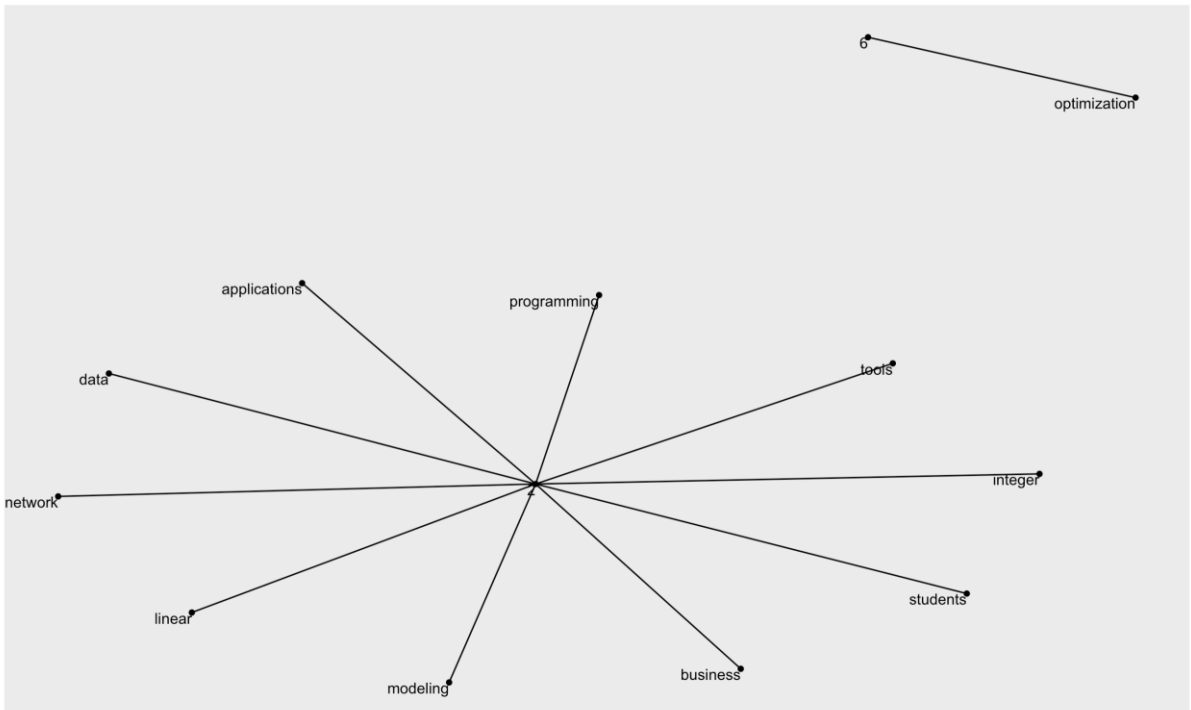


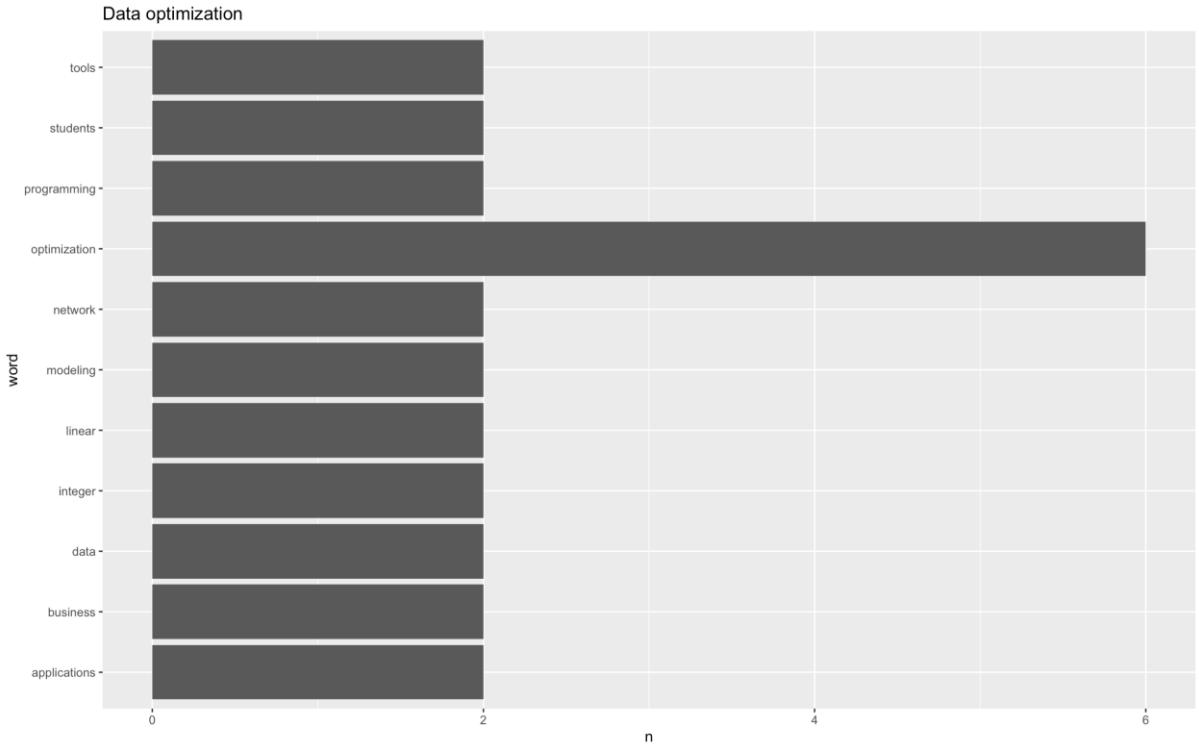
	line	subject	word	sentiment
1	2	Data Management and SQL	management	positive
2	2	Data Management and SQL	management	trust
3	2	Data Management and SQL	understanding	positive
4	2	Data Management and SQL	understanding	trust
5	2	Data Management and SQL	lead	positive
6	2	Data Management and SQL	actionable	anger
7	2	Data Management and SQL	actionable	disgust
8	2	Data Management and SQL	actionable	negative
9	2	Data Management and SQL	manage	positive
10	2	Data Management and SQL	manage	trust
11	2	Data Management and SQL	system	trust
12	2	Data Management and SQL	information	positive
13	2	Data Management and SQL	offer	positive
14	2	Data Management and SQL	include	positive
15	2	Data Management and SQL	develop	anticipation
16	2	Data Management and SQL	develop	positive
17	2	Data Management and SQL	proficiency	anticipation
18	2	Data Management and SQL	proficiency	joy
19	2	Data Management and SQL	proficiency	positive
20	2	Data Management and SQL	proficiency	surprise

Showing 1 to 21 of 29 entries, 4 total columns



Data optimization

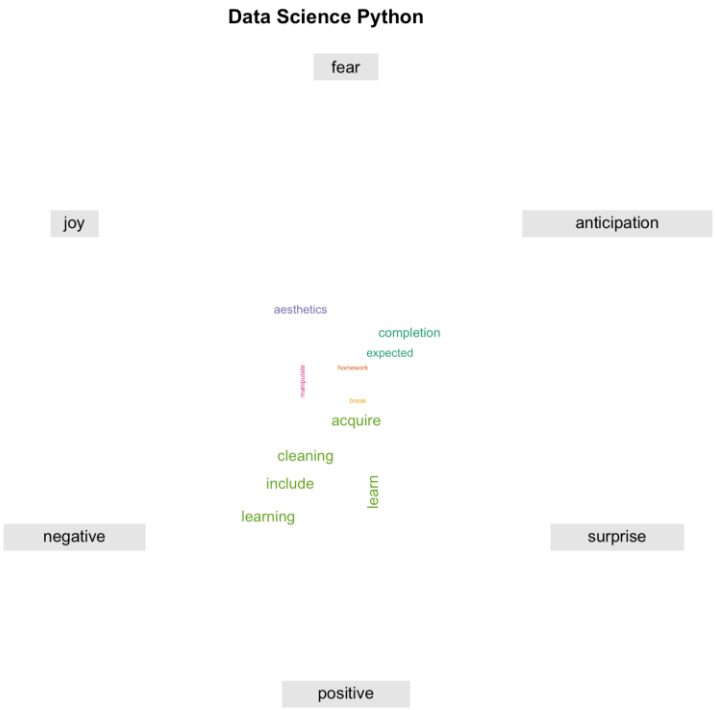




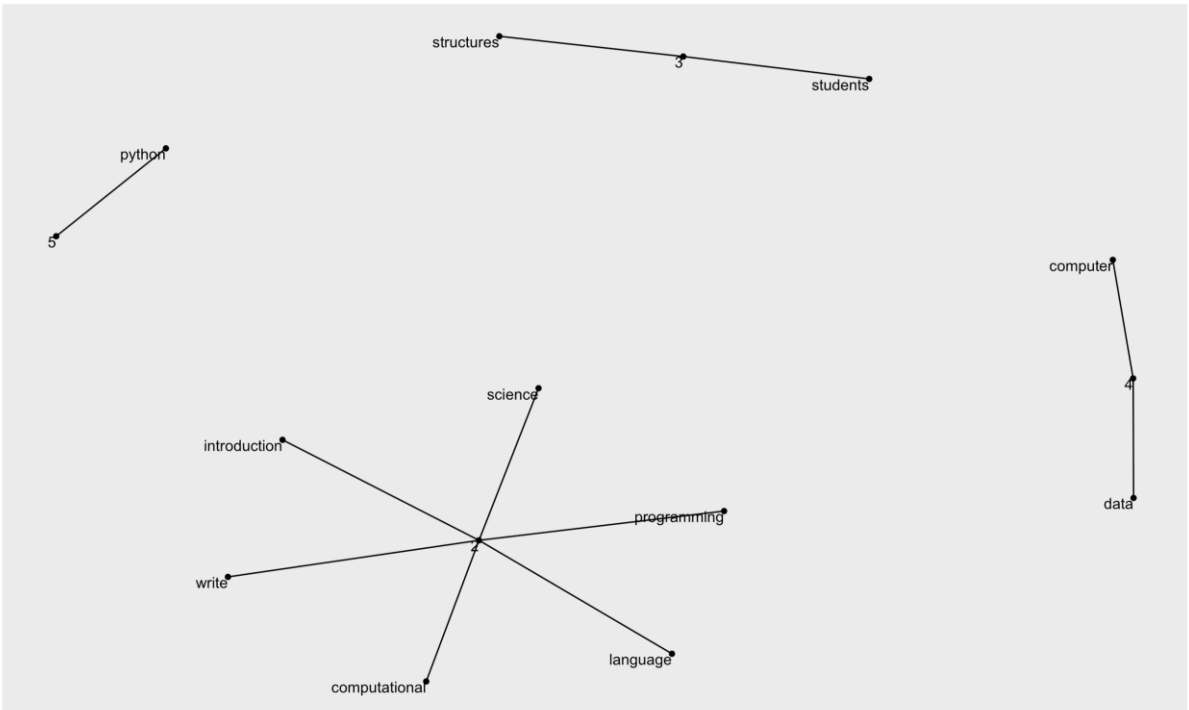
Filter

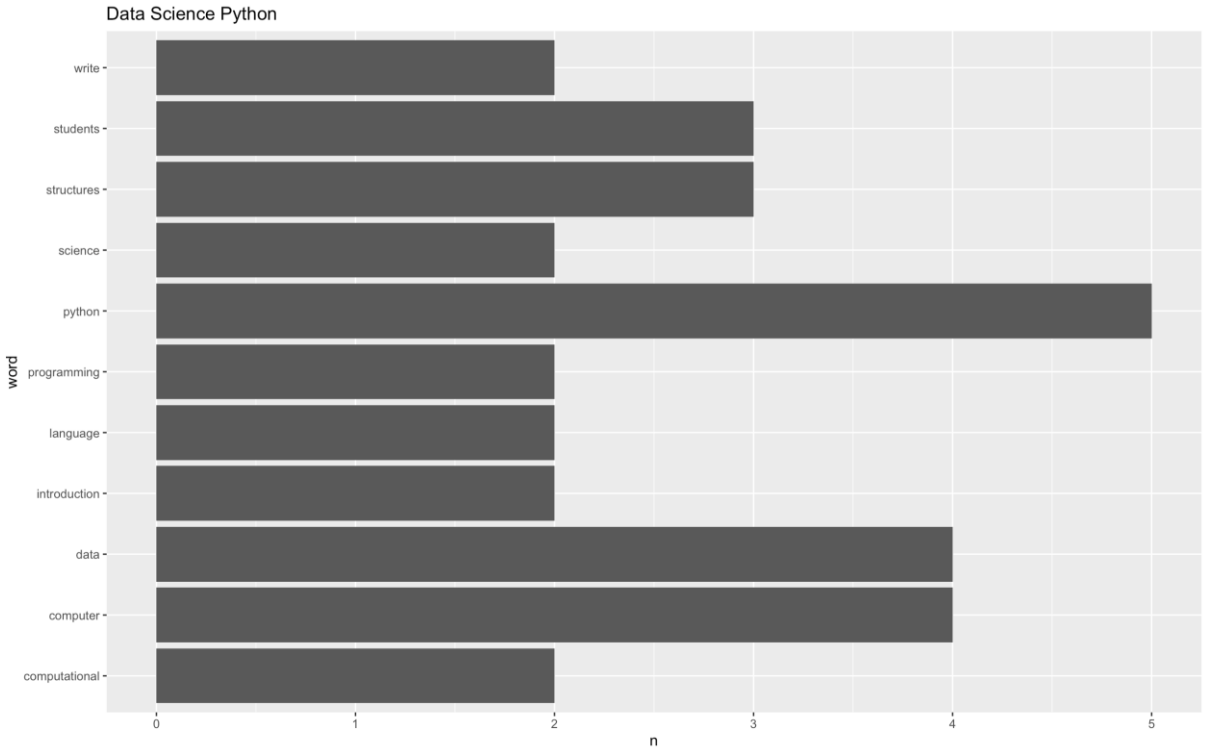
	line	subject	word	sentiment
1	3	Data optimization	theory	anticipation
2	3	Data optimization	theory	trust
3	3	Data optimization	include	positive
4	3	Data optimization	advanced	positive
5	3	Data optimization	network	anticipation
6	3	Data optimization	cover	trust
7	3	Data optimization	network	anticipation
8	3	Data optimization	management	positive
9	3	Data optimization	management	trust
10	3	Data optimization	study	positive
11	3	Data optimization	time	anticipation
12	3	Data optimization	considerable	positive
13	3	Data optimization	attention	positive
14	3	Data optimization	subject	negative
15	3	Data optimization	limited	anger
16	3	Data optimization	limited	negative
17	3	Data optimization	limited	sadness
18	3	Data optimization	uncertain	anger
19	3	Data optimization	uncertain	disgust
20	3	Data optimization	uncertain	fear

Showing 1 to 21 of 25 entries, 4 total columns



Data Science Python

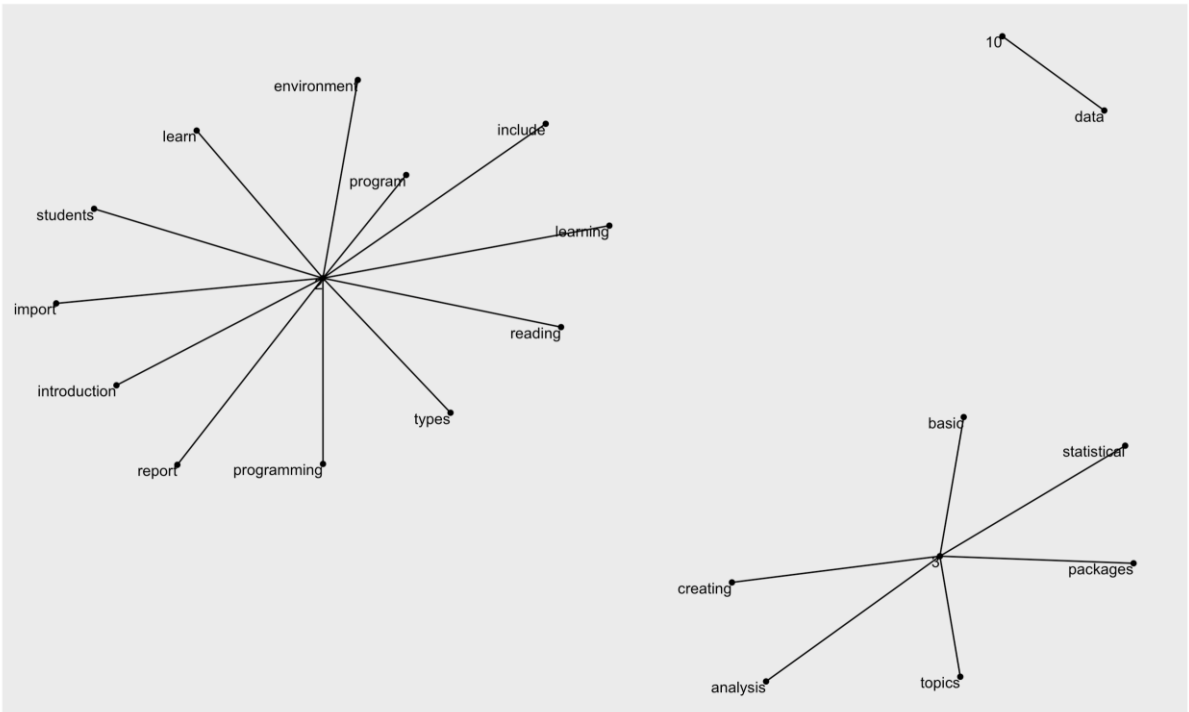


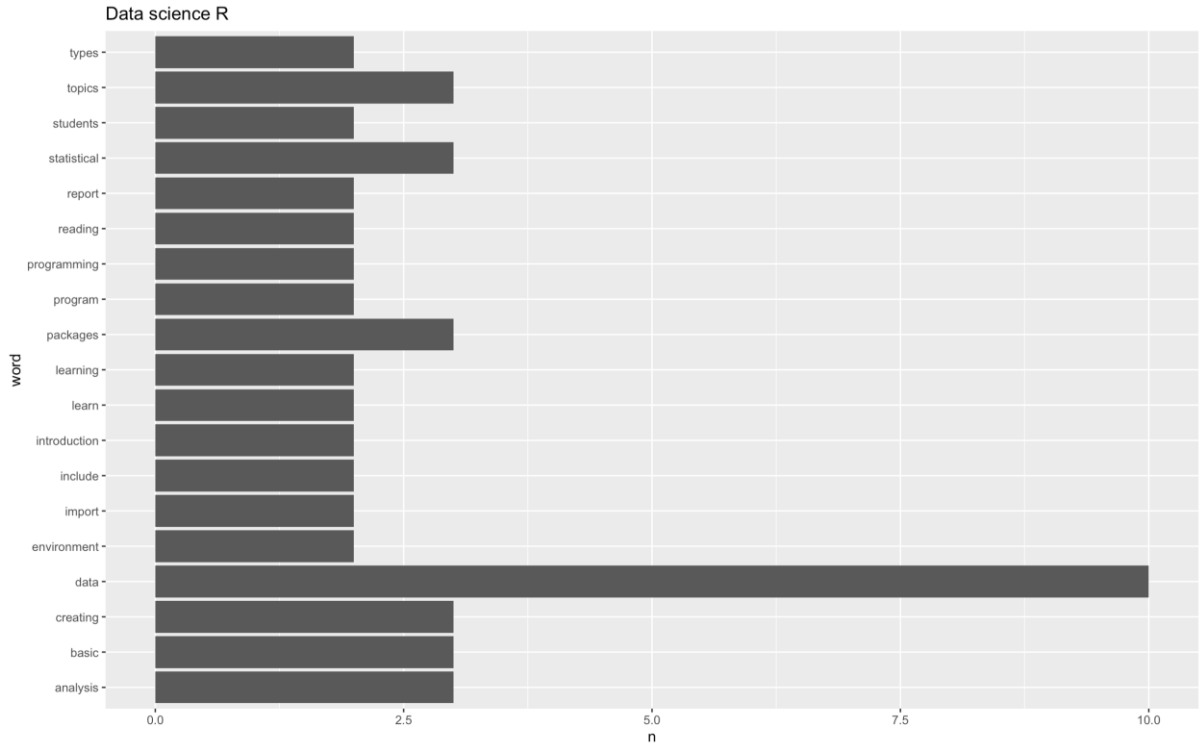


	line	subject	word	sentiment
1	4	Data Science Python	learn	positive
2	4	Data Science Python	homework	fear
3	4	Data Science Python	break	surprise
4	4	Data Science Python	aesthetics	joy
5	4	Data Science Python	aesthetics	positive
6	4	Data Science Python	include	positive
7	4	Data Science Python	cleaning	positive
8	4	Data Science Python	learning	positive
9	4	Data Science Python	completion	anticipation
10	4	Data Science Python	completion	joy
11	4	Data Science Python	completion	positive
12	4	Data Science Python	expected	anticipation
13	4	Data Science Python	acquire	positive
14	4	Data Science Python	manipulate	negative

Showing 1 to 14 of 14 entries, 4 total columns

Data science R

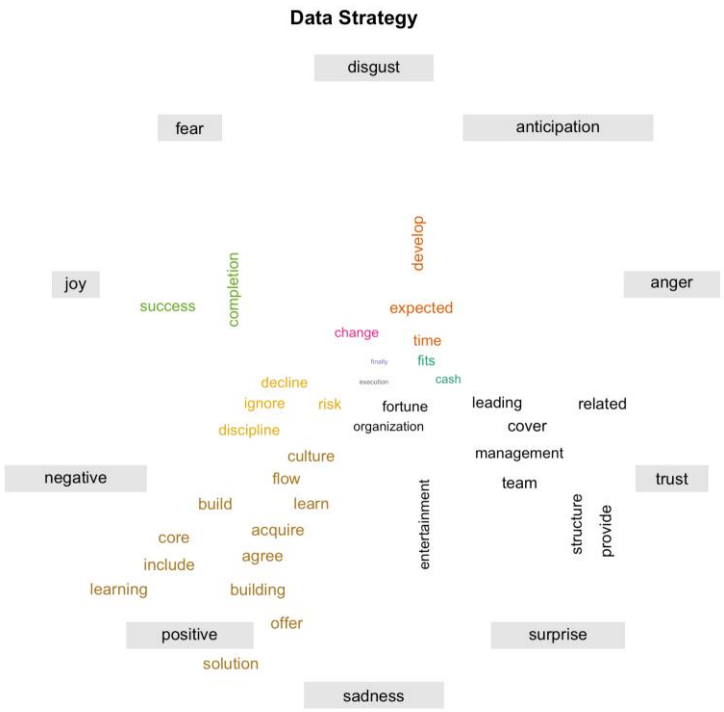




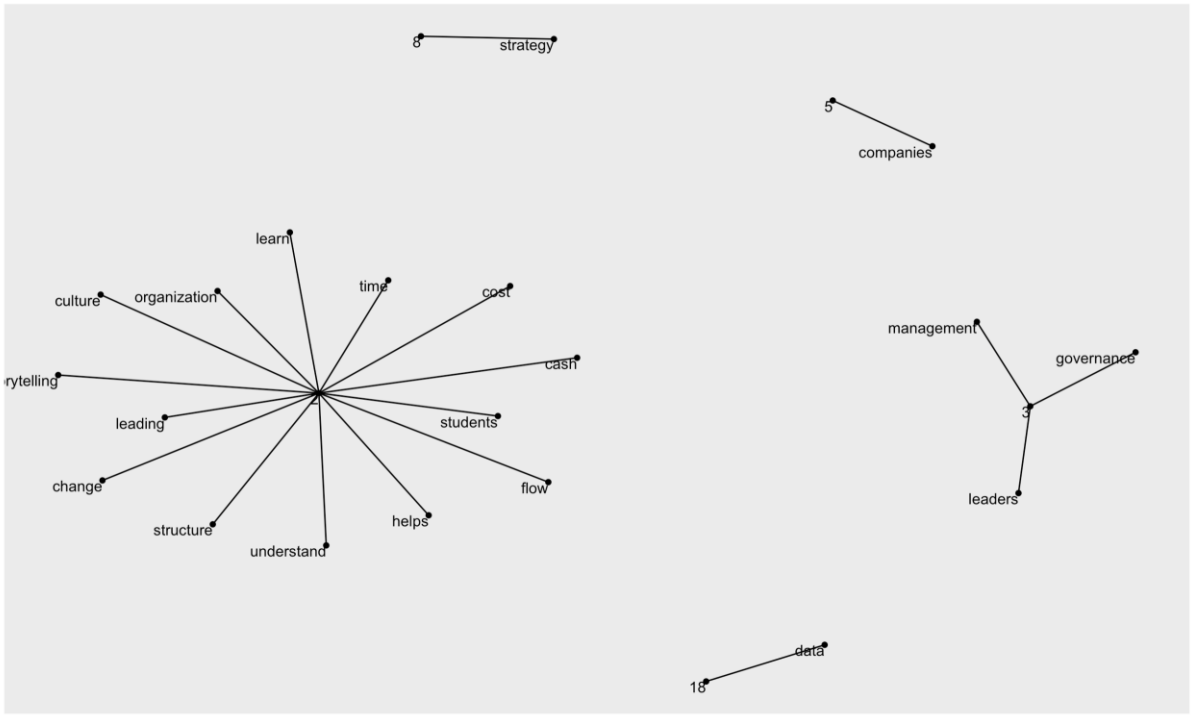
Filter

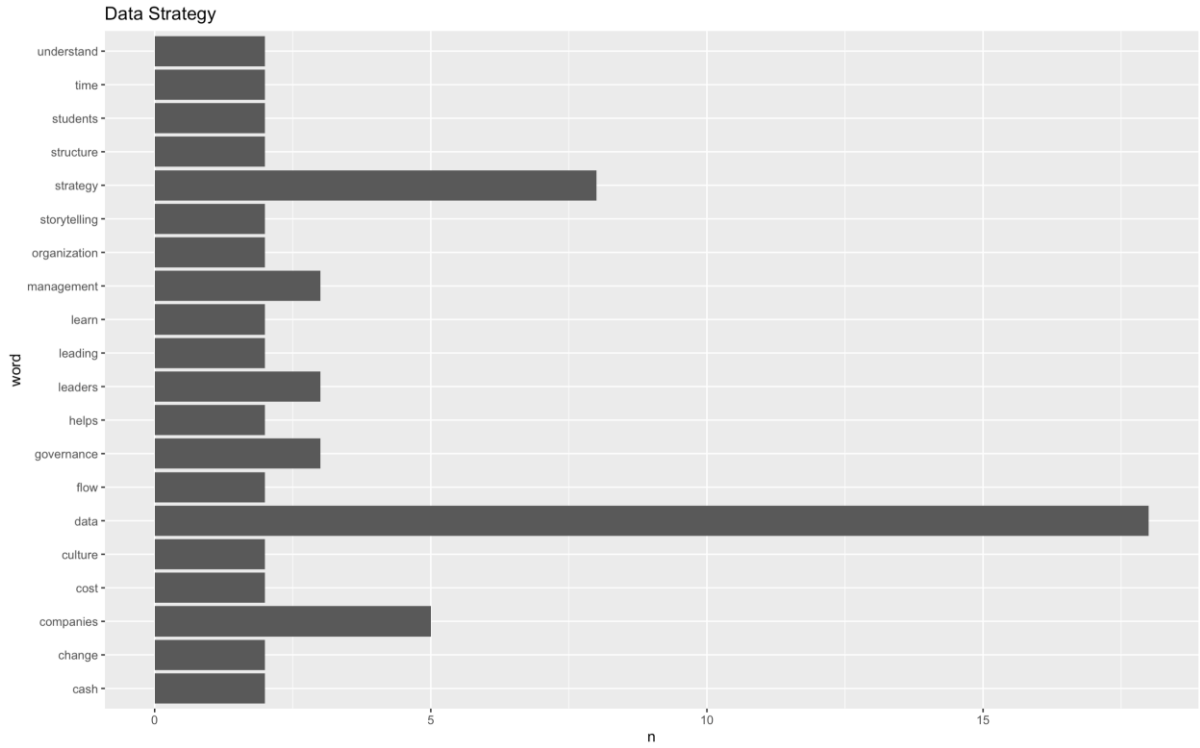
	line	subject	word	sentiment
1	5	Data science R	include	positive
2	5	Data science R	reading	positive
3	5	Data science R	learn	positive
4	5	Data science R	statistical	trust
5	5	Data science R	reading	positive
6	5	Data science R	statistical	trust
7	5	Data science R	provide	positive
8	5	Data science R	provide	trust
9	5	Data science R	include	positive
10	5	Data science R	machine	trust
11	5	Data science R	learning	positive
12	5	Data science R	learning	positive
13	5	Data science R	completion	anticipation
14	5	Data science R	completion	joy
15	5	Data science R	completion	positive
16	5	Data science R	expected	anticipation
17	5	Data science R	learn	positive
18	5	Data science R	effective	positive
19	5	Data science R	effective	trust
20	5	Data science R	resources	joy

Showing 1 to 21 of 24 entries, 4 total columns



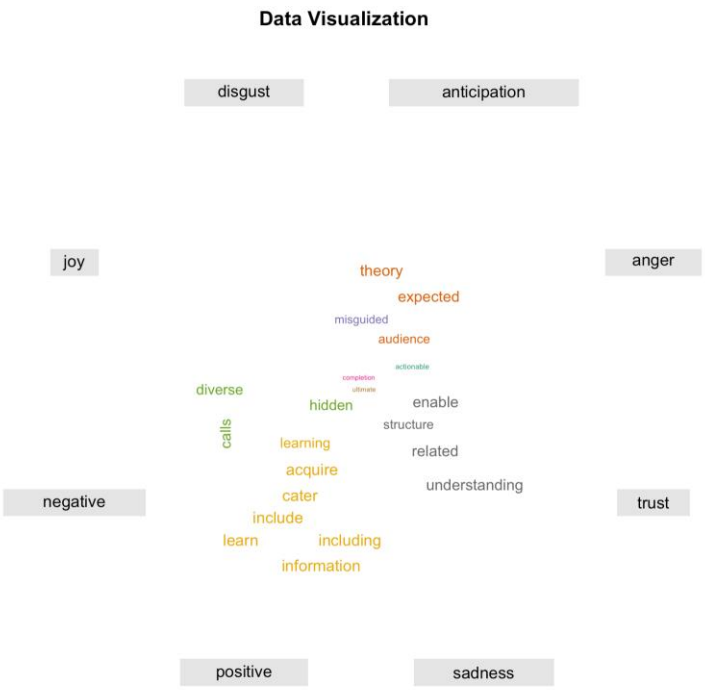
Data Strategy



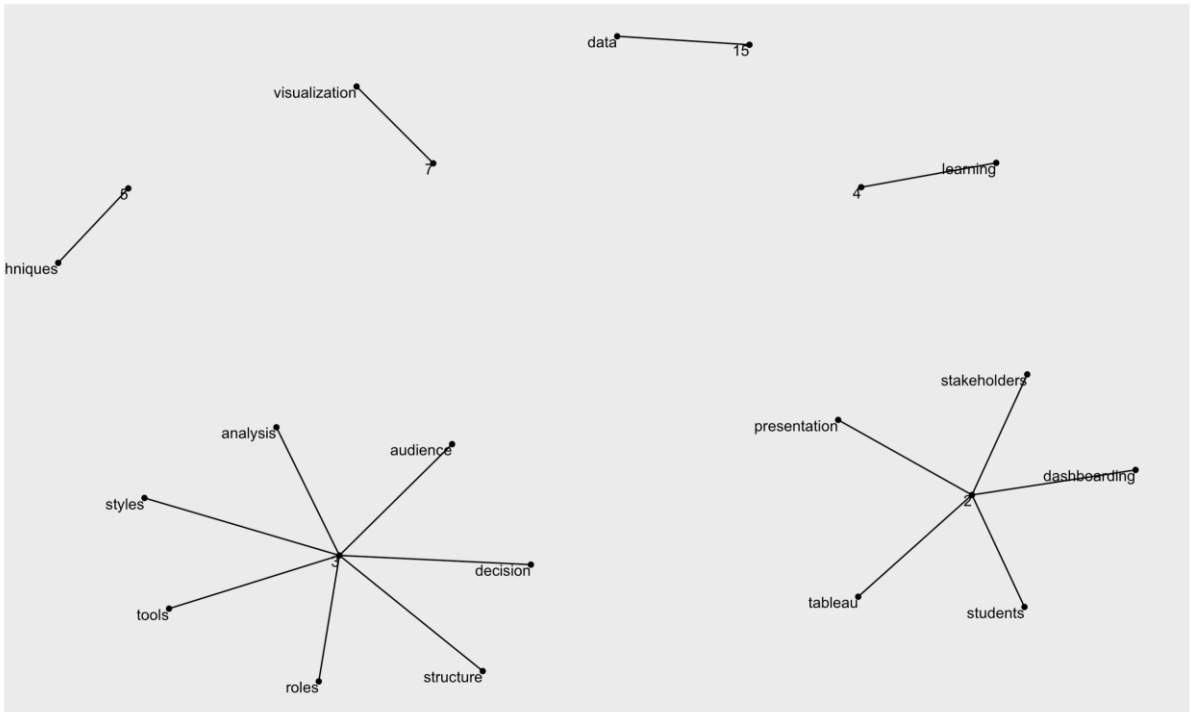


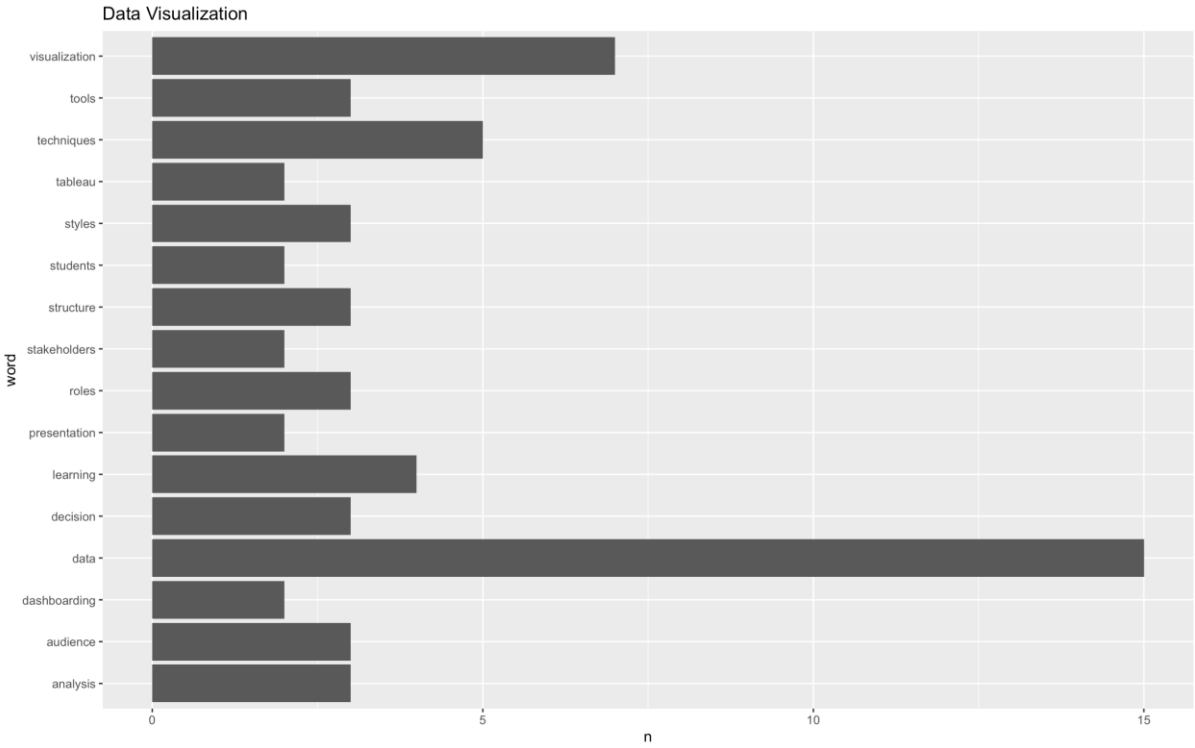
	line	subject	word	sentiment
1	6	Data Strategy	success	anticipation
2	6	Data Strategy	success	joy
3	6	Data Strategy	success	positive
4	6	Data Strategy	offer	positive
5	6	Data Strategy	provide	positive
6	6	Data Strategy	provide	trust
7	6	Data Strategy	solution	positive
8	6	Data Strategy	time	anticipation
9	6	Data Strategy	core	positive
10	6	Data Strategy	cash	anger
11	6	Data Strategy	cash	anticipation
12	6	Data Strategy	cash	fear
13	6	Data Strategy	cash	joy
14	6	Data Strategy	cash	positive
15	6	Data Strategy	cash	trust
16	6	Data Strategy	flow	positive
17	6	Data Strategy	cash	anger
18	6	Data Strategy	cash	anticipation
19	6	Data Strategy	cash	fear
20	6	Data Strategy	cash	joy

Showing 1 to 21 of 101 entries, 4 total columns



Data Visualization

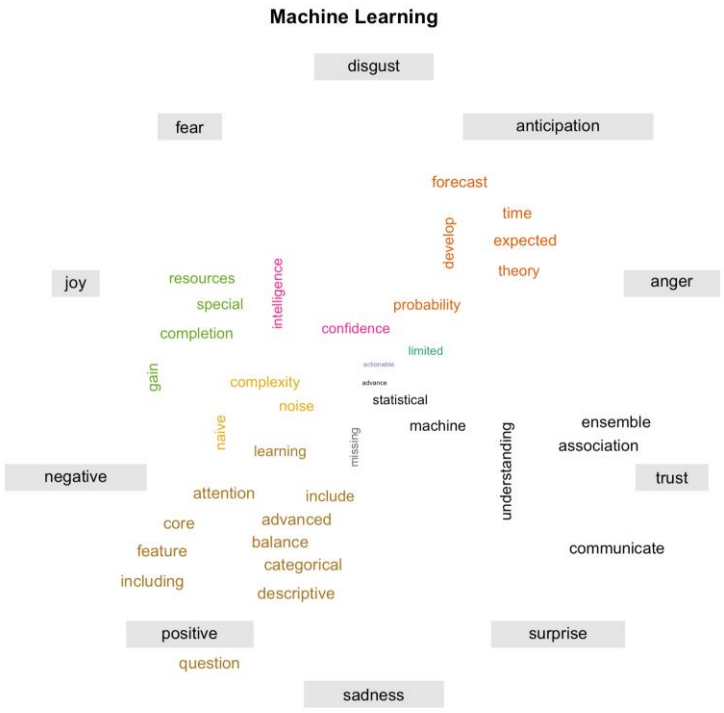




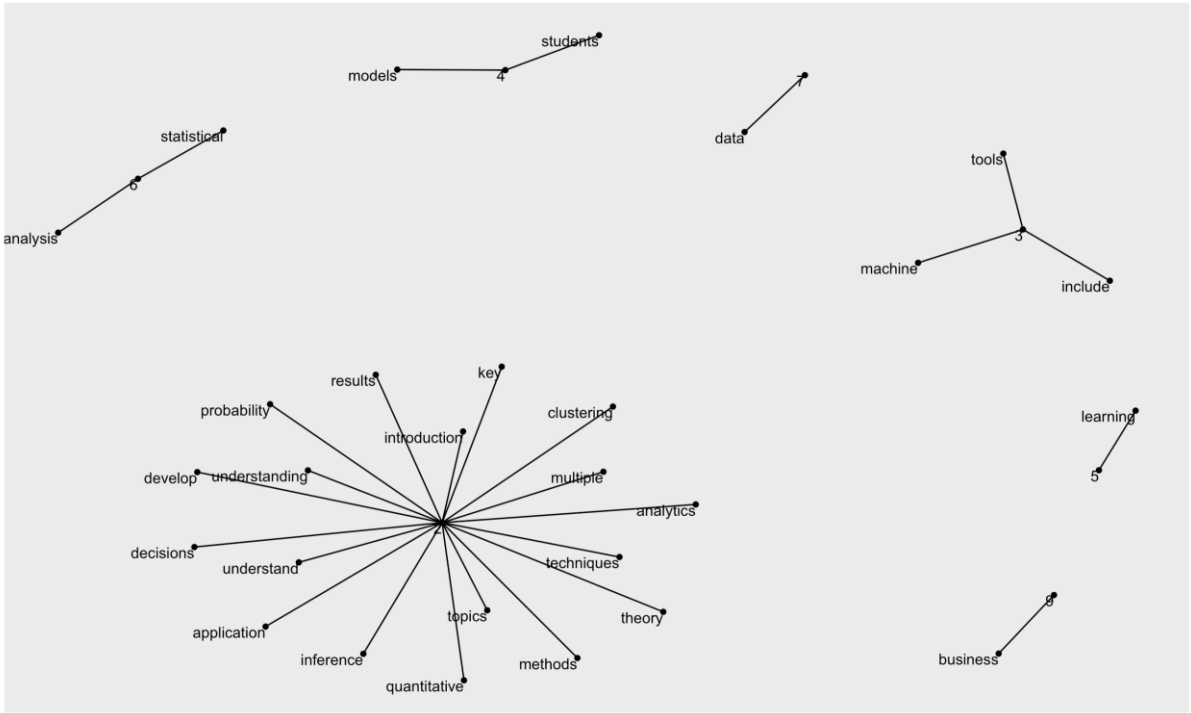
Filter

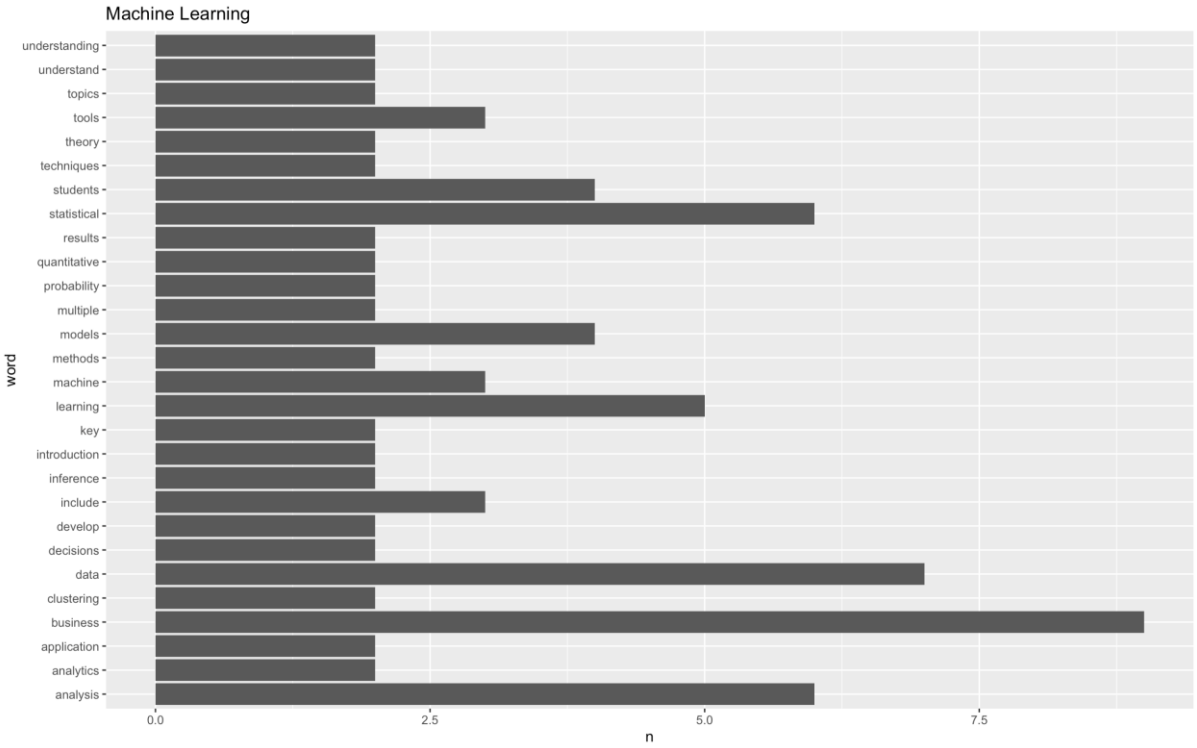
	line	subject	word	sentiment
1	7	Data Visualization	ultimate	anticipation
2	7	Data Visualization	ultimate	sadness
3	7	Data Visualization	calls	anticipation
4	7	Data Visualization	calls	negative
5	7	Data Visualization	calls	trust
6	7	Data Visualization	hidden	negative
7	7	Data Visualization	enable	positive
8	7	Data Visualization	enable	trust
9	7	Data Visualization	actionable	anger
10	7	Data Visualization	actionable	disgust
11	7	Data Visualization	actionable	negative
12	7	Data Visualization	misguided	disgust
13	7	Data Visualization	misguided	negative
14	7	Data Visualization	including	positive
15	7	Data Visualization	structure	positive
16	7	Data Visualization	structure	trust
17	7	Data Visualization	theory	anticipation
18	7	Data Visualization	theory	trust
19	7	Data Visualization	structure	positive
20	7	Data Visualization	structure	trust

Showing 1 to 21 of 43 entries, 4 total columns



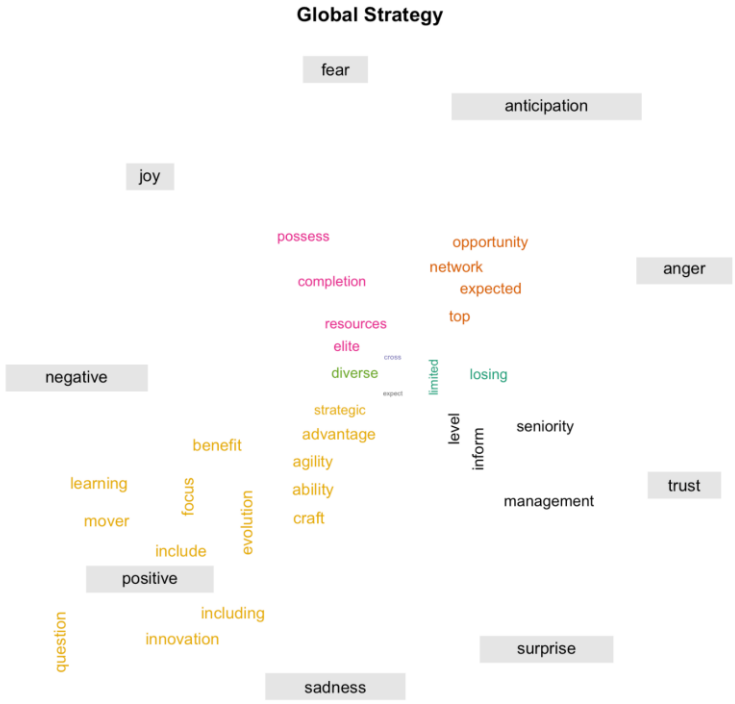
Machine Learning

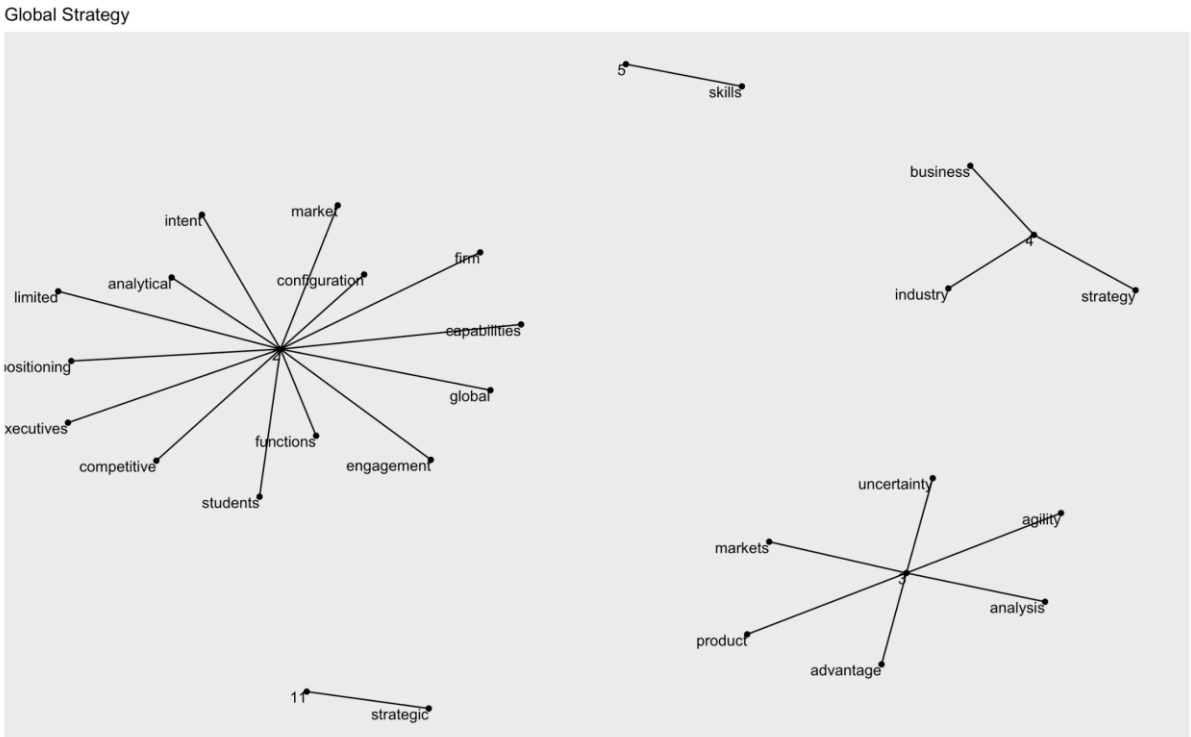


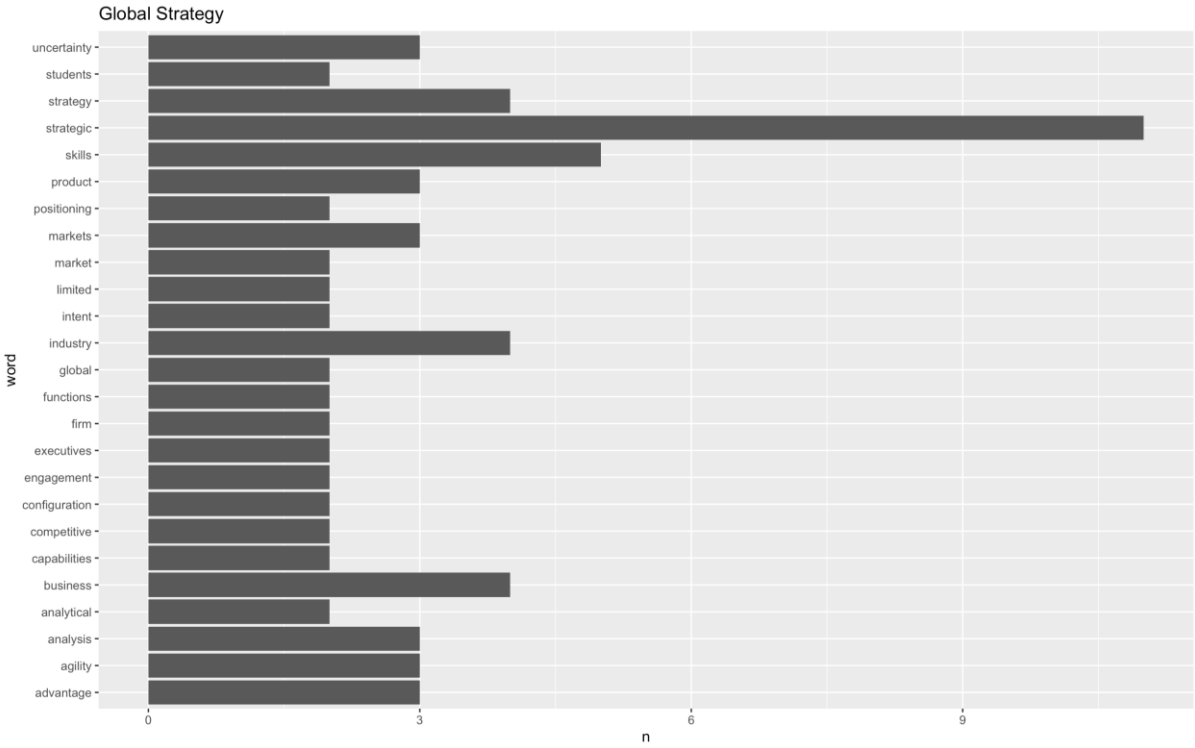

Machine Learning
understanding
understand
topics
tools
theory
techniques
students
statistical
results
quantitative
probability
multiple
models
methods
machine
learning
key
introduction
inference
include
develop
decisions
data
clustering
business
application
analytics
analysis
word
0.0
2.5
5.0
7.5
n

	line	subject	word	sentiment
1	8	Machine Learning	machine	trust
2	8	Machine Learning	learning	positive
3	8	Machine Learning	core	positive
4	8	Machine Learning	theory	anticipation
5	8	Machine Learning	theory	trust
6	8	Machine Learning	feature	positive
7	8	Machine Learning	include	positive
8	8	Machine Learning	naive	negative
9	8	Machine Learning	association	trust
10	8	Machine Learning	ensemble	positive
11	8	Machine Learning	ensemble	trust
12	8	Machine Learning	machine	trust
13	8	Machine Learning	learning	positive
14	8	Machine Learning	statistical	trust
15	8	Machine Learning	include	positive
16	8	Machine Learning	learning	positive
17	8	Machine Learning	machine	trust
18	8	Machine Learning	learning	positive
19	8	Machine Learning	gain	anticipation
20	8	Machine Learning	gain	joy

Showing 1 to 22 of 85 entries, 4 total columns


Global Strategy
fear
anticipation
joy
possess
opportunity
network
anger
completion
expected
top
resources
elite
diverse
negative
limited
losing
strategic
level
inform
seniority
benefit
advantage
agility
learning
focus
evolution
ability
trust
management
mover
craft
include
positive
including
question
innovation
surprise
sadness

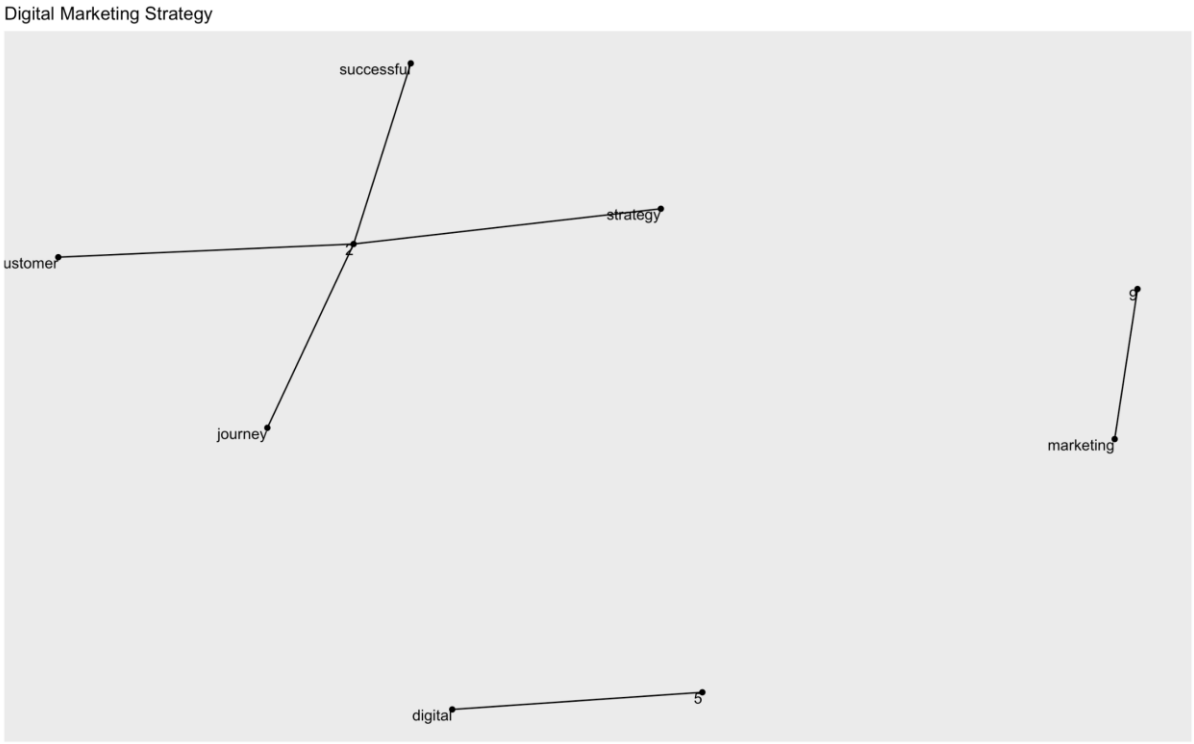

Global Strategy
5
skills
business
4
industry
strategy
intent
market
firm
analytical
configuration
limited
capabilities
positioning
global
executives
functions
competitive
engagement
students
uncertainty
agility
markets
3
analysis
product
advantage
11
strategic

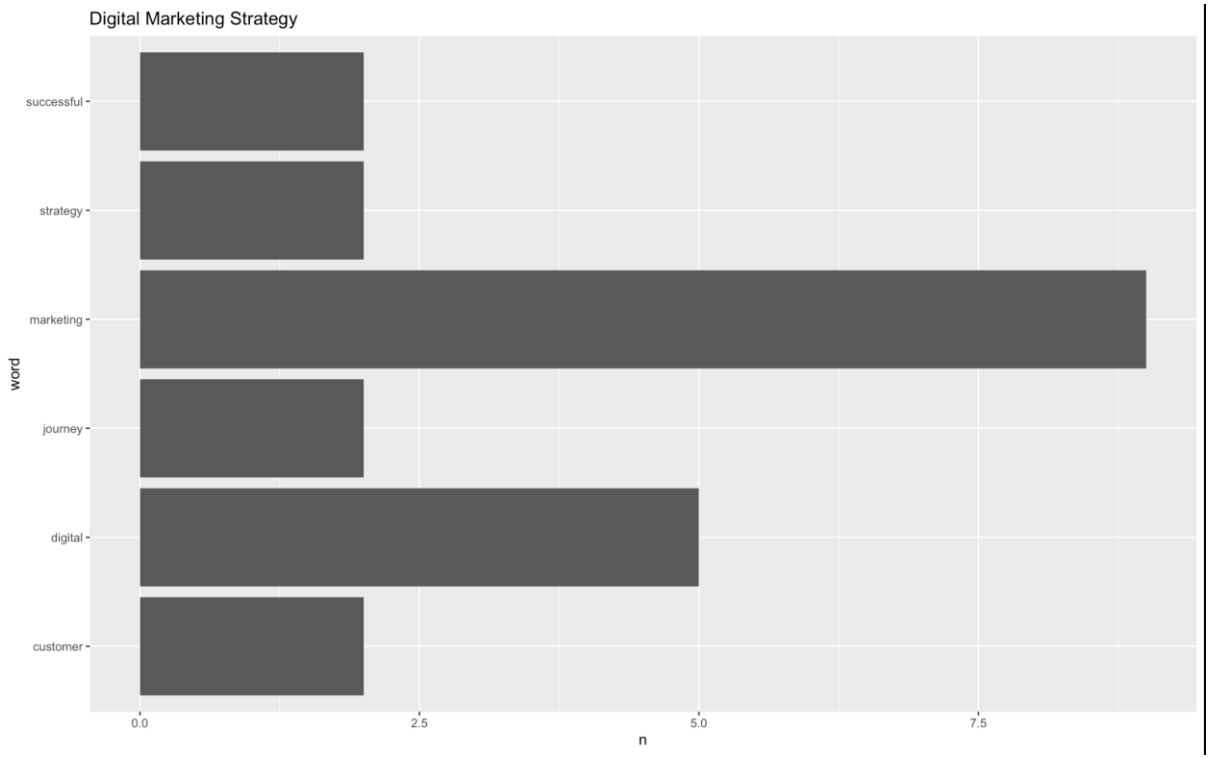


	line	subject	word	sentiment
1	9	Global Strategy	strategic	positive
2	9	Global Strategy	seniority	positive
3	9	Global Strategy	seniority	trust
4	9	Global Strategy	benefit	positive
5	9	Global Strategy	limited	anger
6	9	Global Strategy	limited	negative
7	9	Global Strategy	limited	sadness
8	9	Global Strategy	elite	anticipation
9	9	Global Strategy	elite	joy
10	9	Global Strategy	elite	positive
11	9	Global Strategy	elite	trust
12	9	Global Strategy	top	anticipation
13	9	Global Strategy	top	positive
14	9	Global Strategy	top	trust
15	9	Global Strategy	expect	anticipation
16	9	Global Strategy	expect	positive
17	9	Global Strategy	expect	surprise
18	9	Global Strategy	expect	trust
19	9	Global Strategy	level	positive
20	9	Global Strategy	level	trust

Showing 1 to 22 of 75 entries, 4 total columns


Digital Marketing Strategy
fear
joy
anticipation
grow
completion
successful
develop
create
journey
success
lacking
expected
measure
planning
effective
expertise
agile
include
understanding
customer
negative
trust
growth
learning
strategic
positive

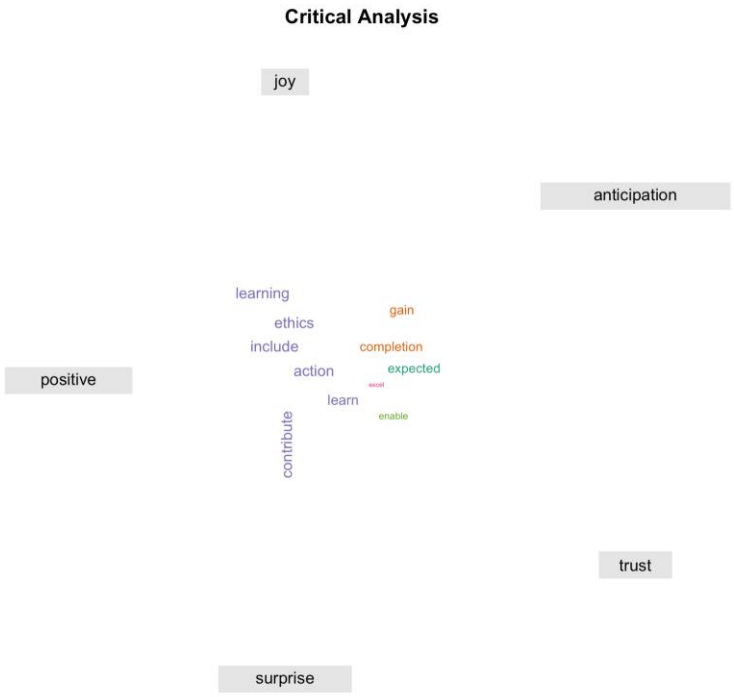

Digital Marketing Strategy
successful
strategy
ustomer
2
journey
9
marketing
digital
5

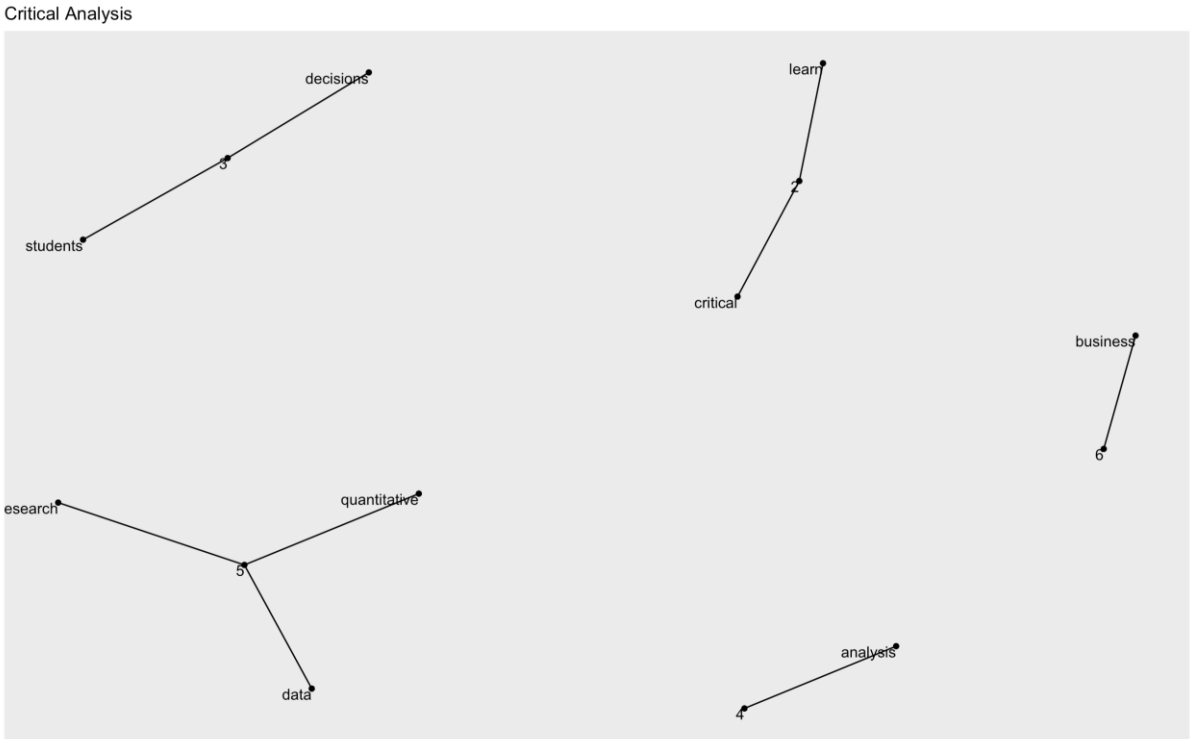


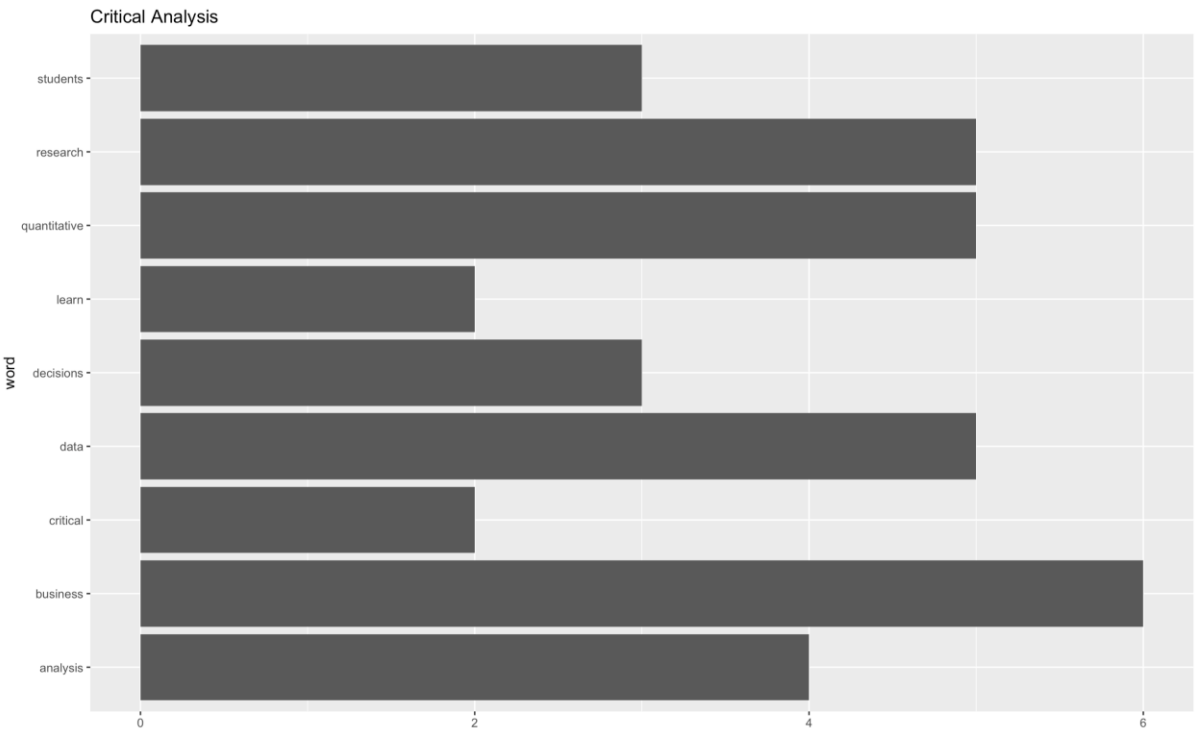
Filter

	line	subject	word	sentiment
48	10	Digital Marketing Strategy	understanding	positive
49	10	Digital Marketing Strategy	understanding	trust
40	10	Digital Marketing Strategy	successful	anticipation
41	10	Digital Marketing Strategy	successful	joy
42	10	Digital Marketing Strategy	successful	positive
43	10	Digital Marketing Strategy	successful	trust
44	10	Digital Marketing Strategy	successful	anticipation
45	10	Digital Marketing Strategy	successful	joy
46	10	Digital Marketing Strategy	successful	positive
47	10	Digital Marketing Strategy	successful	trust
37	10	Digital Marketing Strategy	success	anticipation
38	10	Digital Marketing Strategy	success	joy
39	10	Digital Marketing Strategy	success	positive
36	10	Digital Marketing Strategy	strategic	positive
33	10	Digital Marketing Strategy	planning	anticipation
34	10	Digital Marketing Strategy	planning	positive
35	10	Digital Marketing Strategy	planning	trust
32	10	Digital Marketing Strategy	measure	trust
31	10	Digital Marketing Strategy	learning	positive
30	10	Digital Marketing Strategy	lacking	negative

Showing 1 to 22 of 49 entries, 4 total columns


Critical Analysis
joy
anticipation
learning
gain
ethics
include
completion
action
expected
positive
learn
enable
contribute
trust
surprise


Critical Analysis
decisions
3
students
learn
2
critical
business
6
esearch
quantitative
5
data
analysis
4



	line	subject	word	sentiment
1	11	Critical Analysis	enable	positive
2	11	Critical Analysis	enable	trust
3	11	Critical Analysis	gain	anticipation
4	11	Critical Analysis	gain	joy
5	11	Critical Analysis	gain	positive
6	11	Critical Analysis	contribute	positive
7	11	Critical Analysis	learn	positive
8	11	Critical Analysis	learn	positive
9	11	Critical Analysis	include	positive
10	11	Critical Analysis	ethics	positive
11	11	Critical Analysis	excel	anticipation
12	11	Critical Analysis	excel	joy
13	11	Critical Analysis	excel	positive
14	11	Critical Analysis	excel	surprise
15	11	Critical Analysis	excel	trust
16	11	Critical Analysis	learning	positive
17	11	Critical Analysis	completion	anticipation
18	11	Critical Analysis	completion	joy
19	11	Critical Analysis	completion	positive
20	11	Critical Analysis	expected	anticipation
21	11	Critical Analysis	action	positive

CODE

INPUT

```
############################################################
# HOMEWORK:
# Neil Parekh
# Student Id :4647284
############################################################
install.packages('pdftools')
install.packages('shapeR')
install.packages('tidytext')
install.packages('tidyverse')
install.packages("textreadr")
install.packages("textshape")
install.packages("dplyr")
install.packages("textdata")
install.packages("reshape2")
install.packages("wordcloud")
install.packages("igraph")
install.packages("ggraph")

library(pdftools)
library(shapeR)
library(tidytext)
library(tidyverse)
library(textreadr)
library(textshape)
library(dplyr)
library(scales)
library(tidyr)
library(dplyr)
library(tidytext)
library(stringr)
library(ggplot2)
library(textdata)
library(reshape2)
library(wordcloud)
library(igraph)
library(ggraph)

# Clean environment
rm(list = ls())

# Read in text and save it to variable my_pdf_text as dataframe
pwd <- "/Users/neilparekh/Desktop/Business Analytics/TA/PDF"
setwd(pwd)
```

```
nm <- list.files(path=pwd)

my_pdf_text <- data.frame(do.call(rbind, lapply(nm,function(x) pdf_text(x))))
# Transpose
subject_names <- c('Text Analysis & Natural Language Processing','Data Management and SQL','Data
optimization',
            'Data Science Python','Data science R','Data Strategy','Data Visualization','Machine Learning')
subject_names <- c('Text Analysis & Natural Language Processing',
            'Data Management and SQL',
            'Data optimization',
            'Data Science Python',
            'Data science R',
            'Data Strategy',
            'Data Visualization',
            'Machine Learning',
            'Global Strategy',
            'Digital Marketing Strategy',
            'Critical Analysis')
my_pdf_text <- data_frame(line=1:ncol(my_pdf_text),text=t(as.matrix(my_pdf_text)),subject=subject_names)

# Custom stopwords
custom_stop_words <- c('recommendations','regression')

#Looping over each subject
for (val in 1:nrow(my_pdf_text)){
#for (val in 11:11){
 subject_pdf_text <- my_pdf_text[val,]
 subject_name <- subject_names[[val]]
 #Token
 token_list <- subject_pdf_text %>%
  unnest_tokens(word,text)
 print(token_list)

 #Remove stop words
 tokens_nostop <- token_list %>%
  anti_join(stop_words) %>%
  filter(!word %in% custom_stop_words)  #here's where we remove  custom tokens

 frequencies_tokens_nostop <- tokens_nostop %>%
  count(word, sort=TRUE) %>%
  filter(n>1)
 print(frequencies_tokens_nostop)

 #token frequency histogram
 freq_hist <- frequencies_tokens_nostop  %>%
  ggplot(aes(word, n))+
  ggtitle(subject_name)+
```

```
  geom_col()+
  coord_flip()
print(freq_hist)

# Calculate token graph
token_graph <- frequencies_tokens_nostop %>%
  graph_from_data_frame()

token_graph_plot <- ggraph(token_graph, layout = "fr")+
  ggtitle(subject_name)+
  geom_edge_link()+
  geom_node_point()+
  geom_node_text(aes(label=name), vjust=1, hjust=1)
print(token_graph_plot)

#Business Analysis
nrc <- get_sentiments("nrc")
#table(nrc$sentiment)
sentiments <- bind_rows(
  #(mutate(afinn,lexicon="afinn")),
  (mutate(nrc,lexicon="nrc")),
  #(mutate(bing,lexicon="bing"))
)

#Sentiment analysis
my_sentiment_nrc <- tokens_nostop %>%
  inner_join(get_sentiments("nrc"))
my_sentiment_bing <- tokens_nostop %>%
  inner_join(get_sentiments("bing"))
my_sentiment_afinn <- tokens_nostop %>%
  inner_join(get_sentiments("afinn"))

my_sentiment_afinn_mean_value <- my_sentiment_afinn %>%
  summarise(mean(value)) #Mean per subject
view(my_sentiment_afinn_mean_value)
table(my_sentiment_bing$sentiment)
View(my_sentiment_nrc)
#Graph for negative and positive
my_sentiment_nrc_plot<-my_sentiment_nrc %>%
  group_by(sentiment) %>%
  count(word, sentiment, sort=T)%>%
  top_n(10) %>%
  ungroup() %>%
  mutate(word=reorder(word, n))

my_sentiment_nrc_plot%>%
  ggplot(aes(word, n, fill=sentiment)) +
```

```
  ggtitle(subject_name)+
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y")+
  labs(y="different types of sentiment", x=NULL)+
  coord_flip()

 # Wordcloud to give accucurate data
 my_sentiment_nrc %>%
  count(word, sentiment, sort=TRUE) %>%
  acast(word ~sentiment, value.var="n", fill=0) %>%
  comparison.cloud(max.words = 100,
                   scale = c(0.4, 1),
                   fixed.asp=TRUE,
                   use.r.layout=TRUE,
                   match.colors=FALSE,
                   random.order=FALSE,
                   title.size=1) %>%
  title(main=paste("\n", subject_name, sep=""))

 ###############################################################
 ###### What if we are interested in the most common #######
 ################ 2 consecutive words - bi-gram ########
 ###############################################################

 # We want to see the bigrams (words that appear together, "pairs")
 subject_bigrams <- subject_pdf_text %>%
  unnest_tokens(bigram, text, token = "ngrams", n=2)

 subject_bigrams_separated <- subject_bigrams %>%
  separate(bigram, c("word1", "word2"), sep = " ")

 subject_bigrams_filtered <- subject_bigrams_separated %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word) %>%
  filter(!word1 %in% custom_stop_words) %>%
  filter(!word2 %in% custom_stop_words)

 # Creating new bigram, "no-stop-words":
 subject_bigram_counts <- subject_bigrams_filtered %>%
  count(word1, word2, sort = TRUE)

 # Want to see the new bigrams
 subject_bigram_counts

#Quadrograms

 subject_quadrograms <- subject_pdf_text %>%
```

```
  unnest_tokens(quadrogram, text, token = "ngrams", n=4)

 subject_quadrograms_separated <- subject_quadrograms %>%
  separate(quadrogram, c("word1", "word2", "word3", "word4"), sep = " ")

 subject_quadrograms_filtered <- subject_quadrograms_separated %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word) %>%
  filter(!word3 %in% stop_words$word) %>%
  filter(!word4 %in% stop_words$word) %>%
  filter(!word1 %in% custom_stop_words) %>%
  filter(!word2 %in% custom_stop_words) %>%
  filter(!word3 %in% custom_stop_words) %>%
  filter(!word4 %in% custom_stop_words)

 # Creating new quadrogram, "no-stop-words":
 subject_quadrogram_counts <- subject_quadrograms_filtered %>%
  count(word1, word2, word3, word4, sort = TRUE)

 # Want to see the new quadrograms
 subject_quadrogram_counts
}
```

OUTPUT

```
library(pdftools)
> library(shapeR)
> library(tidytext)
> library(tidyverse)
── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──
✓ ggplot2 3.2.1     ✓ purrr   0.3.3
✓ tibble  2.1.3     ✓ dplyr   0.8.4
✓ tidyr   1.0.0     ✓ stringr 1.4.0
✓ readr   1.3.1     ✓ forcats 0.4.0
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✕ dplyr::filter() masks stats::filter()
✕ dplyr::lag()    masks stats::lag()
> library(textreadr)
> library(textshape)

Attaching package: 'textshape'

The following object is masked from 'package:dplyr':

    combine

The following object is masked from 'package:purrr':

    flatten

The following object is masked from 'package:tibble':

    column_to_rownames

> library(dplyr)
> library(scales)

Attaching package: 'scales'

The following object is masked from 'package:purrr':

    discard

The following object is masked from 'package:readr':

    col_factor

> library(tidyr)
```

```
> library(dplyr)
> library(tidytext)
> library(stringr)
> library(ggplot2)
> library(textdata)
> library(reshape2)

Attaching package: 'reshape2'

The following object is masked from 'package:tidyr':

    smiths

> library(wordcloud)
Loading required package: RColorBrewer
> library(igraph)

Attaching package: 'igraph'

The following object is masked from 'package:textshape':

    ends

The following objects are masked from 'package:dplyr':

    as_data_frame, groups, union

The following objects are masked from 'package:purrr':

    compose, simplify

The following object is masked from 'package:tidyr':

    crossing

The following object is masked from 'package:tibble':

    as_data_frame

The following objects are masked from 'package:stats':

    decompose, spectrum

The following object is masked from 'package:base':

    union
```

```
> library(ggraph)
> # Clean environment
> rm(list = ls())
> # Read in text and save it to variable my_pdf_text as dataframe
> pwd <- "/Users/neilparekh/Desktop/Business Analytics/TA/PDF"
> setwd(pwd)
> nm <- list.files(path=pwd)
> my_pdf_text <- data.frame(do.call(rbind, lapply(nm,function(x) pdf_text(x))))
> # Transpose
> subject_names <- c('Text Analysis & Natural Language Processing','Data Management and SQL','Data
optimization',
+               'Data Science Python','Data science R','Data Strategy','Data Visualization','Machine Learning')
> subject_names <- c('Text Analysis & Natural Language Processing',
+               'Data Management and SQL',
+               'Data optimization',
+               'Data Science Python',
+               'Data science R',
+               'Data Strategy',
+               'Data Visualization',
+               'Machine Learning',
+               'Global Strategy',
+               'Digital Marketing Strategy',
+               'Critical Analysis')
> my_pdf_text <- data_frame(line=1:ncol(my_pdf_text),text=t(as.matrix(my_pdf_text)),subject=subject_names)
Warning message:
`data_frame()` is deprecated, use `tibble()`.
This warning is displayed once per session.
> # Custom stopwords
> custom_stop_words <- c('recommendations','regression')
> #Looping over each subject
> for (val in 1:nrow(my_pdf_text)){
+ #for (val in 11:11){
+   subject_pdf_text <- my_pdf_text[val,]
+   subject_name <- subject_names[[val]]
+   #Token
+   token_list <- subject_pdf_text %>%
+     unnest_tokens(word,text)
+   print(token_list)
+
+   #Remove stop words
+   tokens_nostop <- token_list %>%
+     anti_join(stop_words) %>%
+     filter(!word %in% custom_stop_words)  #here's where we remove  custom tokens
+
+   frequencies_tokens_nostop <- tokens_nostop %>%
+     count(word, sort=TRUE) %>%
+     filter(n>1)
```

```
+  print(frequencies_tokens_nostop)
+
+  #token frequency histogram
+  freq_hist <- frequencies_tokens_nostop  %>%
+    ggplot(aes(word, n))+
+    ggtitle(subject_name)+
+    geom_col()+
+    coord_flip()
+  print(freq_hist)
+
+  # Calculate token graph
+  token_graph <- frequencies_tokens_nostop %>%
+    graph_from_data_frame()
+
+  token_graph_plot <- ggraph(token_graph, layout = "fr")+
+    ggtitle(subject_name)+
+    geom_edge_link()+
+    geom_node_point()+
+    geom_node_text(aes(label=name), vjust=1, hjust=1)
+  print(token_graph_plot)
+
+  #Business Analysis
+  nrc <- get_sentiments("nrc")
+  #table(nrc$sentiment)
+  sentiments <- bind_rows(
+    #(mutate(afinn,lexicon="afinn")),
+    (mutate(nrc,lexicon="nrc")),
+    #(mutate(bing,lexicon="bing"))
+  )
+
+  #Sentiment analysis
+  my_sentiment_nrc <- tokens_nostop %>%
+    inner_join(get_sentiments("nrc"))
+  my_sentiment_bing <- tokens_nostop %>%
+    inner_join(get_sentiments("bing"))
+  my_sentiment_afinn <- tokens_nostop %>%
+    inner_join(get_sentiments("afinn"))
+
+  my_sentiment_afinn_mean_value <- my_sentiment_afinn %>%
+    summarise(mean(value)) #Mean per subject
+  view(my_sentiment_afinn_mean_value)
+  table(my_sentiment_bing$sentiment)
+  View(my_sentiment_nrc)
+  #Graph for negative and positive
+  my_sentiment_nrc_plot<-my_sentiment_nrc %>%
+    group_by(sentiment) %>%
+    count(word, sentiment, sort=T)%>%
```

```
+    top_n(10) %>%
+    ungroup() %>%
+    mutate(word=reorder(word, n))
+
+  my_sentiment_nrc_plot%>%
+    ggplot(aes(word, n, fill=sentiment)) +
+    ggtitle(subject_name)+
+    geom_col(show.legend = FALSE) +
+    facet_wrap(~sentiment, scales = "free_y")+
+    labs(y="different types of sentiment", x=NULL)+
+    coord_flip()
+
+  # Wordcloud to give accucurate data
+  my_sentiment_nrc %>%
+    count(word, sentiment, sort=TRUE) %>%
+    acast(word ~sentiment, value.var="n", fill=0) %>%
+    comparison.cloud(max.words = 100,
+                     scale = c(0.4, 1),
+                     fixed.asp=TRUE,
+                     use.r.layout=TRUE,
+                     match.colors=FALSE,
+                     random.order=FALSE,
+                     title.size=1) %>%
+    title(main=paste("\n", subject_name, sep=""))
+
+  ###############################################################
+  ###### What if we are interested in the most common #######
+  ################# 2 consecutive words - bi-gram ########
+  ###############################################################
+
+  # We want to see the bigrams (words that appear together, "pairs")
+  subject_bigrams <- subject_pdf_text %>%
+    unnest_tokens(bigram, text, token = "ngrams", n=2)
+
+  subject_bigrams_separated <- subject_bigrams %>%
+    separate(bigram, c("word1", "word2"), sep = " ")
+
+  subject_bigrams_filtered <- subject_bigrams_separated %>%
+    filter(!word1 %in% stop_words$word) %>%
+    filter(!word2 %in% stop_words$word) %>%
+    filter(!word1 %in% custom_stop_words) %>%
+    filter(!word2 %in% custom_stop_words)
+
+  # Creating new bigram, "no-stop-words":
+  subject_bigram_counts <- subject_bigrams_filtered %>%
+    count(word1, word2, sort = TRUE)
+
```

```
+  # Want to see the new bigrams
+  subject_bigram_counts
+
+  ###############################################################
+  ###### What if we are interested in the most common #######
+  ################# 4 consecutive words - quadro-gram ########
+  ###############################################################
+
+  subject_quadrograms <- subject_pdf_text %>%
+    unnest_tokens(quadrogram, text, token = "ngrams", n=4)
+
+  subject_quadrograms_separated <- subject_quadrograms %>%
+    separate(quadrogram, c("word1", "word2", "word3", "word4"), sep = " ")
+
+  subject_quadrograms_filtered <- subject_quadrograms_separated %>%
+    filter(!word1 %in% stop_words$word) %>%
+    filter(!word2 %in% stop_words$word) %>%
+    filter(!word3 %in% stop_words$word) %>%
+    filter(!word4 %in% stop_words$word) %>%
+    filter(!word1 %in% custom_stop_words) %>%
+    filter(!word2 %in% custom_stop_words) %>%
+    filter(!word3 %in% custom_stop_words) %>%
+    filter(!word4 %in% custom_stop_words)
+
+  # Creating new quadrogram, "no-stop-words":
+  subject_quadrogram_counts <- subject_quadrograms_filtered %>%
+    count(word1, word2, word3, word4, sort = TRUE)
+
+  # Want to see the new quadrograms
+  subject_quadrogram_counts
+ }
# A tibble: 84 x 3
   line subject                          word
  <int> <chr>                            <chr>
 1     1 Text Analysis & Natural Language Processing text
 2     1 Text Analysis & Natural Language Processing analysis
 3     1 Text Analysis & Natural Language Processing natural
 4     1 Text Analysis & Natural Language Processing language
 5     1 Text Analysis & Natural Language Processing processing
 6     1 Text Analysis & Natural Language Processing this
 7     1 Text Analysis & Natural Language Processing course
 8     1 Text Analysis & Natural Language Processing is
 9     1 Text Analysis & Natural Language Processing a
10     1 Text Analysis & Natural Language Processing deep
# … with 74 more rows
Joining, by = "word"
# A tibble: 4 x 2
```

```
  word        n
  <chr>    <int>
1 text        8
2 analysis    5
3 python      2
4 topics      2
Joining, by = "word"
Joining, by = "word"
Joining, by = "word"
Selecting by n
# A tibble: 132 x 3
   line subject              word
  <int> <chr>                <chr>
 1     2 Data Management and SQL data
 2     2 Data Management and SQL management
 3     2 Data Management and SQL and
 4     2 Data Management and SQL sql
 5     2 Data Management and SQL understanding
 6     2 Data Management and SQL and
 7     2 Data Management and SQL analyzing
 8     2 Data Management and SQL data
 9     2 Data Management and SQL is
10     2 Data Management and SQL a
# … with 122 more rows
Joining, by = "word"
# A tibble: 8 x 2
  word          n
  <chr>     <int>
1 data          8
2 database      3
3 relational    3
4 basic         2
5 databases     2
6 language      2
7 management    2
8 sql           2
Joining, by = "word"
Joining, by = "word"
Joining, by = "word"
Selecting by n
# A tibble: 112 x 3
   line subject          word
  <int> <chr>            <chr>
 1     3 Data optimization data
 2     3 Data optimization optimization
 3     3 Data optimization this
 4     3 Data optimization course
```

```
 5     3 Data optimization will
 6     3 Data optimization introduce
 7     3 Data optimization the
 8     3 Data optimization students
 9     3 Data optimization to
10     3 Data optimization the
# … with 102 more rows
Joining, by = "word"
# A tibble: 11 x 2
   word           n
   <chr>      <int>
 1 optimization     6
 2 applications     2
 3 business         2
 4 data             2
 5 integer          2
 6 linear           2
 7 modeling         2
 8 network          2
 9 programming      2
10 students         2
11 tools            2
Joining, by = "word"
Joining, by = "word"
Joining, by = "word"
Selecting by n
# A tibble: 153 x 3
    line subject           word
   <int> <chr>             <chr>
 1     4 Data Science Python data
 2     4 Data Science Python science
 3     4 Data Science Python python
 4     4 Data Science Python an
 5     4 Data Science Python introduction
 6     4 Data Science Python to
 7     4 Data Science Python the
 8     4 Data Science Python principles
 9     4 Data Science Python and
10     4 Data Science Python techniques
# … with 143 more rows
Joining, by = "word"
# A tibble: 11 x 2
   word           n
   <chr>      <int>
 1 python           5
 2 computer         4
 3 data             4
```

```
 4 structures       3
 5 students         3
 6 computational    2
 7 introduction     2
 8 language         2
 9 programming      2
10 science          2
11 write            2
Joining, by = "word"
Joining, by = "word"
Joining, by = "word"
Selecting by n
# A tibble: 212 x 3
   line subject       word
  <int> <chr>         <chr>
 1     5 Data science R data
 2     5 Data science R science
 3     5 Data science R r
 4     5 Data science R an
 5     5 Data science R introduction
 6     5 Data science R to
 7     5 Data science R the
 8     5 Data science R principles
 9     5 Data science R and
10     5 Data science R techniques
# … with 202 more rows
Joining, by = "word"
# A tibble: 19 x 2
   word           n
   <chr>      <int>
 1 data          10
 2 analysis       3
 3 basic          3
 4 creating       3
 5 packages       3
 6 statistical    3
 7 topics         3
 8 environment    2
 9 import         2
10 include        2
11 introduction   2
12 learn          2
13 learning       2
14 program        2
15 programming    2
16 reading        2
17 report         2
```

```
18 students        2
19 types           2
Joining, by = "word"
Joining, by = "word"
Joining, by = "word"
Selecting by n
# A tibble: 380 x 3
   line subject      word
  <int> <chr>        <chr>
 1     6 Data Strategy data
 2     6 Data Strategy strategy
 3     6 Data Strategy data
 4     6 Data Strategy plays
 5     6 Data Strategy a
 6     6 Data Strategy critical
 7     6 Data Strategy role
 8     6 Data Strategy in
 9     6 Data Strategy the
10     6 Data Strategy success
# … with 370 more rows
Joining, by = "word"
# A tibble: 20 x 2
   word            n
   <chr>       <int>
 1 data           18
 2 strategy        8
 3 companies       5
 4 governance      3
 5 leaders         3
 6 management      3
 7 cash            2
 8 change          2
 9 cost            2
10 culture         2
11 flow            2
12 helps           2
13 leading         2
14 learn           2
15 organization    2
16 storytelling    2
17 structure       2
18 students        2
19 time            2
20 understand      2
Joining, by = "word"
Joining, by = "word"
Joining, by = "word"
```

```
Selecting by n
# A tibble: 254 x 3
   line subject          word
  <int> <chr>            <chr>
 1     7 Data Visualization data
 2     7 Data Visualization visualization
 3     7 Data Visualization the
 4     7 Data Visualization ultimate
 5     7 Data Visualization value
 6     7 Data Visualization of
 7     7 Data Visualization data
 8     7 Data Visualization is
 9     7 Data Visualization to
10     7 Data Visualization help
# … with 244 more rows
Joining, by = "word"
# A tibble: 16 x 2
   word            n
   <chr>       <int>
 1 data           15
 2 visualization     7
 3 techniques        5
 4 learning          4
 5 analysis          3
 6 audience          3
 7 decision          3
 8 roles             3
 9 structure         3
10 styles            3
11 tools             3
12 dashboarding      2
13 presentation      2
14 stakeholders      2
15 students          2
16 tableau           2
Joining, by = "word"
Joining, by = "word"
Joining, by = "word"
Selecting by n
# A tibble: 351 x 3
   line subject         word
  <int> <chr>           <chr>
 1     8 Machine Learning machine
 2     8 Machine Learning learning
 3     8 Machine Learning this
 4     8 Machine Learning course
 5     8 Machine Learning focuses
```

```
 6     8 Machine Learning on
 7     8 Machine Learning the
 8     8 Machine Learning core
 9     8 Machine Learning theory
10     8 Machine Learning and
# … with 341 more rows
Joining, by = "word"
# A tibble: 28 x 2
   word          n
   <chr>     <int>
 1 business      9
 2 data          7
 3 analysis      6
 4 statistical   6
 5 learning      5
 6 models        4
 7 students      4
 8 include       3
 9 machine       3
10 tools         3
# … with 18 more rows
Joining, by = "word"
Joining, by = "word"
Joining, by = "word"
Selecting by n
# A tibble: 295 x 3
    line subject        word
   <int> <chr>          <chr>
 1     9 Global Strategy global
 2     9 Global Strategy strategy
 3     9 Global Strategy course
 4     9 Global Strategy description
 5     9 Global Strategy strategic
 6     9 Global Strategy skills
 7     9 Global Strategy are
 8     9 Global Strategy a
 9     9 Global Strategy key
10     9 Global Strategy asset
# … with 285 more rows
Joining, by = "word"
# A tibble: 25 x 2
   word         n
   <chr>    <int>
 1 strategic   11
 2 skills       5
 3 business     4
 4 industry     4
```

```
 5 strategy     4
 6 advantage    3
 7 agility      3
 8 analysis     3
 9 markets      3
10 product      3
# … with 15 more rows
Joining, by = "word"
Joining, by = "word"
Joining, by = "word"
Selecting by n
# A tibble: 139 x 3
   line subject                word
  <int> <chr>                  <chr>
 1    10 Digital Marketing Strategy digital
 2    10 Digital Marketing Strategy marketing
 3    10 Digital Marketing Strategy strategy
 4    10 Digital Marketing Strategy course
 5    10 Digital Marketing Strategy description
 6    10 Digital Marketing Strategy the
 7    10 Digital Marketing Strategy range
 8    10 Digital Marketing Strategy of
 9    10 Digital Marketing Strategy digital
10    10 Digital Marketing Strategy marketing
# … with 129 more rows
Joining, by = "word"
# A tibble: 6 x 2
  word          n
  <chr>     <int>
1 marketing     9
2 digital       5
3 customer      2
4 journey       2
5 strategy      2
6 successful    2
Joining, by = "word"
Joining, by = "word"
Joining, by = "word"
Selecting by n
# A tibble: 168 x 3
   line subject          word
  <int> <chr>            <chr>
 1    11 Critical Analysis critical
 2    11 Critical Analysis analysis
 3    11 Critical Analysis course
 4    11 Critical Analysis description
 5    11 Critical Analysis this
```

```
 6    11 Critical Analysis course
 7    11 Critical Analysis is
 8    11 Critical Analysis designed
 9    11 Critical Analysis to
10    11 Critical Analysis enable
# … with 158 more rows
Joining, by = "word"
# A tibble: 9 x 2
  word           n
  <chr>      <int>
1 business       6
2 data           5
3 quantitative   5
4 research       5
5 analysis       4
6 decisions      3
7 students       3
8 critical       2
9 learn          2
Joining, by = "word"
Joining, by = "word"
Joining, by = "word"
Selecting by n
There were 50 or more warnings (use warnings() to see the first 50)
>
>
```