STAT 3011 project

2023-02-07

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

```
## Warning: package 'rvest' was built under R version 4.2.2
```

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

```
## Warning: package 'dplyr' was built under R version 4.2.2
```

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

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

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

```
# 1 Download Html files and combine them into one file
path="C:/Users/22600/3011/bookcamp_code/Case_Study4/job_postings"
setwd(path)

filestoread <- list.files(path=path,pattern="\\.html$")
htmlfiles <- lapply(filestoread,function(x)try(read_html(x,encoding="UTF-8")))

l=length(htmlfiles)
cat("We have loaded",l, "HTML files.")
```

```
## We have loaded 1458 HTML files.
```

```
# 2 Parse html files

#install.packages("XML")
#install.packages("bitops")
#install.packages("RCurl")

library(XML)
```

```
## Warning: package 'XML' was built under R version 4.2.2
```

```
library(bitops)
library(RCurl)
```

```
## Warning: package 'RCurl' was built under R version 4.2.2
```

```
library(xml2)

setwd(dir=path)
soup <- lapply(filestoread,function(x)try(htmlParse(read_html(x))))
```

```
# 3 Parse html files into titles and bodies

setwd(path)
library(rvest)
library(bitops)
library(RCurl)
library(xml2)
library(dplyr)
html_title <- c()
html_body <- c()
for(i in 1:l)
{
  ##[[]] in list can get content
  title_now <- htmlfiles[[i]]  %>% html_nodes("title") %>% html_text()
  body_now <- htmlfiles[[i]]  %>% html_nodes("body") %>% html_text()
  if(is.na(title_now)||is.na(body_now)) next#vector function can do more research
  html_title <- c(html_title,title_now)
  html_body <- c(html_body,body_now)
}
```

```
# 4 find duplicated data

setwd(path)

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

```
## Warning: package 'psych' was built under R version 4.2.2
```

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

```
## Warning: package 'Hmisc' was built under R version 4.2.2
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Loading required package: ggplot2

##
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':
##
##     %+%, alpha

##
## Attaching package: 'Hmisc'

## The following object is masked from 'package:psych':
##
##     describe

## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
#di<-Hmisc::describe(html_title) #list #di[[1]]
Hmisc::describe(html_title)[[1]]
```

```
## [1] "html_title"
```

```
Hmisc::describe(html_title)[[4]]
```

```
##        n  missing distinct
##     1458        0     1364
```

```
Hmisc::describe(html_body)[[1]]
```

```
## [1] "html_body"
```

```
Hmisc::describe(html_body)[[4]]
```

```
##        n  missing distinct
##     1458        0     1458
```

```
cat("No duplicated jd")
```

```
## No duplicated jd
```

```r
#5 View the jobs ad
library(htmltools)#
```

```
## Warning: package 'htmltools' was built under R version 4.2.2
```

```r
rstudioapi::viewer(filestoread[[1]])
```

```
## NULL
```

```r
rstudioapi::viewer(filestoread[[2]])
```

```
## NULL
```

```r
# 6 get html bullets contents
html_bullets<-c()
all_bullets<-c()
for(i in 1:l)
{
  content_now<-htmlfiles[[i]] %>% html_nodes("li") %>% html_text()
  html_bullets<-c(html_bullets,list(content_now))
  text_now<-cbind(rep(i,length(content_now)),content_now)
  all_bullets<-rbind(all_bullets,text_now)
}
all_bullets<-as.data.frame(all_bullets)
names(all_bullets)[1]<-"ID"
```

```r
#7 Measuring the percent of bulleted postings
bullet_posting_count<-l
for(i in 1:l)
  if(identical(html_bullets[[i]], character(0))){
    bullet_posting_count=bullet_posting_count-1
  }
percentage=paste(round(100*bullet_posting_count/l, 2), "%", sep="")
cat("We have",percentage,"postings have bullets.")
```

```
## We have 90.53% postings have bullets.
```

```r
#8. Examining the top-ranked words in the HTML bullet

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

```
## Warning: package 'superml' was built under R version 4.2.2
```

```
## Loading required package: R6
```

```r
memory.limit(102400)
```

```
## Warning: 'memory.limit()' is no longer supported
```

```
## [1] Inf
```

```
tfv1 <- TfIdfVectorizer$new(remove_stopwords = TRUE)
tf_mat <- tfv1$fit_transform(html_bullets[1:100])
# As R is not so fast, I only use 1:100 to train dataset

sumtfidf<-apply(tf_mat,2,sum)
sort_sumtfidf<-sort(sumtfidf,decreasing = TRUE)
print(sort_sumtfidf[1:20])
```

```
##        data  experience           0   character    learning    analysis
##    9.120730    8.200802    5.098987    5.043706    4.344551    3.527487
##      skills     ability     machine    business         etc statistical
##    3.416570    3.373315    3.353241    3.348667    3.047458    3.030214
##        work           s   knowledge     science       tools       using
##    2.909485    2.815685    2.682757    2.658730    2.655961    2.591210
##           c      models
##    2.571423    2.484782
```

```
# data  experience           0   character    learning    analysis
#    9.120730    8.200802    5.098987    5.043706    4.344551    3.527487
#      skills     ability     machine    business         etc statistical
#    3.416570    3.373315    3.353241    3.348667    3.047458    3.030214
#        work           s   knowledge     science       tools       using
#    2.909485    2.815685    2.682757    2.658730    2.655961    2.591210
#           c      models
#    2.571423    2.484782

#Because in R, the stopwords may be a little different, there are some
#strange words not removed. But, totally, the top-ranked words are similar
#to those in python, like data, experienced, skills, ability, and work.
#They are in top 20.
```

```
#9. Examining the top-ranked words in the HTML bodies
library(stringr)
```

```
## Warning: package 'stringr' was built under R version 4.2.2
```

```
a= html_body[1:100]
for(i in 1:100){
  if(length(html_bullets[[i]])==0) next
  for(j in 1:length(html_bullets[[i]]))
  {
      a[i]=gsub(pattern=html_bullets[[i]][j],'',a[i],fixed=TRUE)
      #fixed=TRUE can deal with ()
  }
}
#Here a is 1:100 html body without bullets
```

```
#9.5 Apply
memory.limit(102400)
```

```
## Warning: 'memory.limit()' is no longer supported
```

```
## [1] Inf
```

```
tfv1 <- TfIdfVectorizer$new(remove_stopwords = TRUE)
tf_mat1 <- tfv1$fit_transform(a[1:100])
# As R is not so fast, I only use 1:100 to train dataset

sumtfidf1<-apply(tf_mat1,2,sum)
sort_sumtfidf1<-sort(sumtfidf1,decreasing = TRUE)
print(sort_sumtfidf1[1:20])
```

```
##           data           will              s      scientist           team
##       8.035170       4.390359       3.092829       3.003871       2.961091
##       business           work     experience        company         skills
##       2.948415       2.897609       2.772783       2.411203       2.335719
##         people            job       learning          world       research
##       2.333163       2.277602       2.214172       2.175225       2.159508
##        science          using       required qualifications             ca
##       2.130768       2.052147       2.045169       2.005219       1.991207
```

```
#  data           will              s      scientist           team
#       8.035170       4.390359       3.092829       3.003871       2.961091
#       business           work     experience        company         skills
#       2.948415       2.897609       2.772783       2.411203       2.335719
#         people            job       learning          world       research
#       2.333163       2.277602       2.214172       2.175225       2.159508
#        science          using       required qualifications             ca
#       2.130768       2.052147       2.045169       2.005219       1.991207
```

```
#Listing 17. 10. Checking titles for references to data science positions
regex="Data Scien(ce|tist)"
index_non_ds_jobs=which(grepl(regex,html_title)==FALSE)
l_non_ds=length(index_non_ds_jobs) #error in R
percentage1=paste(round(100*l_non_ds/l, 2), "%", sep="")
#error is 0.5%, can be ignored
cat(percentage1,"% of the job posting titles do not mention a",
       "data science position. Below is a sample of such titles:\n")
```

```
## 64.81% % of the job posting titles do not mention a data science position. Below is a sample of such
```

```
for(i in index_non_ds_jobs[1:10])
print(html_title[i])
```

```
## [1] "Political Staffer - San Francisco Bay Area, CA"
## [1] "Patient Care Assistant / PCA - Med/Surg (Fayette, AL) - Fayette, AL"
## [1] "Data Manager / Analyst - Oakland, CA"
## [1] "Scientific Programmer - Berkeley, CA"
## [1] "JD Digits - AI Lab Research Intern - Mountain View, CA"
## [1] "Operations and Technology Summer 2020 Internship-West Coast - Universal City, CA"
## [1] "Data and Reporting Analyst - Olympia, WA 98501"
## [1] "Senior Manager Advanced Analytics - Walmart Media Group - San Bruno, CA"
## [1] "Data Specialist, Product Support Operations - Sunnyvale, CA"
## [1] "Deep Learning Engineer - Westlake, TX"
```

```
#Listing 17. 11. Sampling bullets from a non-data science job
for(i in 1:5)
print(html_bullets[index_non_ds_jobs[2]][[1]][i])
```

```
## [1] "Provides all personal care services in accordance with the plan of treatment assigned by the reg
## [1] "Accurately documents care provided"
## [1] "Applies safety principles and proper body mechanics to the performance of specific techniques o
## [1] "Participates in economical utilization of supplies and ensures that equipment and nursing units
## [1] "Routinely follows and adheres to all policies and procedures"
```

```
#Listing 17. 12. Loading the resume
#Listing 17. 13. Loading the table-of-content
#Read text

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

```
## Warning: package 'readr' was built under R version 4.2.2
```

```
##
## Attaching package: 'readr'
```

```
## The following object is masked from 'package:rvest':
##
##     guess_encoding
```

```
resume <- read_csv("C:/Users/22600/3011/bookcamp_code/Case_Study4/resume.txt")
```

```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
## Rows: 12 Columns: 1
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Experience
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
table_of_contents <- read_csv("C:/Users/22600/3011/bookcamp_code/Case_Study4/table_of_contents.txt")
```

```
## Rows: 80 Columns: 1
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Case Study 1: Finding the Winning Strategy in a Card Game.
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
existing_skills<-list(c(resume,table_of_contents))
```

```
#Listing 17. 14. Combining skills into a single string
#Listing 17. 15. Vectorizing our skills and the job-posting data
text_list<-c(html_body[1:100],existing_skills)
tfv1 <- TfIdfVectorizer$new(remove_stopwords = TRUE)
tf_mat2 <- tfv1$fit_transform(text_list)
l2<-length(text_list)
```

```
#Listing 17. 16. Computing skill-based cosine similarities

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

```
## Warning: package 'lsa' was built under R version 4.2.2
```

```
## Loading required package: SnowballC
```

```
cos_similarities = cosine(t(tf_mat2))
cos_similarities[l2,]
```

```
##   [1] 0.034316218 0.124924228 0.044943478 0.043303811 0.011411250 0.025530657
##   [7] 0.019554149 0.026240827 0.037725757 0.051098077 0.050280595 0.062031367
##  [13] 0.075033490 0.047379429 0.044995141 0.040576769 0.035757645 0.072044657
##  [19] 0.071434453 0.035270012 0.064493499 0.042051428 0.040080747 0.024748968
##  [25] 0.043255531 0.015190429 0.019599638 0.061577173 0.122991906 0.020587877
##  [31] 0.106016678 0.135691254 0.043564060 0.056872884 0.065794022 0.102796831
##  [37] 0.000000000 0.050969765 0.039420084 0.030818141 0.018955089 0.046390630
##  [43] 0.019342747 0.050842570 0.006241941 0.032862586 0.052939461 0.054580019
##  [49] 0.025701684 0.065626265 0.112175870 0.093920073 0.099455478 0.023741000
##  [55] 0.042392491 0.042940156 0.012600616 0.040872410 0.009817392 0.047966034
##  [61] 0.051559292 0.011126951 0.039793247 0.027283696 0.012329599 0.017348100
##  [67] 0.032020095 0.033463652 0.044503016 0.027259266 0.045956361 0.035010359
##  [73] 0.086311762 0.023528557 0.019106916 0.039861189 0.025504346 0.054454852
##  [79] 0.038587682 0.079467383 0.018682299 0.025330474 0.058632066 0.016242662
##  [85] 0.048239732 0.040860531 0.044928673 0.025695747 0.051513745 0.051925645
##  [91] 0.060751296 0.017964261 0.034020011 0.028519966 0.052500036 0.060109875
##  [97] 0.066621883 0.063780953 0.024822028 0.054999515 1.000000000
```

```
#16.5 set relevance_matrix
relevance=cos_similarities[l2,][-l2]
ID=1:(l2-1)
relevance_matrix=t(rbind(ID,relevance))
relevance_matrix=relevance_matrix[order(relevance_matrix[,"relevance"],decreasing="T"),]
index_relevance<-relevance_matrix[,"ID"]
```

```
#Listing 17. 17. Printing the 20 least-relevant jobs
print(html_title[index_relevance[80:100]])
```

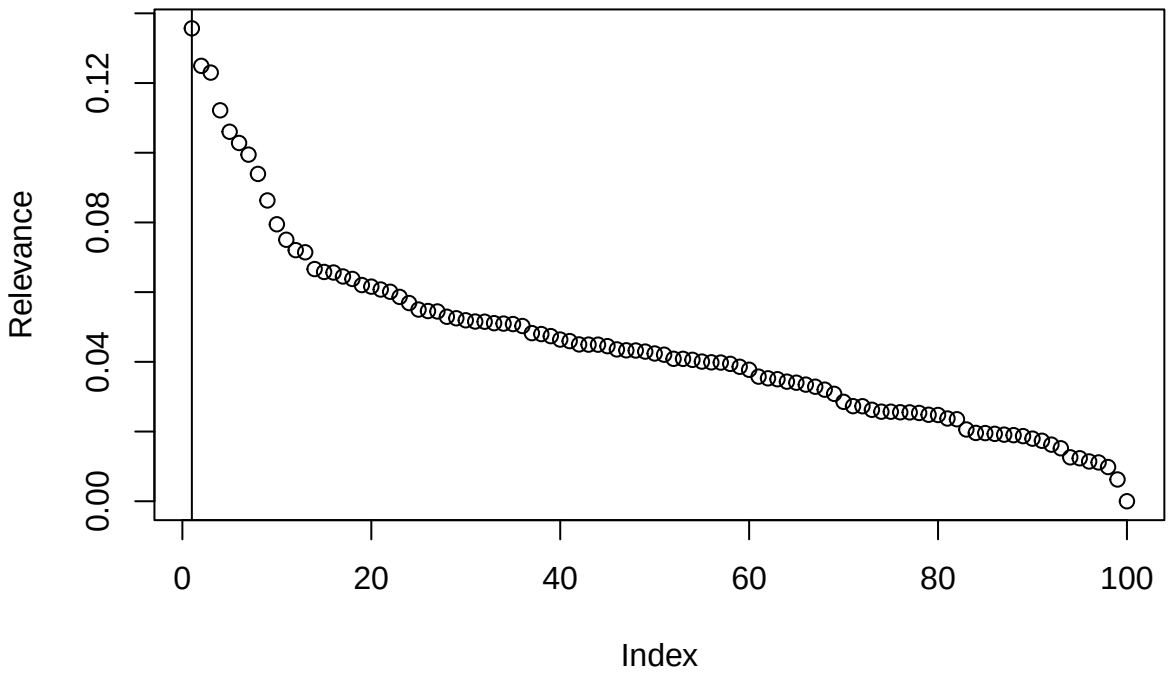
```
##  [1] "Software Developer - Los Gatos, CA 95033"
```

```
##  [2] "Computational Chemist - Menlo Park, CA"
##  [3] "Data Analyst and Compliance Specialist - San Francisco, CA 94103"
##  [4] "Senior Machine Learning (ML) and Computer Vision (CV) Engineer - Denver, CO"
##  [5] "Software Engineering Intern - San Jose, CA"
##  [6] "Scientific Programmer - Berkeley, CA"
##  [7] "Privacy and Data Policy Manager, Instagram - San Francisco, CA"
##  [8] "Certified Nursing Assistant PCA - Mesa, AZ 85206"
##  [9] "Technical Trainer - Redwood City, CA"
## [10] "User Experience Research Intern, Summer 2020 - San Francisco, CA 94105"
## [11] "Walmart Retail Link Associate - MIA - Miami Gardens, FL 33169"
## [12] "Office for New Americans Paid Internship - Salt Lake City, UT 84114"
## [13] "Events & Communication Specialist - - Berkeley, CA"
## [14] "Impact and Learning Manager - Impact, TX"
## [15] "Sustainability Program Manager, Water and Climate - Fremont, CA"
## [16] "Production Analyst - Camarillo, CA"
## [17] "Patient Care Assistant / PCA - Med/Surg (Fayette, AL) - Fayette, AL"
## [18] "Manager- Financial Consulting Valuation Services - New York, NY 10036"
## [19] "Admissions Associate PT - Baldwin Park, CA"
## [20] "Scorekeeper - Oakland, CA 94612"
## [21] "Director of Econometric Modeling - External Careers"
```

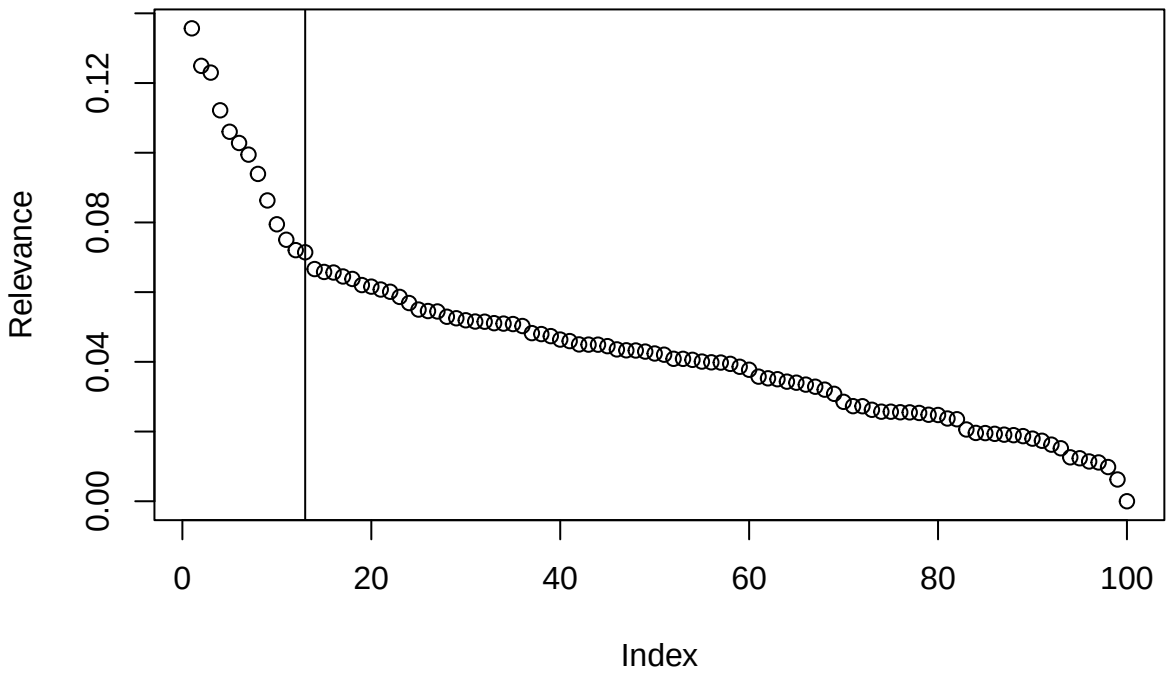
```
#17. 18. Printing the 20 most-relevant jobs
print(html_title[index_relevance[1:20]])
```

```
##  [1] "Santa Clara 4-H Community Ed Specialist 3 - Oakland, CA 94607"
##  [2] "Data Scientist - Beavercreek, OH"
##  [3] "Data Scientist - San Diego, CA"
##  [4] "Data Architect - Raleigh, NC 27609"
##  [5] "Data Science Intern - San Francisco, CA 94105"
##  [6] "Data Scientist - Streetsboro, OH 44241"
##  [7] "Data Scientist - Aliso Viejo, CA"
##  [8] "Data Engineer - Bridgewater, NJ"
##  [9] "Data Modeler - Melbourne, FL"
## [10] "Senior Data Analyst - Los Angeles, CA"
## [11] "Data Specialist, Product Support Operations - Sunnyvale, CA"
## [12] "Data Scientist - Seattle, WA"
## [13] "Data Scientist - Pasadena, CA 91107"
## [14] "Modeling and Simulation Engineer - Level 2 - Seattle, WA"
## [15] "Data Scientist - Fort Lauderdale, FL"
## [16] "Research Associate - Data Science - Washington, DC"
## [17] "Strategy Analyst - San Francisco, CA"
## [18] "Full Time Opportunities for Students or Recent Graduates: Data & Applied Sciences - Redmond, WA
## [19] "Senior Manager Advanced Analytics - Walmart Media Group - San Bruno, CA"
## [20] "Quantitative Business Analyst, Geo - Mountain View, CA"
```

```
#Listing 17. 19. Plotting job-ranking vs relevance
plot(relevance_matrix[,"relevance"],xlab="Index",ylab="Relevance")
abline(v=1)
```




```
#Listing 17. 20. Adding a cutoff to the relevance plot
plot(relevance_matrix[,"relevance"],xlab="Index",ylab="Relevance")
abline(v=13)
```




```
#Listing 17. 21. Printing jobs below the relevance cutoff
print(html_title[index_relevance[1:13]])
```

```
##  [1] "Santa Clara 4-H Community Ed Specialist 3 - Oakland, CA 94607"
##  [2] "Data Scientist - Beavercreek, OH"
##  [3] "Data Scientist - San Diego, CA"
##  [4] "Data Architect - Raleigh, NC 27609"
##  [5] "Data Science Intern - San Francisco, CA 94105"
##  [6] "Data Scientist - Streetsboro, OH 44241"
##  [7] "Data Scientist - Aliso Viejo, CA"
##  [8] "Data Engineer - Bridgewater, NJ"
##  [9] "Data Modeler - Melbourne, FL"
## [10] "Senior Data Analyst - Los Angeles, CA"
## [11] "Data Specialist, Product Support Operations - Sunnyvale, CA"
## [12] "Data Scientist - Seattle, WA"
## [13] "Data Scientist - Pasadena, CA 91107"
```

```
#Listing 17. 22. Printing jobs beyond the relevance cutoff
print(html_title[index_relevance[14:34]])
```

```
##  [1] "Modeling and Simulation Engineer - Level 2 - Seattle, WA"
##  [2] "Data Scientist - Fort Lauderdale, FL"
##  [3] "Research Associate - Data Science - Washington, DC"
##  [4] "Strategy Analyst - San Francisco, CA"
##  [5] "Full Time Opportunities for Students or Recent Graduates: Data & Applied Sciences - Redmond, WA
```

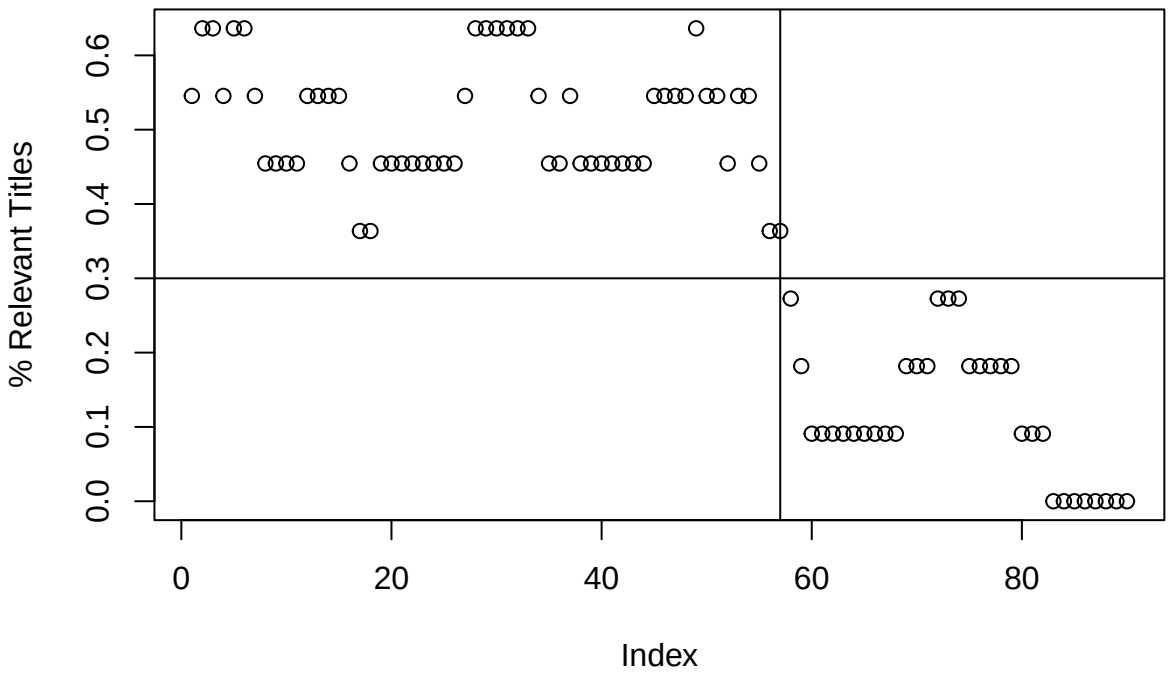
```
##  [6] "Senior Manager Advanced Analytics - Walmart Media Group - San Bruno, CA"
##  [7] "Quantitative Business Analyst, Geo - Mountain View, CA"
##  [8] "Data Scientist - Reston, VA 20192"
##  [9] "Senior/Staff Data Scientist - San Francisco, CA"
## [10] "Data Scientist - Beverly Hills, CA"
## [11] "Data Scientist - Bellevue, WA 98004"
## [12] "Environmental Compliance Coordinator - Raleigh, NC"
## [13] "Senior Data Engineer - Reno, NV 89501"
## [14] "Optimization and KYC Model Risk Analyst, Associate/AVP - San Francisco, CA"
## [15] "TECHNICAL INFORMATION SPECIALIST (WEB SERVICES) - Monterey, CA"
## [16] "Data Analyst (6256U) 1737 - 1737 - Berkeley, CA 94720"
## [17] "Account Executive/Acute Therapies - Sacramento - Sacramento, CA 95819"
## [18] "Director of Marketing Statistics - United States"
## [19] "Data Scientist, Entity Resolution and Data Linking - Alpharetta, GA 30005"
## [20] "PwC Labs - Jr. Data Scientist - Machine Learning (NLP) - Tampa, FL 33607"
## [21] "Data Scientist III - Pasadena, CA 91101"
```

```
#Listing 17. 23. Measuring title relevance in a subset of jobs
percentage_relevant_titles<-function(df_title)
{
regex_relevant="Data (Scien|Analy)" #don't plug in science
regex_irrelevant="\b(Manage)"
match_counts=which(grepl(regex_relevant,df_title)==TRUE&grepl(regex_irrelevant,df_title)==FALSE)
percentage=length(match_counts)/length(df_title) #error in R and
return(percentage)
}
percentage2=percentage_relevant_titles(html_title[index_relevance[14:34]])
percentage2=paste(round(100*percentage2,2), "%", sep="")
# 0.4761905
cat("Approximately",percentage2,"% of job titles between indices ",
       "14 - 34 are relevant")
```

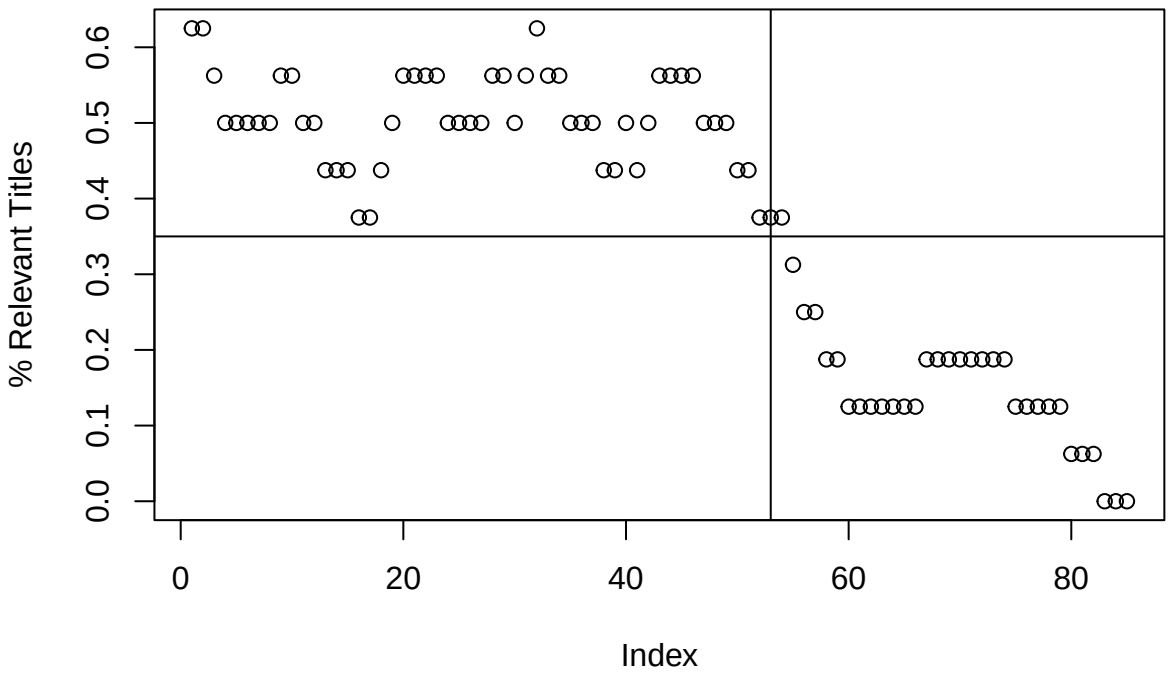
```
## Approximately 47.62% % of job titles between indices  14 - 34 are relevant
```

```
#Because the base is not same large as that in python, it is reasonable.
```

```
#Listing 17. 25. Plotting percent relevance across all title samples
relevant_title_plot<-function(index_range=10,h,v)
{
  percentage3=c()
  start_indices=100-index_range
  for(i in 1:start_indices)
  {
    df_slice = html_title[index_relevance[i:(i+index_range)]]
    ##should include(),otherwise only + 10
    percent= percentage_relevant_titles(df_slice)
    percentage3=c(percentage3,percent)
  }
    plot(1:start_indices,percentage3,xlab="Index",ylab="% Relevant Titles")
    abline(h=h)
    abline(v=v)

}
relevant_title_plot(h=0.3,v=57)
```




```
#Listing 17. 26. Plotting percent relevance across an increased index-range
relevant_title_plot(index_range = 15,v=53,h=0.35)
```




```
# 17. 27. Obtaining bullets from the 30 most-relevant jobs
total_bullets=c()
for(i in index_relevance[1:30])
{
  content_now<-htmlfiles[[i]] %>% html_nodes("li") %>% html_text()
  total_bullets<-c(total_bullets,content_now)
}
#17. 28. Summarizing basic bullet statistic
Hmisc::describe(total_bullets)[[1]]
```

```
## [1] "total_bullets"
```

```
Hmisc::describe(total_bullets)[[4]]
```

```
##        n  missing distinct
##      462        0      455
```

```
#29 Removing duplicates and vectorizing the bullets
total_bullets=sort(total_bullets[!duplicated(total_bullets)])
tfv1 <- TfIdfVectorizer$new(remove_stopwords = TRUE)
tf_mat3 <- tfv1$fit_transform(total_bullets)
n1<-nrow(tf_mat3)
m1<-ncol(tf_mat3)
print(dim(tf_mat3))
```

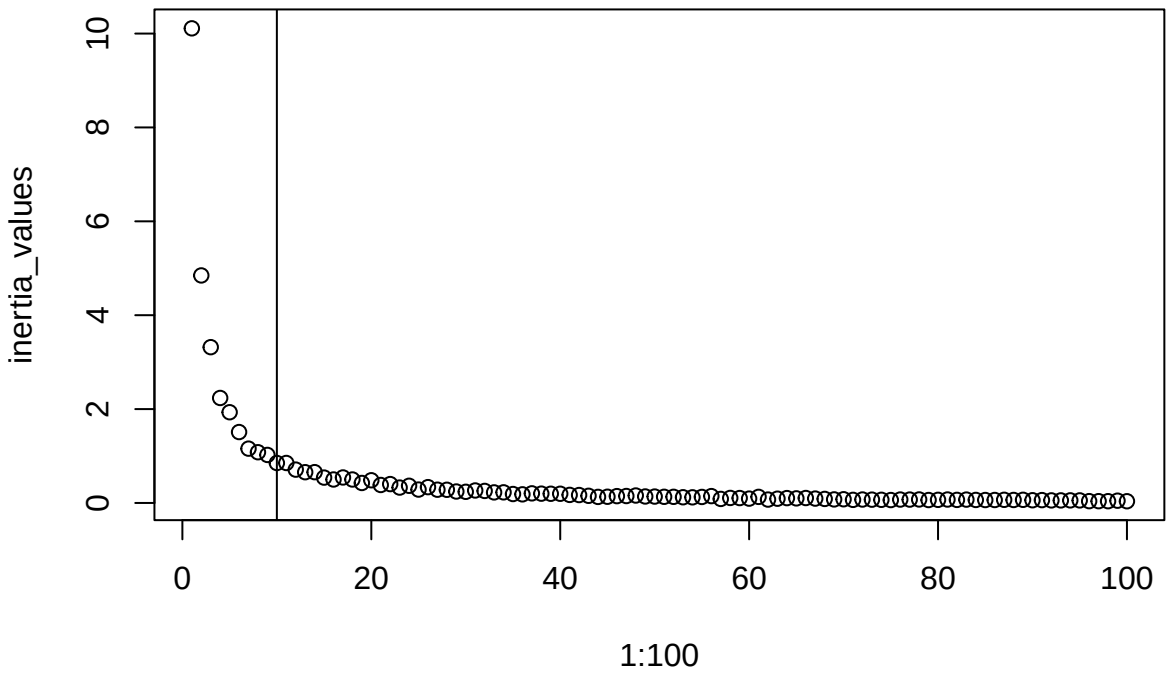
```
## [1]  455 1859
```

```
#30. Dimensionally reducing the TFIDF matrix
s <- svd(tf_mat3)
d <- diag(s$d) #eigenvalue
v <- as.matrix(s$v)
u <- s$u

u2 <- as.matrix(u[,1:100])
d2 <- as.matrix(d[1:100,1:100])
v2 <- as.matrix(v[,1:100])
a2 <- u2 %*% d2

a3 <- normalise2d(a2)
```

```
#31. Plotting an elbow curve using Mini Batch K-Means
library(ClusterR)
```

```
## Warning: package 'ClusterR' was built under R version 4.2.2
```

```
library(cluster)

inertia_values=c()
for(k in 1:100)
{
  temp= MiniBatchKmeans(a3,clusters = k)
  inertia_values=c(inertia_values,mean(temp$WCSS_per_cluster) )
}
plot(1:100,inertia_values)
abline(v=10)
```




```
#Choosing cluter k > 10
```

```
#32. Clustering bullets into 15 clusters
#Choosing k
#install.packages("wordcloud")
library(wordcloud)
```

```
## Warning: package 'wordcloud' was built under R version 4.2.2
```

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

```
k=10
temp= kmeans(a3,k)

cluster<-list()
cluster_index<-list()

#Cluster
for(i in 1:k){
cluster_1_index<-which(temp$cluster==i)
cluster_index<-c(cluster_index,list(cluster_1_index))
}

#wordcloud
```

```
#install.packages("wordcloud2")
library(wordcloud2)
```

```
## Warning: package 'wordcloud2' was built under R version 4.2.2
```

```
for(i in 1:k){
tfv1 <- TfIdfVectorizer$new(remove_stopwords = TRUE)
tf_mat4 <- tfv1$fit_transform(total_bullets[cluster_index[[i]]])
sum_tf_mat4<-apply(tf_mat4,2,sum)
wordcloud(names(sum_tf_mat4),freq=sum_tf_mat4,min.freq = 0.5, random.order=FALSE, rot.per=0.35, colors=
#Set low min frequency
}
```

```
## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## advantages could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## properties could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## concepts could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## applications could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## models could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## scenario could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## boosting could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## features could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## applied could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## principles could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## logistic could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## network could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## engineers could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## research could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## develop could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## spark could not be fit on page. It will not be plotted.
```


```
## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## models could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## performance could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## systems could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## implement could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## technology could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## collaborate could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## stakeholders could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## management could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## experience could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## system could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## services could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## based could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## governance could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## recommendations could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## prioritize could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## improve could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## closely could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## partner could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## supporting could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## monitoring could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## engineers could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## frameworks could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## algorithms could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## required could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## functional could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## advancing could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## reports could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## deploy could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## documentation could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## decision could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## trends could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## service could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## report could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## appropriate could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## optimize could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## various could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## content could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## ensure could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## overall could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## experimentation could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## creating could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## course could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## pipelines could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## helping could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## using could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## cloud could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## quality could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## internal could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## within could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## marketing could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## statistical could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## efforts could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## detection could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## anomaly could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## profitability could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## propose could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## collections could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## amrm could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## driven could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## accuracy could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## engineering could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## analytics could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## changes could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## integration could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## collection could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## machine could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## learning could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## science could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## hypothesis could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## goals could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## provide could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## information could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## professional could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## local could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## change could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## daily could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## architecture could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## alliance could not be fit on page. It will not be plotted.
```


```
## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## mathematics could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## economics could not be fit on page. It will not be plotted.
```

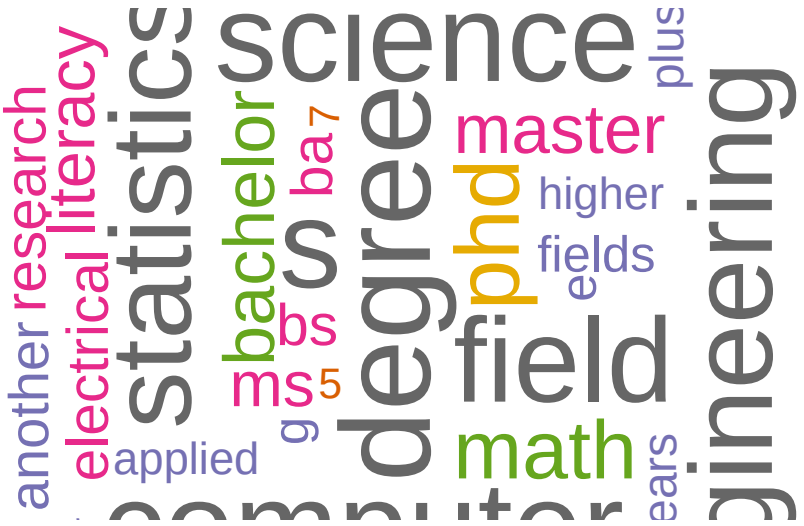
```
## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## discipline could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## physics could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## operations could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## statistical could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## experience could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## masters could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## learning could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## chemistry could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## models could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## manipulating could not be fit on page. It will not be plotted.
```



experience
quisition
nalysis
cleaning
distribute
data
ocessing


experience
data
using
python
visualization
languages
tools
large
statistical
insights
etc
working
analytics
pandas
methods
sql
r
sets
databases
hadoop
matplotlib
manipulate
numpy
hive
tableau
language
packages
services
proficiency
business
architecture
enterprise
visualizations
projects
extracting
pipelines
extensive
learn
excel
models
science
hands
aws
architectures
perl
custom
nosql
tidyverse
apply
creating
target
sas
state
distributed
php
azure
min
test
tool
one
google
presenting
querying
analytical
programming
studio
messy
catalyst
prototyping
testing
rapid
work
facebook
reporting
draw
least
3rd
adwords
slq
gplot
personality


```
## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## understanding could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## background could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## collaborative could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## development could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## potential could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## collaboration could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## customer could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## experimentation could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## approach could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## concepts could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## results could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## customers could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## solution could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## develop could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## group could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## current could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## probability could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## procedures could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## unlimited could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## understood could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## functional could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## monitoring could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## activity could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## analysis could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## demonstrate could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## client could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## statistical could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## perform could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## libraries could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## flexible could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## relevant could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## scalable could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## spark could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## microsoft could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## manage could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## analyses could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## coding could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## scripting could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## future could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## optimization could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## decision could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## simulation could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## mathematical could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## interpersonal could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## communications could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## retain could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## models could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## supported could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## speea could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## general could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## standard could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## patterns could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## succeed could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## audiences could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## multiple could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## provide could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## expertise could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## plus could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## javascript could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## appropriate could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## office could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## findings could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## account could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## following could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## software could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## students could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## reporting could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## access could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## competence could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## postgresql could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## mysql could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## modeling could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## towards could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## aptitude could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## pitfalls could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## forecast could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## design could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## analytic could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## predicting could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## present could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## likely could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## tuning could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## suspicious could not be fit on page. It will not be plotted.
```


```
## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## management could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## segmentation could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## communicates could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## approaches could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## understand could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## analyses could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## identify could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## intelligence could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## necessary could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## requirements could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## develops could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## outcomes could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## optimize could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## increase could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## experiences could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## development could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## online could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## forecasting could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## technical could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## acumen could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## health could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## enterprise could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## decisions could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## summarizes could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## effectively could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## important could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## engineering could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## marketing could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## specialty could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## nursing could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## analysts could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## shoppers could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## educate could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## company could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(sum_tf_mat4), freq = sum_tf_mat4, min.freq = 0.5, :
## various could not be fit on page. It will not be plotted.
```


```
print(total_bullets[cluster_index[[4]]])
```

```
##  [1] " 5-7 years of experience manipulating data sets and building statistical models, a degree in S
##  [2] "\nBA/BS in Math, Statistics, Economics, Computer Science, or other quantitative field"
##  [3] "\nBachelor's, Master's, or Doctorate or Science degree from an accredited course of study, in
##  [4] "\nWe're looking for someone with 5-7 years of experience manipulating data sets and building s
##  [5] "A Bachelor's degree in quantitative fields, such as economics, mathematics, statistics, and co
##  [6] "A degree in data science or a related field (e.g., computer science, statistics, mathematics,
##  [7] "Bachelor's degree or higher in quantitative or related field"
##  [8] "COMPUTER LITERACY"
##  [9] "Currently has or is in the process of obtaining their BA/BS or Masters in Computer Science, Ma
## [10] "Degree in math, engineering, economics, or a related field"
## [11] "Education: B.S. / M.S. in computer science, physics, electrical engineering, applied mathemati
## [12] "Master's or PhD in Statistics or related field"
## [13] "Minimum Bachelors/Masters in Computer Science, MIS or another quantitative field eg. Statistic
## [14] "MS or PhD degree in a quantitative discipline a plus"
## [15] "MS or PhD in a quantitative discipline (e.g., statistics, operations research, computer scienc
## [16] "MSc or PhD degree in any of the following fields: machine learning or computer science, statis
## [17] "Undergraduate degree in a quantitative discipline (e.g., statistics, operations research, bioi
## [18] "Use a combined knowledge of computer science and applications, modelling, statistics, analytic
```

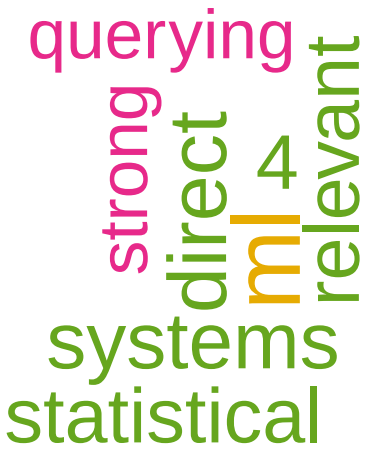
```
#Cluster 4 focused on tech

print(total_bullets[cluster_index[[7]]])
```

```
##  [1] "\n1+ years experience in quantitative analysis"
##  [2] "\n1+ years experience in SQL or other querying language"
##  [3] "1-2+ years prior work experience"
##  [4] "2 to 4 years of relevant work experience in data analysis or related field (e.g., as a statisti
##  [5] "2 years of experience in data analysis, visualization, or related field"
##  [6] "3-5+ years of industry experience with a Phd or 8+ years of industry work experience with a mas
##  [7] "3+ years' experience in specialized data mining, predictive modeling, or other data intensive a
##  [8] "4+ years industry experience in time series modeling or machine learning, with significant pers
##  [9] "Language Fluency: In Java / Python (at least 2 years of experience on commercial projects) and
## [10] "Minimum five years of experience in data warehousing, business intelligence and reporting envi
## [11] "Minimum GPA: 3.4"
## [12] "Minimum of 5 years of experience with at least 2 years of direct related experience in tuning a
## [13] "MS with 2+ years of industry experience or Bachelors with 5+ years of experience in Quantitativ
## [14] "Requires 1 - 3 years administrative/secretarial, environmental compliance management or relatec
## [15] "Strong statistical background and 7+ years of overall experience"
## [16] "Work experience: 5+ years of real-world development experience and 2+ years of experience with
```

```
#Cluster 4 focused on sotf skills
```

```
# 38. Comparing mean resume similarities
tfv1 <- TfIdfVectorizer$new(remove_stopwords = TRUE)
mat5<-c(total_bullets,list(c(resume)))
tf_mat5 <- tfv1$fit_transform(c(total_bullets,list(c(resume))))
l5<-length(c(total_bullets,list(c(resume))))
cos_similarities_mat5 = cosine(t(tf_mat5))
relevance_mat5=cos_similarities_mat5[l5,][-l5]

ID=1:(l2-1)
relevance_matrix=t(rbind(ID,relevance))
relevance_matrix=relevance_matrix[order(relevance_matrix[,"relevance"],decreasing="T"),]

cluster_similarity<-c()
for(i in 1:k){
cluster_temp_simliarity=mean(relevance_mat5[cluster_index[[i]]])
cluster_similarity<-c(cluster_similarity,cluster_temp_simliarity)
}
#39. Sorting subplots by resume similarity
order_similarity <- order(cluster_similarity,decreasing = T)
```

```
#40 Plot following orders
for(i in order_similarity)
{
tfv1 <- TfIdfVectorizer$new(remove_stopwords = TRUE)
tf_mat4 <- tfv1$fit_transform(total_bullets[cluster_index[[i]]])
sum_tf_mat4<-apply(tf_mat4,2,sum)
wordcloud(names(sum_tf_mat4),freq=tf_mat4,max.words=100, random.order=FALSE, rot.per=0.35, colors=brewe
}
```


querying
strong
direct
4
ml
relevant
systems
statistical

abstraction


applicat
designing
scikit
algorithms
analytics
selecting
course


science
statement
demonstrated
automating
map
flink
pregel
signatu
natur
analysis
works
acquisition
cleansi
verifying
scale
reduce
distributed
ming


partnershi
required
highlight
patterns
attractions
skills
methods
complex
communication
historical


working
exploration
bar
studies
developer


deployed
presence
www
mission
effective
year
skills
training
affect
federal
research
data
5


challenging
internal

modelling
5
data
plus
analysis
psychology
bioinformatics


```
#41. Printing sample bullets from Clusters 4 and 7
set.seed(2)

sample_index_cluster4<-sample(1:length(cluster_index[[4]]),size=5)
print(total_bullets[cluster_index[[4]][sample_index_cluster4]])
```

```
## [1] "MS or PhD in a quantitative discipline (e.g., statistics, operations research, computer science
## [2] "A degree in data science or a related field (e.g., computer science, statistics, mathematics, c
## [3] "Undergraduate degree in a quantitative discipline (e.g., statistics, operations research, bioin
## [4] "COMPUTER LITERACY"
## [5] " 5-7 years of experience manipulating data sets and building statistical models, a degree in Sta
```

```
#Cluster 4 focused on tech

sample_index_cluster7<-sample(1:length(cluster_index[[7]]),size=5)
print(total_bullets[cluster_index[[7]][sample_index_cluster7]])
```

```
## [1] "MS with 2+ years of industry experience or Bachelors with 5+ years of experience in Quantitativ
## [2] "\n1+ years experience in quantitative analysis"
## [3] "Minimum of 5 years of experience with at least 2 years of direct related experience in tuning a
## [4] "Work experience: 5+ years of real-world development experience and 2+ years of experience with
## [5] "Language Fluency: In Java / Python (at least 2 years of experience on commercial projects) and
```

```
#Cluster 7 focused on sotf skills

#Clusters 1-6 Tech We can plot
#Clusters 7-10 Soft skills
```

```
#k=10/20. However, R is very slow. So I am not able to analysize big data like in python. Here we just
#We just need to change k to other numbers in chunk 30
```

```
#Here python just analysize 700 jobs like above. The technical part is similar. We just need to change
#We just need to change in index_relevance[1:30] chunk 26 to index_relevance[1:m],
#m be other numbers
```