Webscrapping, data management and visualization project

Introduction:

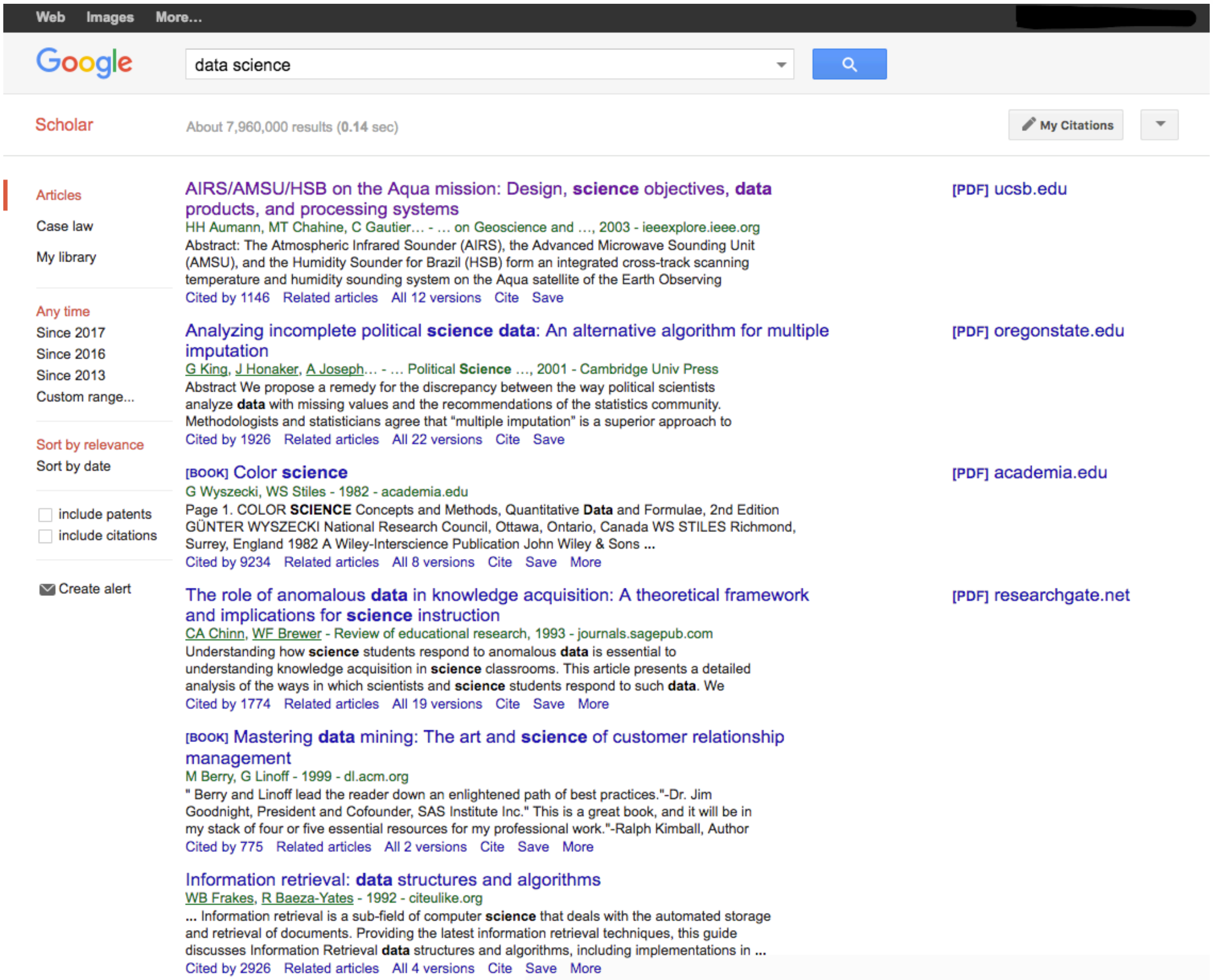
The goal of this project is to demonstrate my ability to scrape information from website, store the scraped data, and conduct analysis using the data. These tasks will be completed with Docker, R and PostgreSQL. The analysis will mainly focus on data exploration and data visualization. My plan for this project is to scrape the Google Scholars search result page, try and store the data in relational databases, and make wordcloud graphs using the data. I am curious about the topic of our current master's program, "data science," and I would like to know popular topics discussed in the publications related to data science.

Web Scraping:

In this section, we will scrape data from the Google Scholar's search results page. The tools to be used in this section are R, namely the `RSelenium` package, and the corresponding Chrome image in Docker.

Inspection of the website:

Looking at the search result page itself, there is the number of search results just below the floating search bar, search filters on the left sidebars, search results in the main body window. The search results contain the type of the results, such as webpage, citation, or books. It also contains the title of the result, general reference of the result which contains the year the paper is published, an excerpt from the summary in which contains the keyword, number of times the paper has been cited, related articles, versions, and some other miscellaneous links. Note that those which has PDFs has a link on the right side of the result that links to the PDFs. There are 10 items per page and the total pages a single search can reach to is 100 pages.



The information that is to be extracted from the results are: title, reference, the year the paper is published, summary, and how many times the paper has been cited. Proceed to scraping the page.

Scraping with RSelenium and Docker chrome image:

First, start the `standalone-chrome-debug` image in Docker, load the `RSelenium` package in R, and initiated a remote driver that connects to the docker image:

```
# load dependency
library(RSelenium)

# define remote driver for selenium
rd <- remoteDriver(remoteServerAddr = "192.168.99.100",
  browserName = 'chrome',
  port = 8080)
rd$open()
```

Inspecting the page elements and finding the corresponding XPATHs of each element that we want, Write a nested "for" loop that scrapes the elements we want from a certain page range. Here, we scrape from page 1 to 20. Note that, because the way google scholar url works, the second page has a `start=10` tag on its url, which means, page 3 has the tag `start=20`, thus page 20 is in fact index number 19 in the loop. It is really nice for google to program their url this way because it makes scraping much easier.

```
# loop and scrape
system.time({

  # initialize the data frame
  ds <- data.frame()

  for (page in 0:19){
    # loop through page
    item_index = page*10
    url <- paste('https://scholar.google.com/scholar?q=data+science&as_vis=1&as_sdt=1,5&start=', item_index, sep = "")
    rd$navigate(url)
    print(paste('At page', page+1))

    # loop through publications
    for (item in 1:10){
      print(paste('     item', item))

      # define xpath
      xp_title <- paste('//*[@id="gs_ccl_results"]/div[', item, ']/div[@class="gs_ri"]/h3/a', sep = "")
      xp_reference <- paste('//*[@id="gs_ccl_results"]/div[', item, ']/div[@class="gs_ri"]/div[1]', sep = "")
      xp_summary <- paste('//*[@id="gs_ccl_results"]/div[', item ,']/div[@class="gs_ri"]/div[2]', sep = "")
      xp_citecount <- paste('//*[@id="gs_ccl_results"]/div[', item, ']/div[@class="gs_ri"]/div[3]/a[1]', sep = "")

      # get element by xpath
      element_title <- rd$findElement(using = 'xpath', xp_title)
      element_reference <- rd$findElement(using = 'xpath', xp_reference)
      element_summary <- rd$findElement(using = 'xpath', xp_summary)
      element_citecount <- rd$findElement(using = 'xpath', xp_citecount)

      # extract text
      title <- element_title$getElementText()[[1]]
      reference <- element_reference$getElementText()[[1]]
      summary <- element_summary$getElementText()[[1]]
      citecount <- element_citecount$getElementText()[[1]]
```

```
      year <- substr(gsub("[^0-9]","", reference), start = 0, stop = 4)

      # bind and feed to data frame
      ds <- rbind(ds, cbind(title, reference, year, summary, citecount))
    }
  }
})
```

This 20 page scrape took about 50 seconds in the first run, and around 40 seconds in the second run, since some of the elements has been cached by the browser.

Data cleaning:

The dataset `ds` that is extracted this way has to be cleaned before being stored in a databse. Since titles, references, and summaries are unique, by default `data.frame` stored these values in factors; we need to correct them to characters. Also, the `citecount` is a string, which we need to extract the number and convert to numeric. Here is the resulting view of the dataset `ds`.

```
# correct field types
ds$title <- as.character(ds$title)
ds$reference <- as.character(ds$reference)
ds$year <- as.numeric(as.character(ds$year))
ds$summary <- as.character(ds$summary)
ds$citecount <- as.numeric(substr(as.character(ds$citecount), start = 10, stop
= 20))
```

	title	reference	year	summary	citecount
1	AIRS/AMSU/HSB on the Aqua mission: Design, scienc...	HH Aumann, MT Chahine, C Gautier... – ... on Geoscie...	2003	Abstract: The Atmospheric Infrared Sounder (AIRS), th...	1146
2	Analyzing incomplete political science data: An altern...	G King, J Honaker, A Joseph... – ... Political Science ...,...	2001	Abstract We propose a remedy for the discrepancy be...	1926
3	Color science	G Wyszecki, WS Stiles – 1982 – academia.edu	1982	Page 1. COLOR SCIENCE Concepts and Methods, Quan...	9234
4	The role of anomalous data in knowledge acquisition:...	CA Chinn, WF Brewer – Review of educational researc...	1993	Understanding how science students respond to ano...	1774
5	Mastering data mining: The art and science of custom...	M Berry, G Linoff – 1999 – dl.acm.org	1999	" Berry and Linoff lead the reader down an enlightene...	775
6	Information retrieval: data structures and algorithms	WB Frakes, R Baeza-Yates – 1992 – citeulike.org	1992	... Information retrieval is a sub-field of computer sci...	2926
7	ENDF/B-VII. 0: next generation evaluated nuclear dat...	MB Chadwick, P Obložinský, M Herman, NM Greene... ...	2006	We describe the next generation general purpose Eval...	1726
8	Reducing the dimensionality of data with neural netw...	GE Hinton, RR Salakhutdinov – science, 2006 – scienc...	2006	Abstract High-dimensional data can be converted to l...	5141
9	Statistics: methods and applications: a comprehensiv...	T Hill, P Lewicki, P Lewicki – 2006 – books.google.com	2006	This-one of a kind-book offers a comprehensive, alm...	1729
10	A survey of data provenance in e-science	YL Simmhan, B Plale, D Gannon – ACM Sigmod Record...	2005	Abstract Data management is growing in complexity ...	1064
11	Categorical data analysis	A Agresti, M Kateri – 2011 – Springer	2011	For categorical data, the binomial (see Binomial Distri...	21057
12	Calibration of the Computer Science and Applications...	PS Freedson, E Melanson, J Sirard – ... and science in ...	1998	... PURPOSE: We established accelerometer count rang...	2050
13	Introductory digital image processing: a remote sensi...	JR Jensen – 1996 – cabdirect.org	1996	The second revised edition of the title book focuses o...	7926
14	Citation indexes for science. A new dimension in doc...	E Garfield – International journal of epidemiology, 20...	2006	... Previous Section. References. ↵ Thomasson P, Stan...	2304
15	Book Review: Corbin, J., & Strauss, A.(2008). Basics of...	RW Service – Organizational Research Methods, 2009 ...	2009	... 318). Section 2 (chapters 8-11) demonstrates, via ...	48959

Then we proceed to import these data into PostgreSQL.

Data Storage and Query:

In this section, we will be importing the data we extracted from the webpage into PostgreSQL. This task will be completed using R, namely the `RPostgresql` package, with a existing PostgreSQL server.

Data storage using PostgreSQL:

First, we define the connection to the PostgreSQL server in R.

```
# load dependency
require(RPostgreSQL)

# initiate database driver and connection
drv <- dbDriver("PostgreSQL")
con <- dbConnect(drv, dbname = "postgres", host = "localhost", port = "5432",
  user = "postgres", password = pw)
```

Then we test if the server is connected.

```
# a test
dbExistsTable(con, "scrape")
# query if table "scrape" exists in table, which we do not have
```

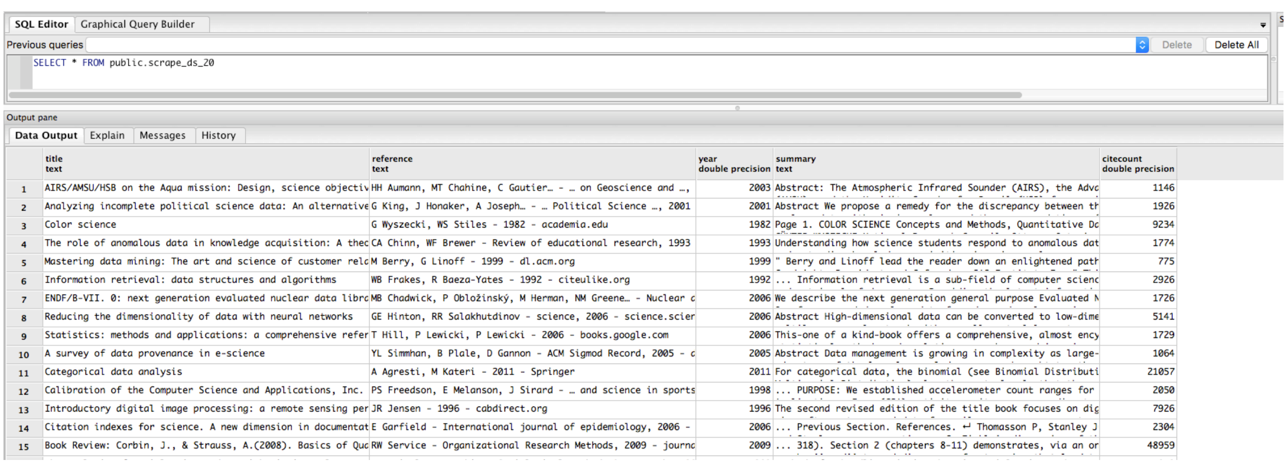
Upon returning `[1] FALSE`, we know that our connection is established.

Then we proceed to import the data into the database.

```
# write data
dbWriteTable(con, "scrape_ds_20", value = ds, append = F, row.names = F)
```

This command imports the data as a table, named "scrape_ds_20", overwriting existing information without including row.names (which we don't really have, but mentioned to ensure integrity anyways.)

Try query the table we just imported in PostgreSQL. We can see that the data has been imported with somewhat correct datatype. We can see that the numbers, year and cite count has wrong datatype.



	title text	reference text	year double precision	summary text	citecount double precision
1	AIRS/AMSU/HSB on the Aqua mission: Design, science objectiv	HH Aumann, MT Chahine, C Gautier… - … on Geoscience and …,	2003	Abstract: The Atmospheric Infrared Sounder (AIRS), the Advc	1146
2	Analyzing incomplete political science data: An alternative	G King, J Honaker, A Joseph… - … Political Science …, 2001	2001	Abstract We propose a remedy for the discrepancy between th	1926
3	Color science	G Wyszecki, WS Stiles - 1982 - academia.edu	1982	Page 1. COLOR SCIENCE Concepts and Methods, Quantitative Dc	9234
4	The role of anomalous data in knowledge acquisition: A thec	CA Chinn, WF Brewer - Review of educational research, 1993	1993	Understanding how science students respond to anomalous dat	1774
5	Mastering data mining: The art and science of customer relc	M Berry, G Linoff - 1999 - dl.acm.org	1999	" Berry and Linoff lead the reader down an enlightened path	775
6	Information retrieval: data structures and algorithms	WB Frakes, R Baeza-Yates - 1992 - citeulike.org	1992	... Information retrieval is a sub-field of computer scienc	2926
7	ENDF/B-VII. 0: next generation evaluated nuclear data librc	MB Chadwick, P Obložinský, M Herman, NM Greene… - Nuclear c	2006	We describe the next generation general purpose Evaluated N	1726
8	Reducing the dimensionality of data with neural networks	GE Hinton, RR Salakhutdinov - science, 2006 - science.scier	2006	Abstract High-dimensional data can be converted to low-dime	5141
9	Statistics: methods and applications: a comprehensive refer	T Hill, P Lewicki, P Lewicki - 2006 - books.google.com	2006	This-one of a kind-book offers a comprehensive, almost ency	1729
10	A survey of data provenance in e-science	YL Simmhan, B Plale, D Gannon - ACM Sigmod Record, 2005 - c	2005	Abstract Data management is growing in complexity as large-	1064
11	Categorical data analysis	A Agresti, M Kateri - 2011 - Springer	2011	For categorical data, the binomial (see Binomial Distributi	21057
12	Calibration of the Computer Science and Applications, Inc.	PS Freedson, E Melanson, J Sirard - … and science in sports	1998	... PURPOSE: We established accelerometer count ranges for	2050
13	Introductory digital image processing: a remote sensing per	JR Jensen - 1996 - cabdirect.org	1996	The second revised edition of the title book focuses on dig	7926
14	Citation indexes for science. A new dimension in documentat	E Garfield - International journal of epidemiology, 2006 -	2006	... Previous Section. References. ↵ Thomasson P, Stanley J	2304
15	Book Review: Corbin, J., & Strauss, A.(2008). Basics of Quc	RW Service - Organizational Research Methods, 2009 - journc	2009	... 318). Section 2 (chapters 8-11) demonstrates, via an or	48959

The following `ALTER TABLE` command fixes the datatype.

```
ALTER TABLE public.scrape_ds_20 ALTER COLUMN year TYPE numeric(4,0);
ALTER TABLE public.scrape_ds_20 ALTER COLUMN citecount TYPE numeric(10,0);
```

Data querying using R:

We can also execute the query in R. Here we query the table and load it back into R as `dataset`.

```
# query the table
dataset <- dbGetQuery(con, "SELECT * FROM public.scrape_ds_20")
```

Checking the datatypes in R reveal no issue.

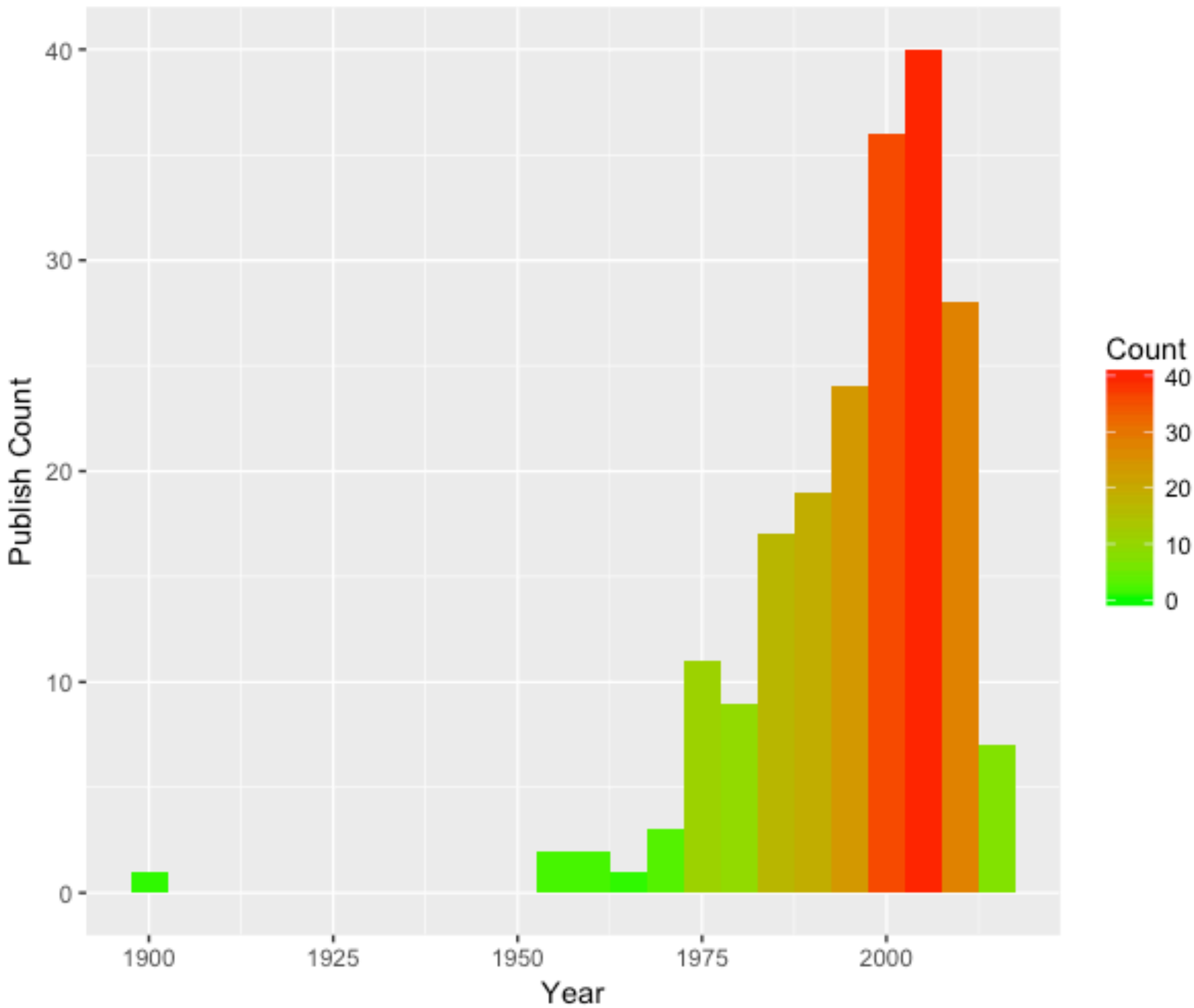
```
> class(dataset$title)
[1] "character"
> class(dataset$reference)
[1] "character"
> class(dataset$year)
[1] "numeric"
> class(dataset$summary)
[1] "character"
> class(dataset$citecount)
[1] "numeric"
```

Data Exploration:

In this section, we will be doing exploratory analysis using the scraped data. We will use `ggplot` , some natural language processing packages, and `wordcloud` .

Simple explorations:

First, we plot the year against how many papers are published. The plot shows that most papers that contains the keywords were published in the past 30 years. With the aid of computers, the research that involves "data" and "science" became more and more popular.

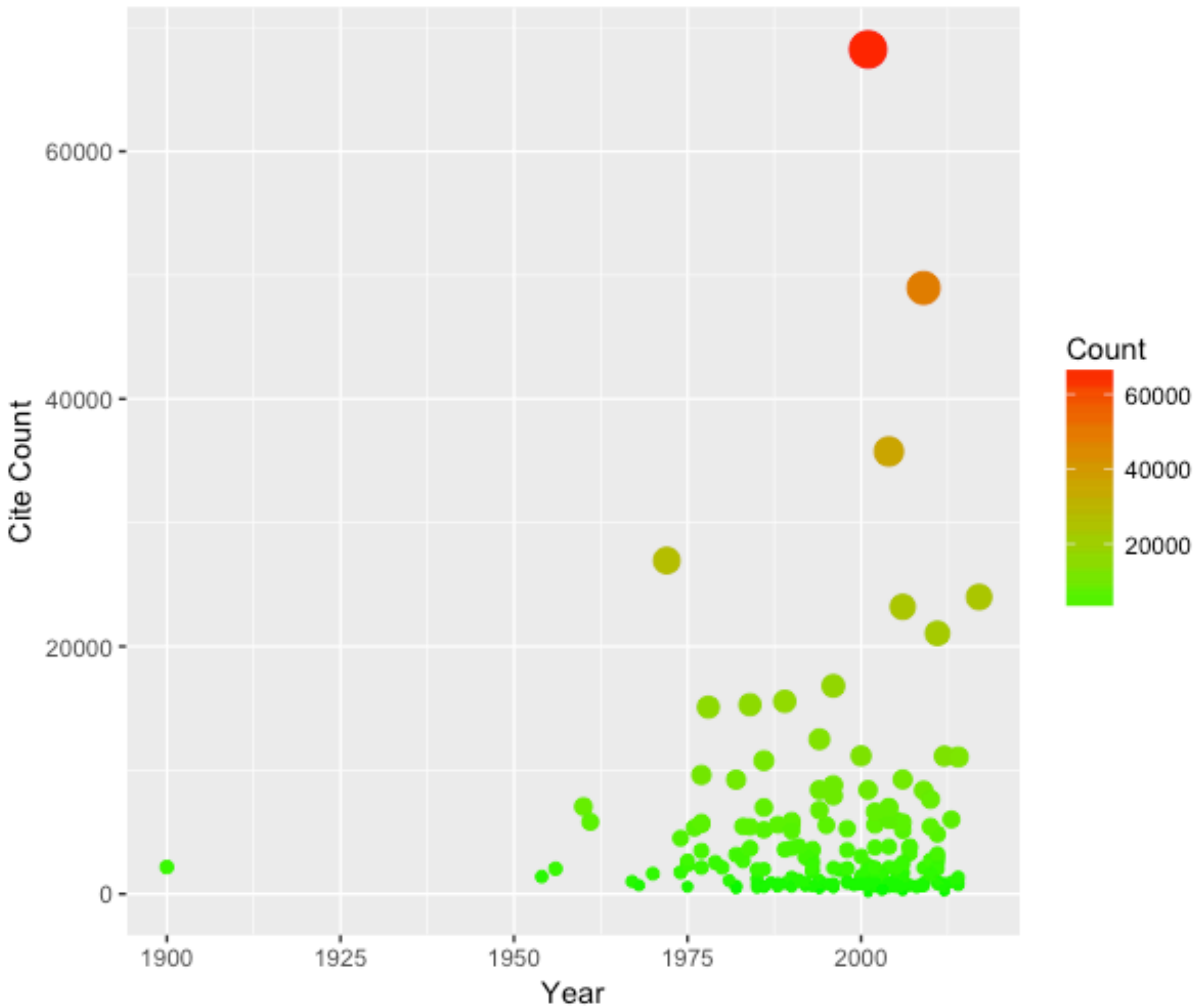


Notice that one paper that was published in 1900, I am curious to see what it is.

```
> dataset$title[which(dataset$year == 1900)]
[1] "The Structure of Science"
```

It is there because it contains the keyword one of the keyword "science".

Now we plot year against the total number of times papers has been cited. The cite count follows the general trend of the previous publish count plot.



I am curious to see what paper got the highest cite count.

```
> dataset$title[which(dataset$citecount == max(dataset$citecount))]
[1] "Analysis of relative gene expression data using real-time quantitative PCR and the 2- ΔΔCT method"
```

The paper was published by Ken Livak and Thomas Schmittgen in 2001, discussing the two most commonly used methods to analyze data from real-time.

Wordclouds:

I am curious to see what terminologies are popular in the searched papers. With the help of natural language processing packages in R, we are able to create beautiful wordclouds to explore.

```
# load dependencies
library(tm)
library(wordcloud)
library(RColorBrewer)

# use NLP to filter words
text <- paste(ml$title, collapse = ' ')
myCorpus <- Corpus(VectorSource(text))
myCorpus = tm_map(myCorpus, content_transformer(tolower))
myCorpus = tm_map(myCorpus, removePunctuation)
myCorpus = tm_map(myCorpus, removeNumbers)
myCorpus = tm_map(myCorpus, removeWords, c(stopwords("SMART"), "data",
"science"))
DTM = TermDocumentMatrix(myCorpus, control = list(minWordLength = 1))
matrix = as.matrix(DTM)
data_processed <- as.data.frame(matrix)
```

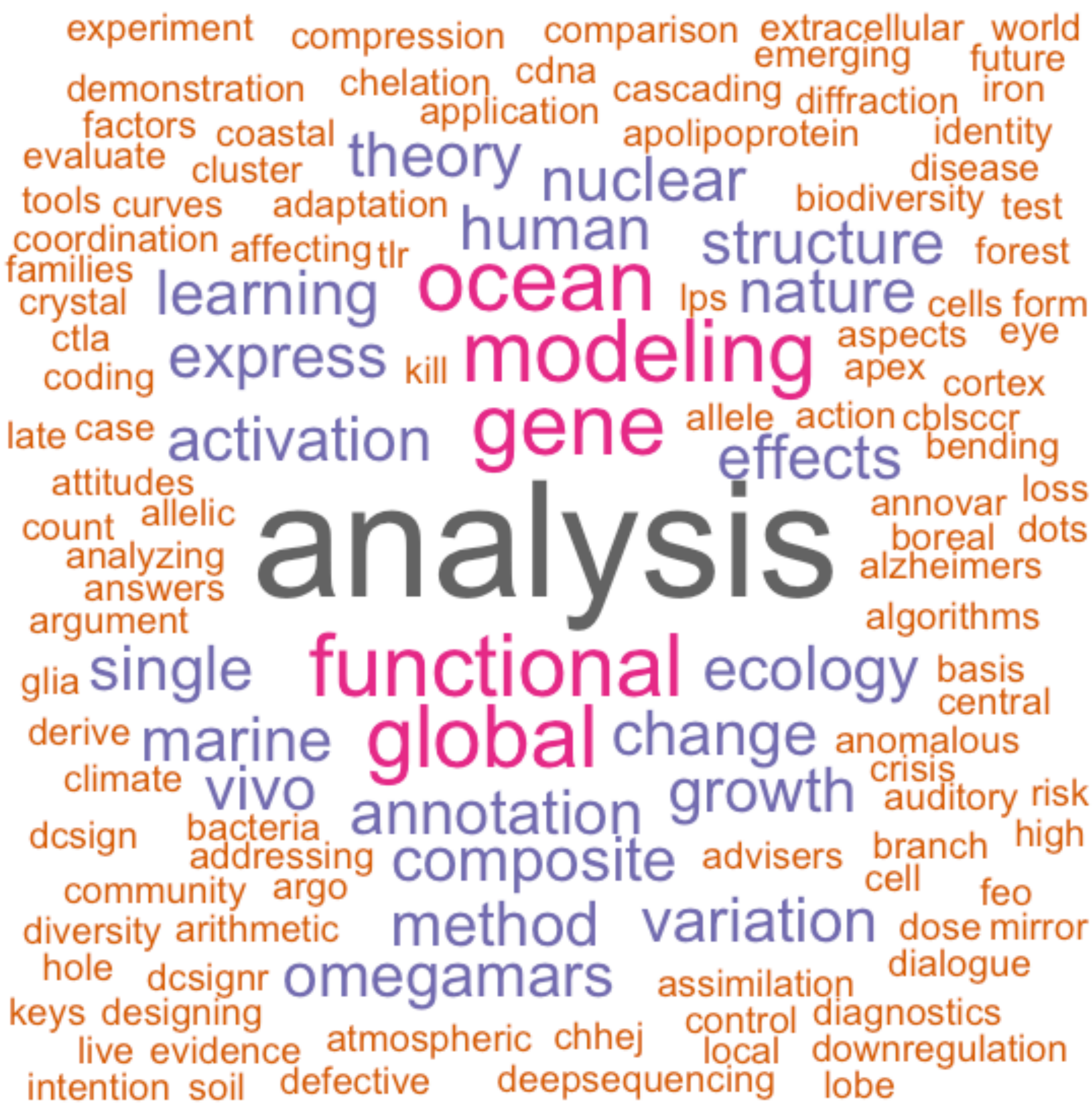
First, combine all the paper titles into one character string. Then, create the corpus object for the character string. After that, transform and clean the corpus, changing all letters to lower case, remove punctuations, remove numbers and remove stop words. We have to specify that "data" and "science" are removed from the corpus, because "data" and "science" will obviously be the most frequently existing words since they are the search keywords. Finally, transform the corpus into a frequency matrix, and then to a `data.frame` for better data structure. If the corpus was kept at matrix form, it would require a sorting before feeding to wordcloud.

```
# plot wordcloud with selected color palette
pal <- brewer.pal(8, "Dark2")
wordcloud(row.names(data_processed), data_processed$'1',
  scale = c(5, 0.4),
  min.freq = 2,
  rot.per = 0,
  max.words = 500,
  random.order = FALSE,
  colors = pal)
```

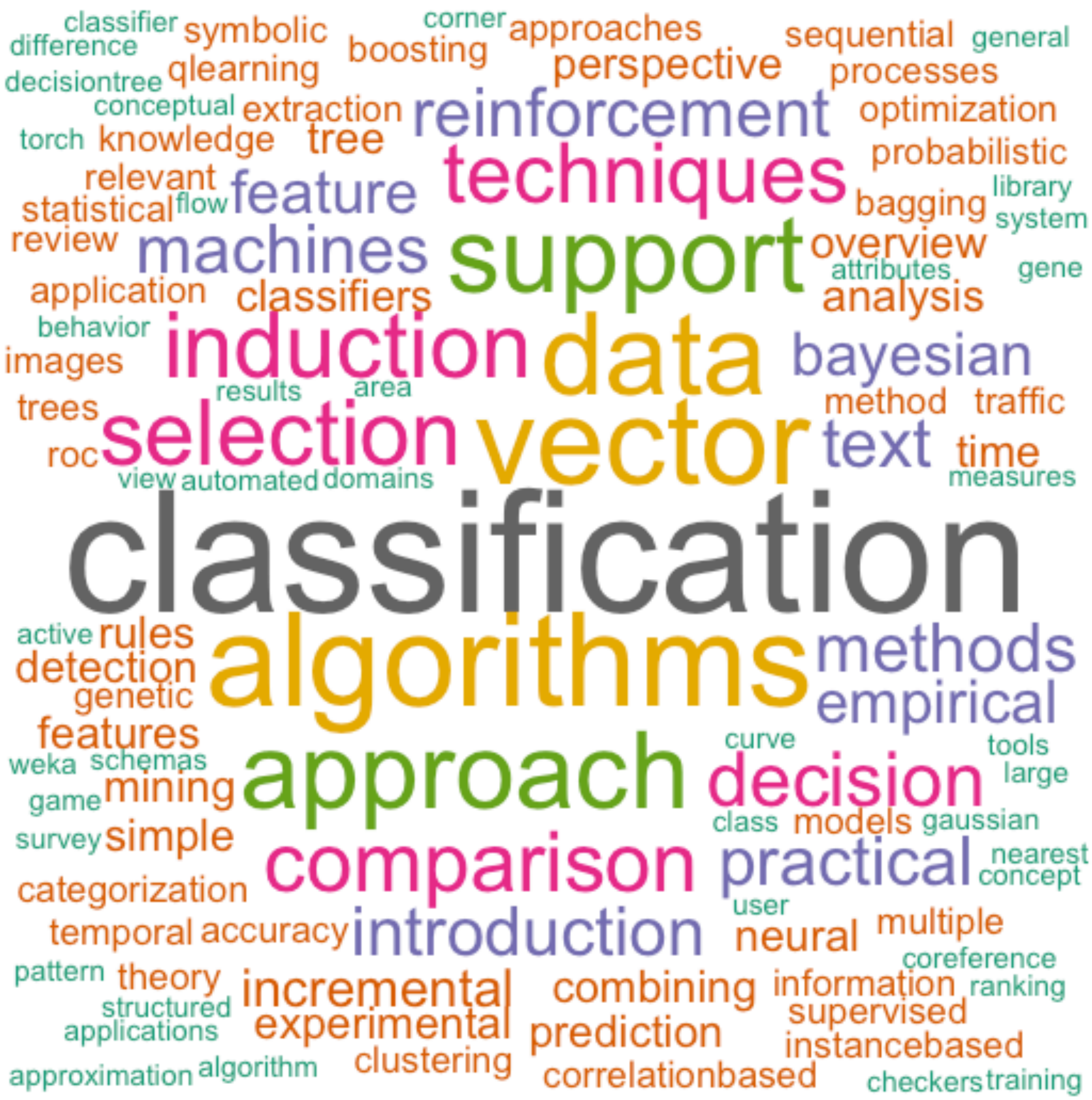
Then create the wordcloud using my preferred color palette.

We can see that "analysis" is the most frequent word in the titles. Other notable keywords are: methods, research, design, algorithms, management, network, etc.

Since the order of our original search result was based on relevence, I was curious to see what was the least relevant results look like. Using the same methods mentioned above, I scraped the last 10 pages of the search result. I was able to create the following wordcloud. Notice that "analysis" is still at the front and center. Notable keywords are: modeling, method, theory, etc.



I also tried my scraping and visualization further with keywords: "machine learning". I scraped the first 20 pages of the search results, and plotted the following wordcloud. The most mentioned keyword in the scrape is "classification". The core problem machine learning tries to solve is the classification problems since it is sometimes really hard to explicitly program the computer to recognize patterns. Other notable keywords are: algorithms, induction, selection, bayesian, decision, approach, etc.



Conclusion:

This is an interesting and rewarding project to do. I was able to experiment many of the web scraping techniques both mentioned in class and based on my research. And it is the first time that I was able to try out natural language processing and create wordclouds.

Google scholars is not a very hard website to scrape. Thanks to its simplistic design, it has easy to understand url format and page structure, which saved me hours to figure out how to scrape it. I initially planned to scrape reddit as mentioned in my project proposal, but it turned out that reddit has a really sophisticated nested webpage structure and comment system. However, I plan on finishing my reddit scrape on a later date.

At the time of writing this report, new ideas about scraping Google Scholars come to mind. I hope I will be able to update this project in the coming summer break.