Project 3 Data Science Skills

Umais Siddiqui,Neil Hwang, Michelle Mondy, Kalyanaraman Parthasarathy and Murali Kunissery

October 17, 2017

Rpubs link: http://rpubs.com/umais/Project3_Data607 |

`http://rpubs.com/neilhwang/group2`

`http://rpubs.com/mkunissery/321374`

`http://rpubs.com/mmondy/320996`

GitHub link: https://github.com/umais/Data607-Project3

Obejective

In this project our goal is to be able to answer the question

"Which are the most valued data science skills?"

Approach

In order to find the answer to our question we have researched a couple of job searching websites and decided to use CyberCoders and Indeed website to scrape the skillsets required for Jobs that had a title of Data Scientist.We will be using the rvest and MySQL libraries. rvest will be used to scrape and parse the HTML data.

```
library(devtools)
library(RMySQL)
```

```
## Loading required package: DBI
```

```
library(arules)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'arules'
```

```
## The following objects are masked from 'package:base':
##
##     abbreviate, write
```

```
library(arulesViz)
```

```
## Loading required package: grid
```

```
library(ggplot2)
library(plyr)
```

Database Schema

CREATE TABLE DataScienceJobs(

```
JobId int auto_increment primary key,

JobTitle nvarchar(255),

JobLocation nvarchar(255),

JobSalary nvarchar(255),

Source nvarchar(255)
```

);

CREATE TABLE DataScienceSkills(

```
SkillId int auto_increment primary key,

JobId int,

SkillName nvarchar(255)
```

);

Data Collection

We seperated out the data collection part in in a R Script file called Project3.R

In that file we are scraping the data from CyberCoders and Indeed website for Jobs that have a title of data scientist and inserting it in to the tables created based on the above schema. Below is the link to that code.

https://github.com/umais/Data607-Project3/blob/master/Project3/Project3.R

Retrieving Data from MySQL

We will be retrieving the data inserted in MYSQL from the scraping done from the R script file and performing some downstream analysis on the data.

```
mydb = dbConnect(MySQL(), user='root', password='Welcome@1', dbname='project3', host='localhost')
#mydb = dbConnect(MySQL(), user='root', password='password', host= '127.0.0.1', port=3306)
#dbSendQuery(mydb, "CREATE DATABASE project3;")
#dbSendQuery(mydb, "USE project3")
  results = dbSendQuery(mydb, "SELECT j.JobTitle,j.JobLocation,s.SkillName FROM DataScienceJobs j INNER

  jobSkills=fetch(results, n=-1)

  head(jobSkills)
```

```
##           JobTitle         JobLocation           SkillName
## 1  Data Scientist       Sunnyvale, CA              Python
## 2  Data Scientist       Sunnyvale, CA                C/C++
## 3  Data Scientist       Sunnyvale, CA        Apache Spark
```

```
## 4  Data Scientist       Sunnyvale, CA              Kafka
## 5  Data Scientist       Sunnyvale, CA      ElasticSearch
## 6  Data Scientist   San Francisco, CA  Postgres/Redshift
```

More Analysis

We can tell by the initial results from group by query that Python is the skill that is most valued.

```
rs = dbSendQuery(mydb, "SELECT SkillName,Count(1) as Total FROM DataScienceJobs j INNER JOIN DataScier

df=fetch(rs, n=-1)



head(df)
```

```
##            SkillName Total
## 1             Python   168
## 2  Machine Learning    122
## 3                  R     82
## 4      Data Science      56
## 5             Hadoop     55
## 6          Big Data      50
```

Looking at Indeed Data

If we look at only Indeed data again we can see that Python , R and Machine Learning are among the top required skills.

```
    rs = dbSendQuery(mydb, "SELECT SkillName,Source,Count(1) as Total FROM DataScienceJobs j INNER JO

      IndeedDF=fetch(rs, n=-1)
      head(IndeedDF)
```

```
##            SkillName Source Total
## 1             Python Indeed    99
## 2                  R  Indeed    82
## 3  Machine Learning  Indeed    61
## 4      Data Science  Indeed    56
## 5             Hadoop  Indeed    55
## 6          Big Data  Indeed    50
```

Looking at CyberCoders Data

Similarly in Cyber Coders data we see the same thing that Python,R,Machine Learning and Hadoop are among the top required skills for data scientist.

```
   rs = dbSendQuery(mydb, "SELECT SkillName,Source,Count(1) as Total FROM DataScienceJobs j INNER JOIN

      CyberCoderDF=fetch(rs, n=-1)
      head(CyberCoderDF)
```

```
##            SkillName      Source Total
## 1             Python CyberCoders    69
```

```
## 2  Machine Learning  CyberCoders    61
## 3                R  CyberCoders    30
## 4           Hadoop  CyberCoders    21
## 5       Data Mining  CyberCoders    21
## 6              SQL  CyberCoders    18
```

```
  rs = dbSendQuery(mydb, "SELECT j.JobId,JobTitle,JobLocation,JobSalary,SkillName,Source FROM DataScien

  AllJobs=fetch(rs, n=-1)

head(AllJobs)
```

```
##   JobId         JobTitle         JobLocation                JobSalary
## 1     1  Data Scientist       Sunnyvale, CA   Full-time $150k - $200k
## 2     1  Data Scientist       Sunnyvale, CA   Full-time $150k - $200k
## 3     1  Data Scientist       Sunnyvale, CA   Full-time $150k - $200k
## 4     1  Data Scientist       Sunnyvale, CA   Full-time $150k - $200k
## 5     1  Data Scientist       Sunnyvale, CA   Full-time $150k - $200k
## 6     2  Data Scientist   San Francisco, CA    Full-time $90k - $130k
##             SkillName      Source
## 1              Python CyberCoders
## 2              C/C++  CyberCoders
## 3       Apache Spark  CyberCoders
## 4              Kafka  CyberCoders
## 5      ElasticSearch  CyberCoders
## 6  Postgres/Redshift  CyberCoders
```

Association Analysis

To explore the data further, we perform an association analysis, which is one of the more popular unsupervised machine learning algorithms, using the package *arules*. To begin, we preoprocess the data to list the skills by specific jobs.

```
df <- data.frame(matrix(ncol = 2, nrow = nrow(AllJobs)))
df[,1] <- factor(AllJobs[,"JobId"])
df[,2] <- factor(AllJobs[,"SkillName"])
temp <- df[,c(1,2)]
first_item <- ddply(temp, .(X1), function(x) x[1, ])
temp2 <- merge(x = temp, y = first_item, by = "X1", all.x = TRUE)
data <- temp2[duplicated(temp2$X1),]
data$X1 <- data$X2.y
data$X2.y <- NULL
names(data) <- c("X", "Y")
m <- as.matrix(data)
l <- lapply(1:nrow(m), FUN = function(i) (m[i, ]))
```

Now, we convert the list into transactions that *arules* can work with in forming aprior rules to identify the most common associations that tend to occur together among skills.

```
transactions <- as(l, "transactions")
```

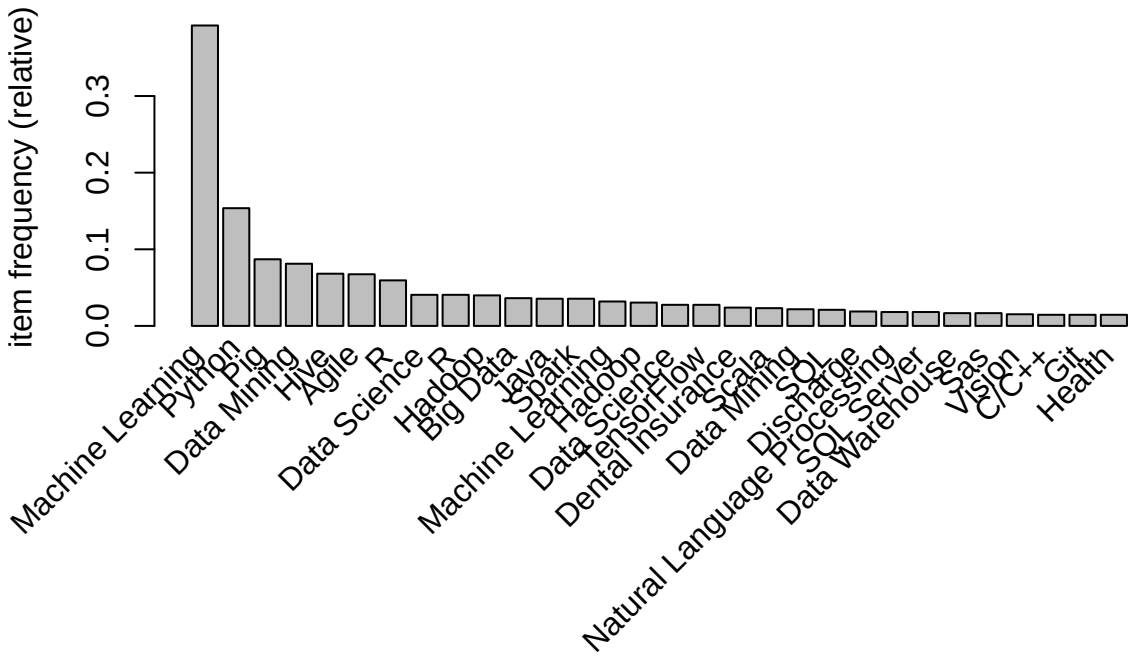
```
## Warning in asMethod(object): removing duplicated items in transactions
```

```
itemsets <- apriori(transactions, parameter = list(target = "frequent",
    supp=0.001, minlen = 2, maxlen=6))
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##          NA    0.1    1 none FALSE            TRUE       5   0.001      2
##  maxlen            target   ext
##       6 frequent itemsets FALSE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 1
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[238 item(s), 1380 transaction(s)] done [0.00s].
## sorting and recoding items ... [146 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 done [0.00s].
## writing ... [169 set(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

This frequency plot confirms our findings earlier that among the most demanded skill of data scientists are python, machine learning, R, data mining, and hadoop.

```
#Top 30 most frequently occurring skills
itemFrequencyPlot(transactions, topN=30)
```



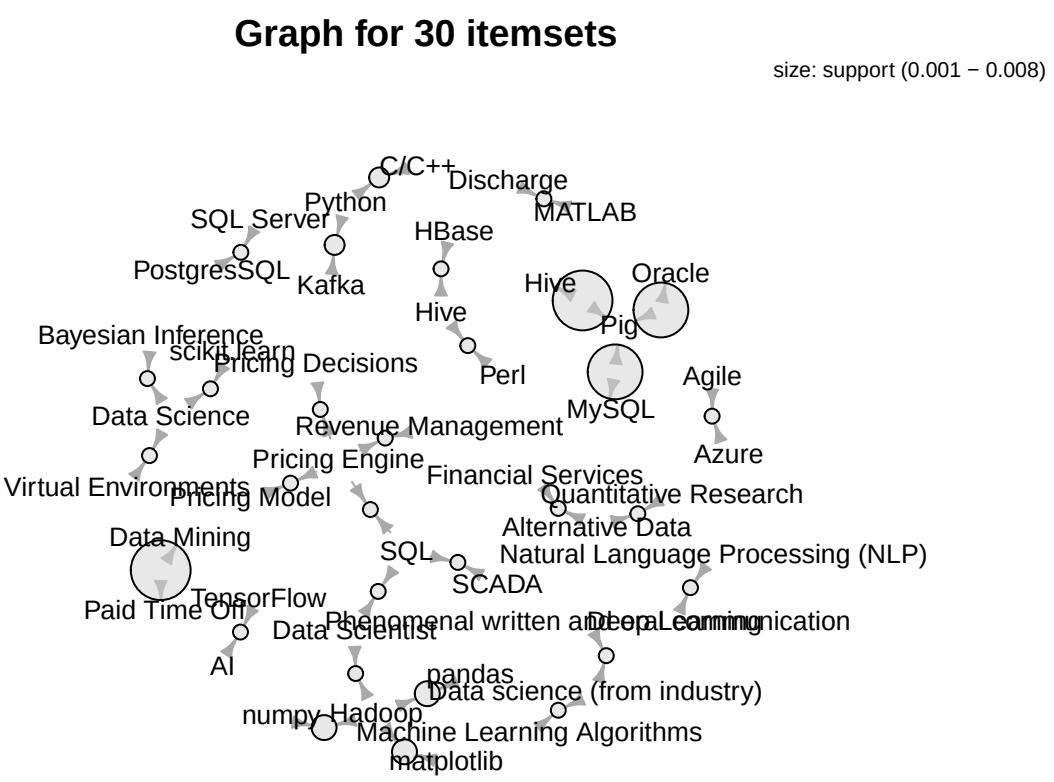
```r
quality(itemsets)$lift <- interestMeasure(itemsets, measure="lift", trans = transactions)

#Top 30 associations
inspect(head(sort(itemsets, by = "lift"), n=30))
```

```
##      items                                                 support       lift
## [1]  { Data science (from industry) ,
##        Phenomenal written and oral communication } 0.001449275 230.000000
## [2]  { Data science (from industry) ,
##        Machine Learning Algorithms }               0.001449275 230.000000
## [3]  { Pricing Decisions ,
##        Pricing Engine }                            0.001449275 172.500000
## [4]  { Pricing Engine ,
##        Pricing Model }                             0.001449275 172.500000
## [5]  { Pricing Engine ,
##        Revenue Management }                        0.001449275 172.500000
## [6]  { Alternative Data ,
##        Quantitative Research }                     0.001449275  55.200000
## [7]  { SCADA ,
##        SQL }                                       0.001449275  47.586207
## [8]  { Alternative Data ,
##        Financial Services }                        0.001449275  46.000000
## [9]  { Bayesian Inference ,
##        Data Science }                              0.001449275  36.315789
## [10] { Data Science ,
##        Virtual Environments }                      0.001449275  36.315789
## [11] { Data Science ,
##        scikit.learn }                              0.001449275  36.315789
## [12] { Deep Learning ,
##        Natural Language Processing (NLP) }         0.001449275  35.844156
## [13] { Hadoop ,
##        numpy }                                     0.002898551  32.857143
## [14] { Hadoop ,
##        matplotlib }                                0.002898551  32.857143
## [15] { Hadoop ,
##        pandas }                                    0.002898551  21.904762
## [16] {  Azure ,
##        Agile }                                     0.001449275  14.838710
## [17] {  HBase ,
##        Hive }                                      0.001449275  14.680851
## [18] {  PostgresSQL ,
##        SQL Server }                                0.001449275  13.800000
## [19] {  AI ,
##        TensorFlow }                                0.001449275  12.105263
## [20] { Pricing Engine ,
##        SQL }                                       0.001449275  11.896552
## [21] {  Oracle ,
##        Pig }                                       0.007246377  10.454545
## [22] {  Perl ,
##        Hive }                                      0.001449275   9.787234
## [23] { Data Scientist ,
##        SQL }                                       0.001449275   9.517241
## [24] {  MATLAB ,
##        Discharge }                                 0.001449275   8.846154
```

```
## [25] {  Hive ,
##        Pig }                                   0.007971014   8.433333
## [26] {  Paid Time Off ,
##        Data Mining }                           0.007971014   7.133459
## [27] {  MySQL ,
##        Pig }                                   0.007246377   6.764706
## [28] { Data Scientist ,
##        Hadoop }                                0.001449275   6.571429
## [29] { C/C++ ,
##       Python}                                  0.002173913   6.509434
## [30] { Kafka ,
##       Python}                                  0.002173913   6.509434
```

```
#Visualization of top associations and skills
plot(head(sort(itemsets, by = "lift"), n=30), method = "graph", control=list(cex=.8))
```



Conclusion

Based on the data collected from CyberCoders and Indeed we can see after doing some analysis that the most valuable skills for a data scientist are Python, R , Machine learning and Hadoop.