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/data607_project3

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

Obejective

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

"Which are the most valued data science skills?"

Approach

We have researched a couple of job searching websites and decided to use CyberCoders website to scrape the skillset 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(rvest)
```

```
## Loading required package: xml2
```

```
library(RMySQL)
```

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

Data Collection

In the step below we are opening the connection to MySQL database that resides locally but can also be a remote database server. We will use this connection to create the schema , Insert data that is collected from the scraping and also fetch the data for analysis.

```
mydb = dbConnect(MySQL(), user='root', password='Welcome@1', dbname='project3', host='localhost')
```

Insert Function

In this step we are defining a function that will accept HTML data that would be parsed and inserted in MySQL tables for the purposes of our analysis

```
InsertData<-function (url){
  selector_name<-".job-details-container"



skills<-html_nodes(x = url, css = selector_name)
for(i in 1:length(skills))
{
  selector_name<-".job-title a"
  JobTitle<-html_nodes(x = skills[i], css = selector_name)%>%
    html_text()
   selector_name<-".location"
  Location<-html_nodes(x = skills[i], css = selector_name)%>%
    html_text()
  selector_name<-".wage"
  salary<-html_nodes(x = skills[i], css = selector_name)%>%
    html_text()

   selector_name<-".skill-name"
  skill<-html_nodes(x = skills[i], css = selector_name)%>%
    html_text()



 dbGetQuery(mydb, paste("INSERT INTO DataScienceJobs(JobTitle,JobLocation,JobSalary) VALUES('" ,JobTitl

last_id = fetch(dbSendQuery(mydb, "SELECT LAST_INSERT_ID();"))

 for(j in 1 : length(skill))
 {


   dbGetQuery(mydb, paste("INSERT INTO  DataScienceSkills(JobId,SkillName) VALUES('"  ,last_id,"','",ski

 }


}

}
```

Creating the MySQL Schema

In this step we are dropping the tables and recreating them so that we can insert the data that is collected in next step. The purpose of dropping the tables is so that we can run the program multiple times and not create duplicate entries.

```
dbGetQuery(mydb, "DROP TABLE IF EXISTS DataScienceJobs;" )
```

```
## data frame with 0 columns and 0 rows
```

```
dbGetQuery(mydb, "DROP TABLE IF exists DataScienceSkills;" )
```

```
## data frame with 0 columns and 0 rows
```

```
dbGetQuery(mydb, "CREATE TABLE DataScienceJobs(
  JobId int auto_increment primary key,
  JobTitle nvarchar(255),
  JobLocation nvarchar(255),
  JobSalary nvarchar(255)

  );" )
```

```
## data frame with 0 columns and 0 rows
```

```
dbGetQuery(mydb, "  CREATE TABLE DataScienceSkills(
  SkillId int auto_increment primary key,
  JobId int,
  SkillName nvarchar(255)

  );" )
```

```
## data frame with 0 columns and 0 rows
```

Data Collection

In this step we will be collecting the data from CyberCoders and calling the function InsertData and passing the HTML as data to the function so that it could be parsed and inserted in to the MySQL tables

```
url<- read_html('https://www.cybercoders.com/jobs/data-scientist-jobs/')

 InsertData(url)

 url<- read_html('https://www.cybercoders.com/jobs/data-scientist-jobs/?page=2')
 InsertData(url)

  url<- read_html('https://www.cybercoders.com/jobs/data-scientist-jobs/?page=3')
 InsertData(url)

  url<- read_html('https://www.cybercoders.com/jobs/data-scientist-jobs/?page=4')
 InsertData(url)

  url<- read_html('https://www.cybercoders.com/jobs/data-scientist-jobs/?page=5')
 InsertData(url)

  url<- read_html('https://www.cybercoders.com/jobs/data-scientist-jobs/?page=6')
 InsertData(url)
#pager-item
  selector_name<-".pager-item span"
  pages<-html_nodes(x = url, css = selector_name)%>%
    html_text()
```

Retrieving Data from MySQL

We will be retrieving the data inserted in previous step and perfomring some downstream analysis on the data.

```
results = dbSendQuery(mydb, "SELECT j.JobTitle,j.JobLocation,s.SkillName FROM DataScienceJobs j INNER J

jobSkills=fetch(results, n=-1)

head(jobSkills)
```

```
##            JobTitle     JobLocation            SkillName
## 1  Data Scientist      Mclean, VA     Machine Learning
## 2  Data Scientist      Mclean, VA          Data Mining
## 3  Data Scientist      Mclean, VA   Data Visualization
## 4  Data Scientist      Mclean, VA               Python
## 5  Data Scientist      Mclean, VA    Raw Data Analysis
## 6  Data Scientist   Cambridge, MA                  SQL
```

More Analysis

We can tell by the initial look from the pie chart and the results from group by quety that Python is the skill that is most valued.

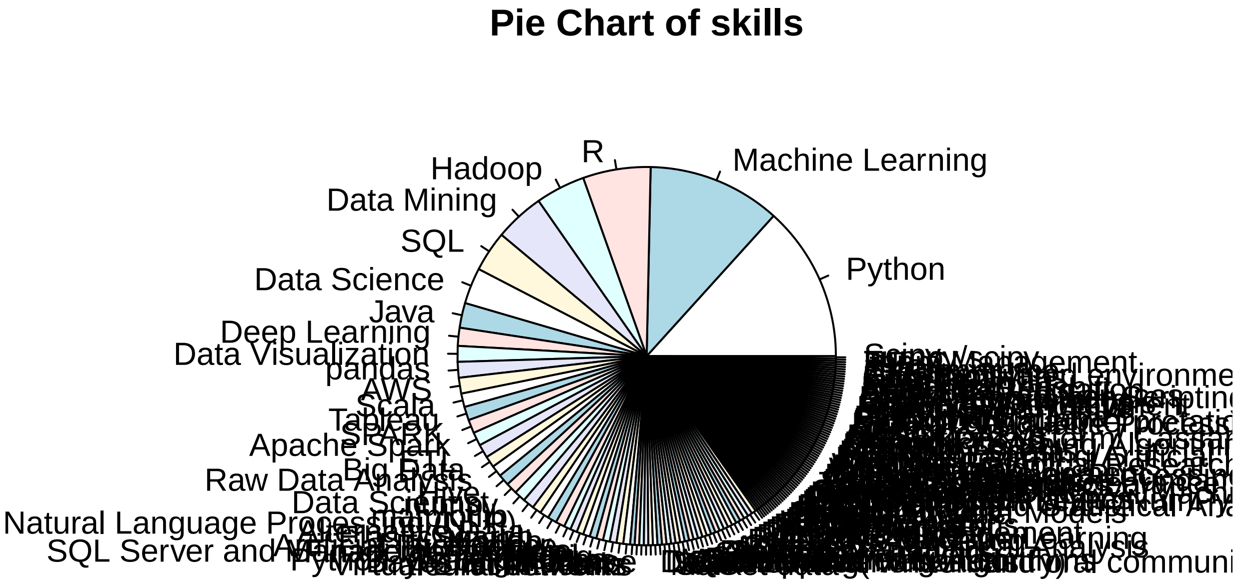
```
rs = dbSendQuery(mydb, "SELECT SkillName,Count(1) as Total FROM DataScienceJobs j INNER JOIN DataScienc

df=fetch(rs, n=-1)

head(df)
```

```
##            SkillName Total
## 1             Python    69
## 2   Machine Learning    59
## 3                  R    30
## 4             Hadoop    22
## 5        Data Mining    22
## 6                SQL    18
```

```
pie(df$Total, labels = df$SkillName, main="Pie Chart of skills")
```


Pie Chart of skills
Python
Machine Learning
R
Hadoop
Data Mining
SQL
Data Science
Java
Deep Learning
Data Visualization
pandas
AWS
Scala
Tableau
SPARK
Apache Spark
Big Data
Raw Data Analysis
Data Scientist
Natural Language Processing
SQL Server and