skills_analysis

GP SINGH

March 25, 2016

Installing the required library

```
suppressWarnings(library(data.table))
suppressWarnings(library(knitr))
suppressWarnings(library(tidyr))
suppressWarnings(require(plyr))

## Loading required package: plyr

suppressWarnings(library(wordcloud))

## Loading required package: RColorBrewer

suppressWarnings(library("RColorBrewer"))
suppressWarnings(library(plotrix))
suppressWarnings(library(plotly))

## Loading required package: ggplot2

##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:graphics':
##
##     layout

suppressWarnings(library(ggplot2))
suppressWarnings(library("devtools"))
```

The data is extracted from csv file that was generated using the articles and web urls.

```
#reading the data from csv file
data1 <-
read.csv("https://raw.githubusercontent.com/RobertSellers/SlackProjects/maste
r/data/Build-URL_DataFrame-Output.csv")

head(data1)

##                    doc_title       skill_name ds_freq
## 1  Big Data Analyst Profile         big data       9
## 2  Big Data Analyst Profile         Research       1
## 3  Big Data Analyst Profile    Story telling       0
## 4  Big Data Analyst Profile     Visual Basic       0
## 5  Big Data Analyst Profile   Technical Zeal       0
## 6  Big Data Analyst Profile Data Warehousing       0

#data1 contains doc_title, skill_name and frequency of occurence of skills in
that document
```

We will remove the column doc_title, as it is unnecessary for Analysis

```
skill <- data1[, 2:3]

head(skill)

##         skill_name ds_freq
## 1         big data       9
## 2         Research       1
## 3    Story telling       0
## 4     Visual Basic       0
## 5   Technical Zeal       0
## 6 Data Warehousing       0

#filtering out unique skills
sapply(skill, function(x) length(unique(x)))

## skill_name    ds_freq
##        149         39

#149 unique skills
# We want to remove ones with zero frequency.

skills <- subset(skill, ds_freq != 0)

# the dataset skills have all the skills with zero frequency removed.

sapply(skills, function(x) length(unique(x)))

## skill_name    ds_freq
##        115         38
```

Collecting the unique skills in all articles and adding up the frequency to create a data frame with unique skills and their count.

```
DT <- data.table(skills)
data_count <-DT[, sum(ds_freq), by = skill_name]
dat <- data.frame(data_count)


head(dat)

##      skill_name  V1
## 1      big data 704
## 2      Research 146
## 3    Statistics 359
## 4   Data Mining 166
## 5             R 323
## 6 communication  81
```

```
dim(dat)
```

```
## [1] 115   2
```

```
df<- dat[order(-dat$V1), ]


head(df)
```

```
##          skill_name  V1
## 1          big data 704
## 3        Statistics 359
## 5                 R 323
## 7  Machine Learning 297
## 17           Hadoop 272
## 9       programming 246
```

```
kable(df)
```

	skill_name	V1
1	big data	704
3	Statistics	359
5	R	323
7	Machine Learning	297
17	Hadoop	272
9	programming	246
19	Python	206
14	Visualization	178
4	Data Mining	166
2	Research	146
27	SQL	143
35	Java	86
6	communication	81
49	C++	77
45	C	75
51	SAS	67
74	Oracle	64

21	Spark	63
44	Business Intelligence	63
23	NoSQL	52
18	apache	51
39	Mathematics	49
61	predictive analytics	44
13	leadership	40
25	regression	39
12	Optimization	38
22	Hive	38
29	MapReduce	38
16	innovation	37
33	Excel	35
20	Pig	34
36	Probability	33
73	Tableau	31
15	unstructured data	28
42	creativity	25
43	Windows	25
26	business acumen	24
40	Curiosity	24
47	Linear Algebra	24
71	D3	22
46	Calculus	21
57	infographic	21
84	consulting	19
50	Matlab	18
52	Hortonworks	18
65	Cloudera	17
30	Hbase	16
38	artificial intelligence	16
53	Curious	16
63	MongoDB	16
41	innovative	15
48	MySQL	15
55	reporting	15

58	Collaboration	15
75	problem solving	15
10	Bayesian	14
32	database management	14
79	SPSS	14
80	Perl	14
95	Matrix	14
62	Apache Hadoop	13
91	Data Warehousing	13
24	pandas	12
28	apache spark	12
59	scripting	12
76	Teradata	12
82	HTML	12
8	Bayesian Statistics	10
37	neural networks	10
54	infographics	10
72	api	10
92	Linux	9
11	text mining	8
31	javascript	8
56	data security	8
69	Ruby	8
70	Unix	8
86	Mahout	8
90	GIS	8
64	Cassandra	7
77	Scala	6
99	Numpy	6
68	Maths	5
81	motivated	5
83	collaborative	5
94	Story telling	5
98	scipy	5
78	Stata	4
100	web scraping	4

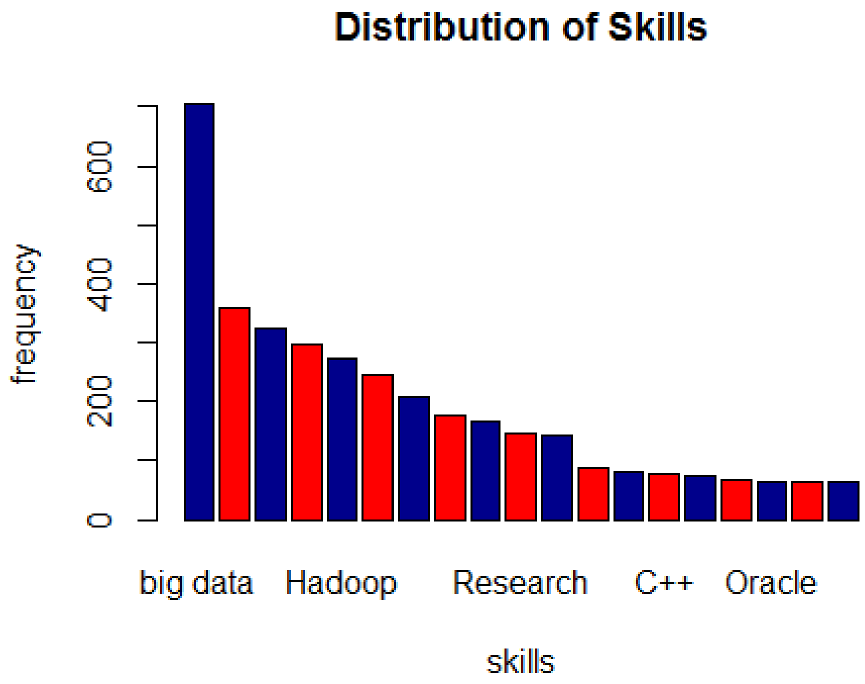
66	BigQuery	3
85	Matrices	3
93	neural network	3
105	cybersecurity	3
111	Story teller	3
34	VBA	2
87	Text Processing	2
88	Weka	2
89	Experimenting	2
104	Data Architecture	2
106	PostgreSQL	2
107	geographic information systems	2
108	motivation	2
60	Flowcharts	1
67	Homegrown	1
96	Geometry	1
97	ERwin	1
101	regular expressions	1
102	SQLite	1
103	Mac OS X	1
109	RDBMS	1
110	Algorithmic Thinking	1
112	Team work	1
113	Data Transformation	1
114	Data Integrity	1
115	machinelearning	1

Big Data, Statistics and R are the top three skills for Data Scientists.

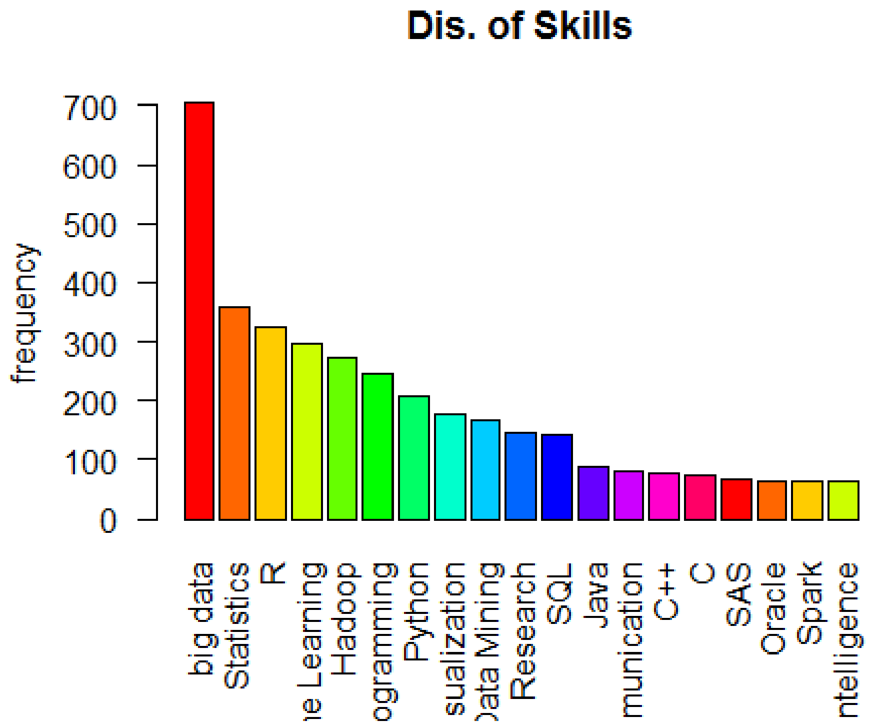
Visualizations

for visualizations we will create a dataframe with skills whose frequency of occurence is 60 or more for data scientists

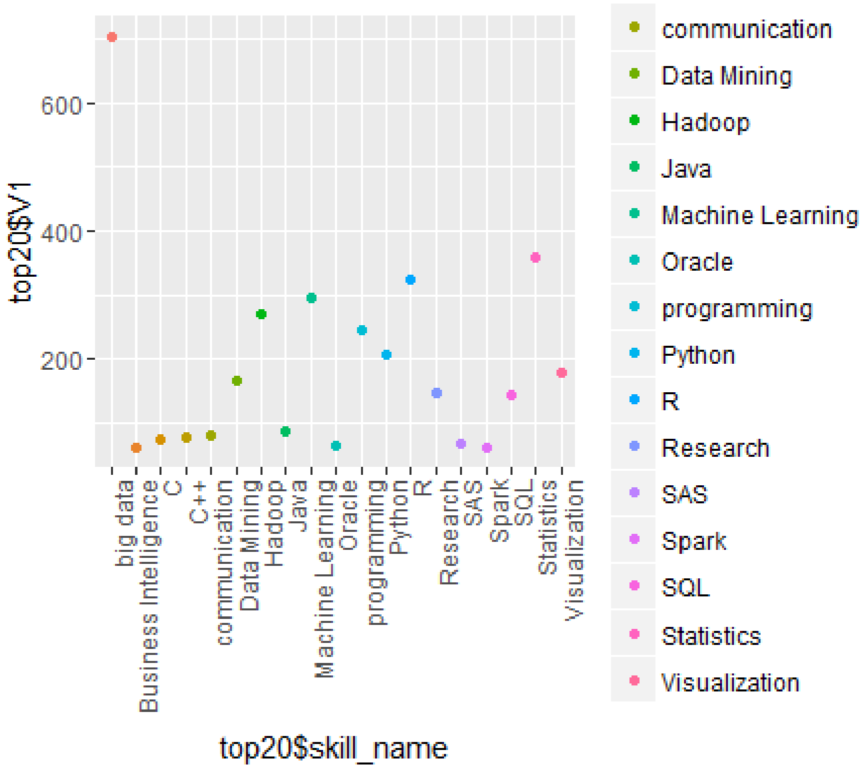
```r
top20 <- subset(df, V1 >= 60)
x <-barplot(top20$V1, main = "Distribution of Skills", xlab = "skills",
ylab="frequency", col=c("darkblue","red"), names.arg=top20$skill_name)
```



```
barplot(top20$V1, main="Dis. of Skills", ylab="frequency",
names.arg=top20$skill_name, las=2, col=rainbow(15))
```



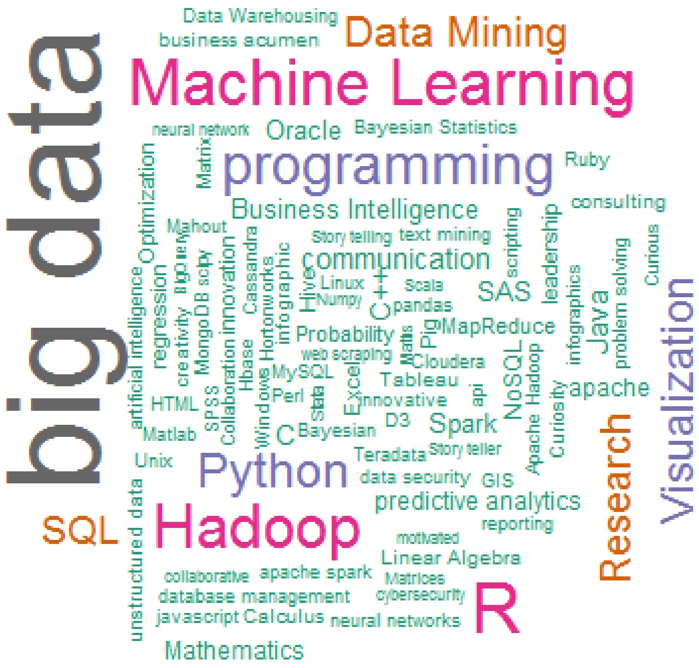
```r
p<-qplot(top20$skill_name, top20$V1, data = top20, color = top20$skill_name)
p + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Wordcloud

```r
set.seed(1234)
wordcloud(words = df$skill_name, freq = df$V1, rot.per=0.45,
          colors=brewer.pal(8, "Dark2"))
```

```
## Warning in wordcloud(words = df$skill_name, freq = df$V1, rot.per = 0.45,
:
## Statistics could not be fit on page. It will not be plotted.
```



Analysis of Soft and Technical Skills

```
data2 <-
read.csv("https://raw.githubusercontent.com/RobertSellers/SlackProjects/maste
r/data/skills_modified.csv")

head(data2)

##                       Skill sc_id
## 1             Adaptability     2
## 2     Algorithmic Thinking     2
## 3 Amazon Elastic MapReduce     1
## 4      Amazon Web Services     1
## 5                   apache     1
## 6            Apache Hadoop     1

dd <-dim(data2)
dd

## [1] 149   2

#We will create two datasets for technical and non technical skills

soft <- subset(data2, sc_id == 2)
soft

##                      Skill sc_id
## 1            Adaptability     2
## 2    Algorithmic Thinking     2
## 16        business acumen     2
## 17  Business Intelligence     2
## 24          Collaboration     2
## 25          collaborative     2
## 26          communication     2
## 27             consulting     2
## 28             creativity     2
## 29              Curiosity     2
## 30                Curious     2
## 38          Experimenting     2
## 54             innovation     2
## 55             innovative     2
## 58             leadership     2
## 74              motivated     2
## 75             motivation     2
## 85              Open Mind     2
## 94        problem solving     2
## 104             reporting     2
## 105              Research     2
## 120          Story teller     2
## 121         Story telling     2
## 123             Team work     2
## 130         Visualization     2
## 137        Technical Zeal     2
```

```
tech <- subset(data2, sc_id == 1)

#checking dimensions of  datasets

ds<- dim(soft)
ds
```

```
## [1] 26  2
```

```
dt<- dim(tech)
dt
```

```
## [1] 123   2
```

```
# percentage of occurence

soft_per <- as.numeric((26/149)*100)
soft_per
```

```
## [1] 17.44966
```

```
tech_per<- as.numeric((123/149)*100)
tech_per
```

```
## [1] 82.55034
```