In [124…
```
import numpy as np
import pandas as pd

import nltk
from nltk.corpus import stopwords
import string
from wordcloud import WordCloud

import seaborn as sns

import matplotlib.pyplot as plt
%matplotlib inline
```

1. Load the data into python

In [125…
```
#reading the data

df = pd.read_excel(r'Resume_Data1.xlsx')
df['Cleaned_Resume'] = ''
df.head()
```

Out[125]:

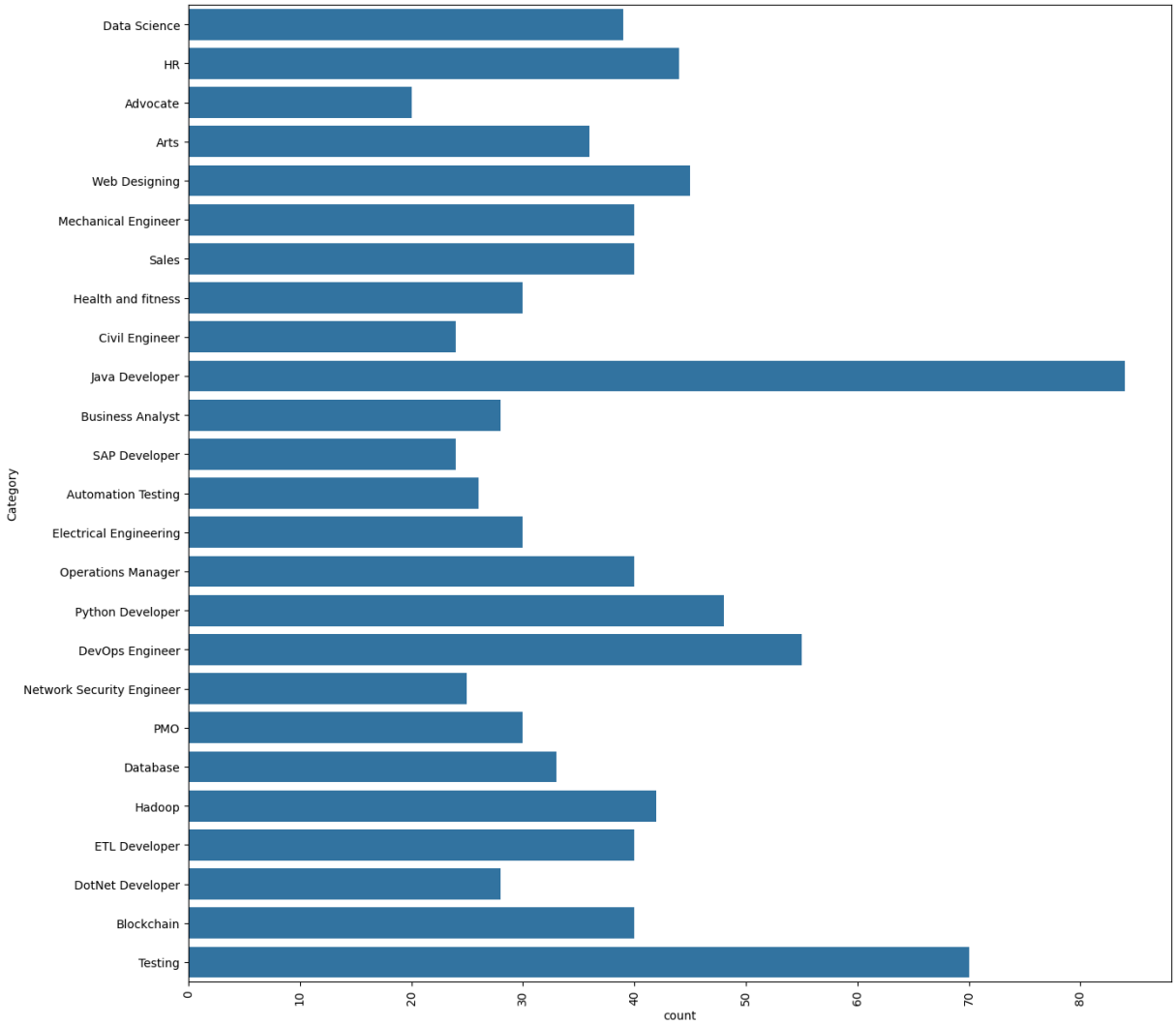
	Category	Resume	Cleaned_Resume
0	NaN	Skills * Programming Languages: Python (pandas...	
1	Data Science	Education Details \nMay 2013 to May 2017 B.E ...	
2	Data Science	Areas of Interest Deep Learning, Control Syste...	
3	Data Science	Skills â?¢ R â?¢ Python â?¢ SAP HANA â?¢ Table...	
4	Data Science	Education Details \n MCA YMCAUST, Faridabad...	

In [126…
```
print ("Resume Categories")
print (df['Category'].value_counts())
```

```
Resume Categories
Java Developer                84
Testing                       70
DevOps Engineer               55
Python Developer              48
Web Designing                 45
HR                            44
Hadoop                        42
Blockchain                    40
ETL Developer                 40
Mechanical Engineer           40
Sales                         40
Operations Manager            40
Data Science                  39
Arts                          36
Database                      33
Electrical Engineering        30
PMO                           30
Health and fitness            30
Business Analyst              28
DotNet Developer              28
Automation Testing            26
Network Security Engineer     25
SAP Developer                 24
Civil Engineer                24
Advocate                      20
Name: Category, dtype: int64
```

In [127…

```
plt.figure(figsize=(15,15))
plt.xticks(rotation=90)
sns.countplot(y="Category", data=df)
plt.show()
```



In [128…

```
df["Resume"][2]
```

Out[128]: 'Areas of Interest Deep Learning, Control System Design, Programming in-Python, Electric Machinery, Web Development, Analytics Technical Activities q Hindustan Aeronautics Limited, Bangalore - For 4 weeks under the guidance of Mr. Satish, Senior Engineer in the hangar of Mirage 2000 fighter aircraft Technical Skills Programming Matlab, Python and Java, LabView, Python WebFrameWork-Django, Flask, LTSPICE-intermediate Languages and and MIPOWER-intermediate, Github (GitBash), Jupyter Notebook, Xampp, MySQL-Basics, Python Software Packages Interpreters-Anaconda, Python2, Python3, Pycharm, Java IDE-Eclipse Operating Systems Windows, Ubuntu, Debian-Kali Linux Education Details \nJanuary 2019 B.Tech. Electrical and Electronics Engineering Manipal Institute of Technology\nJanuary 2015 DEEKSHA CENTER\nJanuary 2013 Little Flower Public School\nAugust 2000 Manipal Academy of Higher\nDATA SCIENCE \n\nDATA SCIENCE AND ELECTRICAL ENTHUSIAST\nSkill Details \nData Analysis- Exprience - Less than 1 year months\nexcel- Exprience - Less than 1 year months\nMachine Learning- Exprience - Less than 1 year months\nmathematics- Exprience - Less than 1 year months\nPython- Exprience - Less than 1 year months\nMatlab- Exprience - Less than 1 year months\nElectrical Engineering- Exprience - Less than 1 year months\nSql- Exprience - Less than 1 year monthsCompany Details \ncompany - THEMATHCOMPANY\ndescription - I am currently working with a Casino based operator(name not to be disclosed) in Macau.I need to segment the customers who visit their property based on the value the patrons bring into the company.Basically prove that the segmentation can be done in much better way than the current system which they have with proper numbers to back it up.Henceforth they can implement target marketing strategy to attract their customers who add value to the business.'

In [129…
```
df["Resume"][500]
```

Out[129]: 'Education Details \nJanuary 2012 to January 2013 B.E. Electrical Shivaji University\nSeptember 2008 HSC Pune, Maharashtra Pune University\nJuly 2006 SSC Pune, Maharashtra Pune University\nElectrical Engineer \n\nElectrical Engineer - R K ELECTRICAL PVT. LTD\nSkill Details \nCompany Details \ncompany - R K ELECTRICAL PVT. LTD\ndescription - Experience:- 1 Year 3 Months\n\nTroubleshooting and Maintenance of following Electrical Equipment:-\nâ?¢ All Type of Maintenance of Utility.\nâ?¢ Electrical and Mechanical Maintenance.\nâ?¢ Two 625 KVA Diesel Generator Set (Kirloskar)\nâ?¢ HT/LT Switchgear With Protection System Using Relays and Provision For Interlocking (C&S, Kirloskar)\nâ?¢ Handling HT Vacuum & SF6 Circuit Breaker, Transformer Up to 5000 KVA, LT Air circuit Breaker 2000A\nâ?¢ Maintenance of STP and WTP Plant.\nâ?¢ Maintenance of Air Blower, Actuators, Soft Starter, EOT Crane, Mono Rail, Centrifugal or Vertical Pumps, Hydraulic Machine, Rolling Machine, Lath Machine, Drill Machine, AHU, HVAC, Chiller etc.\nâ?¢ Basic knowledge of PLC/SCADA Operation.\nâ?¢ Trouble shooting of Switchgear and Control Panel, Pump and Motor\nâ?¢ Maintenance of UPS, Battery Charger and Battery Bank\nâ?¢ Motor Testing Both HT & LT Up to 450 KW\nâ?¢ Monitoring and Controlling the 110V Control Panel and Relays Panel\nâ?¢ Involved in Fault Finding & Relay Resetting\nâ?¢ Monitoring and Correcting Power Factor \nâ?¢ Service and Maintenance of Up to 55 KW Submersible Pump\nâ?¢ Maintenance of MCC and PCC Panel\nâ?¢ Servicing of Motor and Associated Component and Motor Operated Valve\nâ?¢ Problem Solving of Power Contactor, Auxiliary Contactor Relay, CT and PT\nâ?¢ Effecting Preventive/Predictive Maintenance Schedules Equipment in Order to Increase the Uptime/ Reliability\nâ?¢ Maintenance & Operation in Day to Day Activity\nâ?¢ Operation, Preventive Maintenance, Day to Day Breakdown Maintenance Conventional Maintaining of Log Book and Check List.\nâ?¢ 33/22kV Main Feeder & 22/11kV Distribution Line Maint. & H.T/L.T S/S Break Down Work.\n\nELECTRICAL SAFETY (Knowledge of Various Aspect of Safety & Its Application)\nâ?¢ Requirement, Familiar With Various Safety Equipment and Tools\nâ?¢ Lockout, Tag out of Electrical Switchgear During Work\nâ?¢ Issue of Work Permit Line Clearance to Work on Electrical Distribution Network\nâ?¢ Requirement & Proper Usage of Protective Equipment\nâ?¢ Accident Statistics'

2. Cleaning and preprocessing the resume text

In [130…
```
import re
def cleanResume(resumeText):
    resumeText = re.sub('http\S+\s*', ' ', resumeText)  # remove URLs
    resumeText = re.sub('RT|cc', ' ', resumeText)  # remove RT and cc
    resumeText = re.sub('#\S+', '', resumeText)  # remove hashtags
    resumeText = re.sub('@\S+', '  ', resumeText)  # remove mentions
    resumeText = re.sub('[%s]' % re.escape("""!"#$%&'()*+,-./:;<=>?@[\]^_`{|
    resumeText = re.sub(r'[^\x00-\x7f]',r' ', resumeText)
    resumeText = re.sub('\s+', ' ', resumeText)  # remove extra whitespace
    return resumeText

df['Cleaned_Resume'] = df.Resume.apply(lambda x: cleanResume(x))
```

In [131…
```
df['Cleaned_Resume'][2]
```

Out[131]:
```
'Areas of Interest Deep Learning Control System Design Programming in Pyth
on Electric Machinery Web Development Analytics Technical Activities q Hin
dustan Aeronautics Limited Bangalore For 4 weeks under the guidance of Mr
Satish Senior Engineer in the hangar of Mirage 2000 fighter aircraft Techn
ical Skills Programming Matlab Python and Java LabView Python WebFrameWork
Django Flask LTSPICE intermediate Languages and and MIPOWER intermediate G
ithub GitBash Jupyter Notebook Xampp MySQL Basics Python Software Packages
Interpreters Anaconda Python2 Python3 Pycharm Java IDE Eclipse Operating S
ystems Windows Ubuntu Debian Kali Linux Education Details January 2019 B T
ech Electrical and Electronics Engineering Manipal Institute of Technology
January 2015 DEEKSHA CENTER January 2013 Little Flower Public School Augus
t 2000 Manipal Academy of Higher DATA SCIENCE DATA SCIENCE AND ELECTRICAL
ENTHUSIAST Skill Details Data Analysis Exprience Less than 1 year months e
xcel Exprience Less than 1 year months Machine Learning Exprience Less tha
n 1 year months mathematics Exprience Less than 1 year months Python Expri
ence Less than 1 year months Matlab Exprience Less than 1 year months Elec
trical Engineering Exprience Less than 1 year months Sql Exprience Less th
an 1 year monthsCompany Details company THEMATHCOMPANY description I am cu
rrently working with a Casino based operator name not to be disclosed in M
acau I need to segment the customers who visit their property based on the
value the patrons bring into the company Basically prove that the segmenta
tion can be done in much better way than the current system which they hav
e with proper numbers to back it up Henceforth they can implement target m
arketing strategy to attract their customers who add value to the business
'
```

In [132…
```
df.head()
```

Out[132]:

	Category	Resume	Cleaned_Resume
0	NaN	Skills * Programming Languages: Python (pandas...	Skills Programming Languages Python pandas num...
1	Data Science	Education Details \nMay 2013 to May 2017 B.E ...	Education Details May 2013 to May 2017 B E UIT...
2	Data Science	Areas of Interest Deep Learning, Control Syste...	Areas of Interest Deep Learning Control System...
3	Data Science	Skills â?¢ R â?¢ Python â?¢ SAP HANA â?¢ Table...	Skills R Python SAP HANA Tableau SAP HANA SQL ...
4	Data Science	Education Details \n MCA YMCAUST, Faridabad...	Education Details MCA YMCAUST Faridabad Haryan...

In [133…
```
len(df)
```

Out[133]:
```
962
```

3. Performing the NLP tasks on the cleaned text

In [134…
```
#getting the entire resume text
```

```
corpus=" "

for i in range(0,962):
    corpus= corpus+ df["Cleaned_Resume"][i]
```

In [135…
```
corpus[1000:2500]
```

Out[135]:
```
'review process and run analytics and generate reports Core member of a te
am helped in developing automated review platform tool from scratch for as
sisting E discovery domain this tool implements predictive coding and topi
c modelling by automating reviews resulting in reduced labor costs and tim
e spent during the lawyers review Understand the end to end flow of the so
lution doing research and development for classification models predictive
analysis and mining of the information present in text data Worked on anal
yzing the outputs and precision monitoring for the entire tool TAR assists
in predictive coding topic modelling from the evidence by following EY sta
ndards Developed the classifier models in order to identify red flags and
fraud related issues Tools Technologies Python scikit learn tfidf word2vec
doc2vec cosine similarity Na ve Bayes LDA NMF for topic modelling Vader an
d text blob for sentiment analysis Matplot lib Tableau dashboard for repor
ting MULTIPLE DATA SCIENCE AND ANALYTIC PROJECTS USA CLIENTS TEXT ANALYTIC
S MOTOR VEHICLE CUSTOMER REVIEW DATA Received customer feedback survey dat
a for past one year Performed sentiment Positive Negative Neutral and time
series analysis on customer comments across all 4 categories Created heat
map of terms by survey category based on frequency of words Extracted Posi
tive and Negative words across all the Survey categories and plotted Word
cloud Created customized tableau dashboards for effective reporting and vi
sualizations CHAT'
```

In [136…
```
#Creating the tokenizer
tokenizer = nltk.tokenize.RegexpTokenizer('\w+')

#Tokenizing the text
tokens = tokenizer.tokenize(corpus)

len(tokens)
```

Out[136]:
```
411913
```

In [137…
```
#now we shall make everything lowercase for uniformity
#to hold the new lower case words

words = []

# Looping through the tokens and make them lower case
for word in tokens:
    words.append(word.lower())
```

In [138…
```
#import nltk
#nltk.download('stopwords')
```

In [139…
```
#Stop words are generally the most common words in a language.
#English stop words from nltk.

ltk.corpus.stopwords.words('english')
```

```
words_new = []

#Now we need to remove the stop words from the words variable
#Appending to words_new all words that are in words but not in sw

for word in words:
    if word not in stopwords:
        words_new.append(word)
```

In [140…
```
len(words_new)
```

Out[140]: 318305

Lemmatization

Lemmatization is the process of grouping together the different inflected forms of a word so they can be analysed as a single item. Lemmatization is similar to stemming but it brings context to the words. So it links words with similar meaning to one word. Lemmatization is preferred over Stemming because lemmatization does morphological analysis of the words.

In [141…
```
#import nltk
#nltk.download('wordnet')
```

In [142…
```
from nltk.stem import WordNetLemmatizer

wn = WordNetLemmatizer()

lem_words=[]

for word in words_new:
    word=wn.lemmatize(word)
    lem_words.append(word)
```

In [143…
```
len(lem_words)
```

Out[143]: 318305

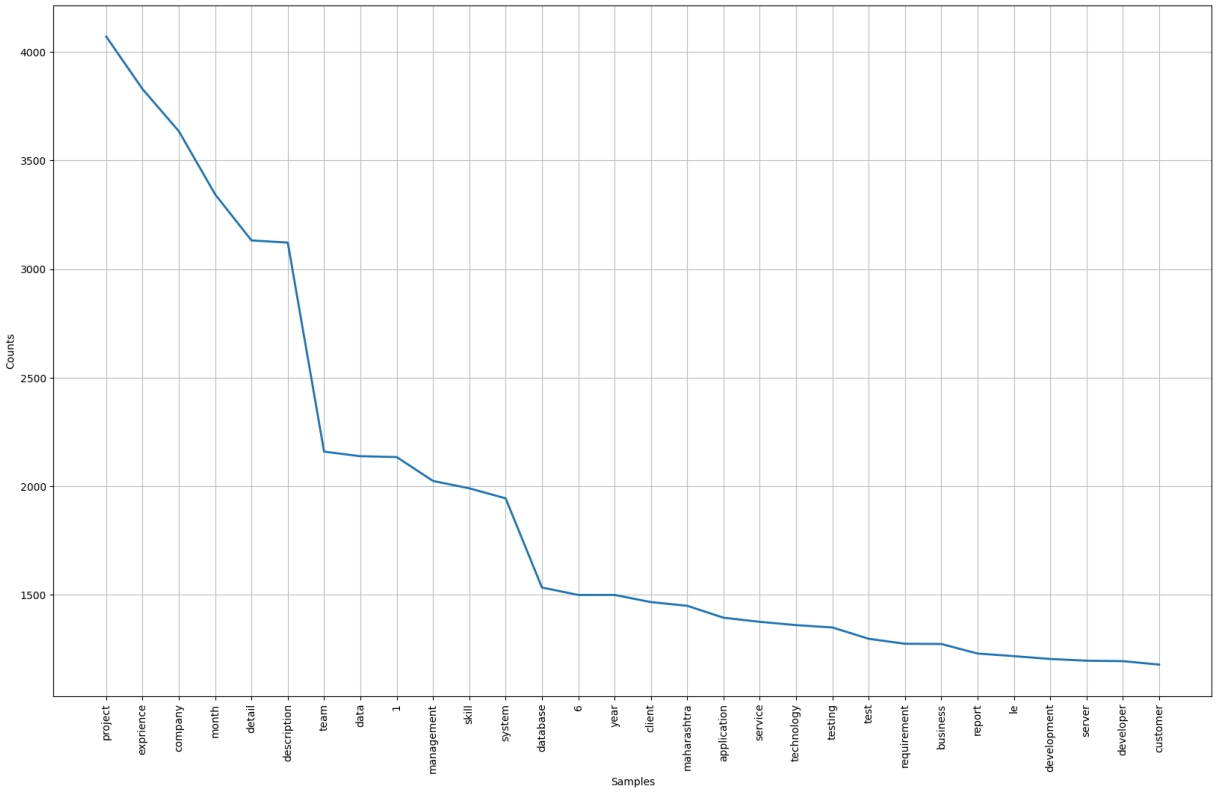
In [144…
```
same=0
diff=0

for i in range(0,1832):
    if(lem_words[i]==words_new[i]):
        same=same+1
    elif(lem_words[i]!=words_new[i]):
        diff=diff+1

print('Number of words Lemmatized=', diff)
print('Number of words not Lemmatized=', same)
```

```
Number of words Lemmatized= 294
Number of words not Lemmatized= 1538
```

4. Find the frequency distribution of the words

In [145…
```
#The frequency distribution of the words
freq_dist = nltk.FreqDist(lem_words)
```

In [146…
```
#Frequency Distribution Plot
plt.subplots(figsize=(20,12))
freq_dist.plot(30)
```



Out[146]: <Axes: xlabel='Samples', ylabel='Counts'>

In [147…
```
len(freq_dist)
```

Out[147]: 6769

In [148…
```
mostcommon = freq_dist.most_common(50)
```

In [149…
```
mostcommon
```

```
Out[149]: [('project', 4071),
           ('exprience', 3829),
           ('company', 3635),
           ('month', 3344),
           ('detail', 3132),
           ('description', 3122),
           ('team', 2159),
           ('data', 2138),
           ('1', 2134),
           ('management', 2024),
           ('skill', 1990),
           ('system', 1944),
           ('database', 1533),
           ('6', 1499),
           ('year', 1499),
           ('client', 1466),
           ('maharashtra', 1449),
           ('application', 1394),
           ('service', 1375),
           ('technology', 1360),
           ('testing', 1349),
           ('test', 1297),
           ('requirement', 1274),
           ('business', 1273),
           ('report', 1229),
           ('le', 1217),
           ('development', 1204),
           ('server', 1196),
           ('developer', 1194),
           ('customer', 1178),
           ('ltd', 1177),
           ('process', 1163),
           ('responsibility', 1137),
           ('using', 1124),
           ('sql', 1120),
           ('january', 1090),
           ('java', 1076),
           ('engineering', 1055),
           ('work', 1038),
           ('pune', 1026),
           ('role', 969),
           ('c', 951),
           ('user', 916),
           ('operation', 895),
           ('software', 886),
           ('pvt', 879),
           ('sale', 845),
           ('activity', 832),
           ('environment', 800),
           ('design', 786)]
```

5. Building the word cloud with the corpus

In [150…
```
#converting into string

res=' '.join([i for i in lem_words if not i.isdigit()])
```

In [151…
```
plt.subplots(figsize=(16,10))
wordcloud = WordCloud(
                          background_color='black',
                          max_words=100,
                          width=1400,
                          height=1200
                         ).generate(res)


plt.imshow(wordcloud)
plt.title('Resume Text WordCloud (100 Words)')
plt.axis('off')
plt.show()
```

Resume Text WordCloud (100 Words)

In [152…
```
plt.subplots(figsize=(20,15))
wordcloud = WordCloud(
                          background_color='black',
                          max_words=200,
```

```
                                width=1400,
                                height=1200
                               ).generate(res)


plt.imshow(wordcloud)
plt.title('Resume Text WordCloud (200 Words)')
plt.axis('off')
plt.show()
```

6. Filter the resume data for a specific category of Data Science

In [153…
```
data_science= df[df["Category"]=="Data Science"]
```

In [154…
```
data_science.head()
```

Out[154]:

	Category	Resume	Cleaned_Resume
1	Data Science	Education Details \nMay 2013 to May 2017 B.E ...	Education Details May 2013 to May 2017 B E UIT...
2	Data Science	Areas of Interest Deep Learning, Control Syste...	Areas of Interest Deep Learning Control System...
3	Data Science	Skills â?¢ R â?¢ Python â?¢ SAP HANA â?¢ Table...	Skills R Python SAP HANA Tableau SAP HANA SQL ...
4	Data Science	Education Details \n MCA YMCAUST, Faridabad...	Education Details MCA YMCAUST Faridabad Haryan...
5	Data Science	SKILLS C Basics, IOT, Python, MATLAB, Data Sci...	SKILLS C Basics IOT Python MATLAB Data Science...

In [155…
```
len(data_science)
```

Out[155]: 39

In [156…
```
data_science["Cleaned_Resume"]
```

Out[156]:
```
1     Education Details May 2013 to May 2017 B E UIT...
2     Areas of Interest Deep Learning Control System...
3     Skills R Python SAP HANA Tableau SAP HANA SQL ...
4     Education Details MCA YMCAUST Faridabad Haryan...
5     SKILLS C Basics IOT Python MATLAB Data Science...
6     Skills Python Tableau Data Visualization R Stu...
7     Education Details B Tech Rayat and Bahra Insti...
8     Personal Skills Ability to quickly grasp techn...
9     Expertise Data and Quantitative Analysis Decis...
10    Skills Programming Languages Python pandas num...
11    Education Details May 2013 to May 2017 B E UIT...
12    Areas of Interest Deep Learning Control System...
13    Skills R Python SAP HANA Tableau SAP HANA SQL ...
14    Education Details MCA YMCAUST Faridabad Haryan...
15    SKILLS C Basics IOT Python MATLAB Data Science...
16    Skills Python Tableau Data Visualization R Stu...
17    Education Details B Tech Rayat and Bahra Insti...
18    Personal Skills Ability to quickly grasp techn...
19    Expertise Data and Quantitative Analysis Decis...
20    Skills Programming Languages Python pandas num...
21    Education Details May 2013 to May 2017 B E UIT...
22    Areas of Interest Deep Learning Control System...
23    Skills R Python SAP HANA Tableau SAP HANA SQL ...
24    Education Details MCA YMCAUST Faridabad Haryan...
25    SKILLS C Basics IOT Python MATLAB Data Science...
26    Skills Python Tableau Data Visualization R Stu...
27    Education Details B Tech Rayat and Bahra Insti...
28    Personal Skills Ability to quickly grasp techn...
29    Expertise Data and Quantitative Analysis Decis...
30    Skills Programming Languages Python pandas num...
31    Education Details May 2013 to May 2017 B E UIT...
32    Areas of Interest Deep Learning Control System...
33    Skills R Python SAP HANA Tableau SAP HANA SQL ...
34    Education Details MCA YMCAUST Faridabad Haryan...
35    SKILLS C Basics IOT Python MATLAB Data Science...
36    Skills Python Tableau Data Visualization R Stu...
37    Education Details B Tech Rayat and Bahra Insti...
38    Personal Skills Ability to quickly grasp techn...
39    Expertise Data and Quantitative Analysis Decis...
Name: Cleaned_Resume, dtype: object
```

7. Create a corpus for data science resume text.

In [157…
```
data_science_corpus = " "

for index, row in data_science.iterrows():
    data_science_corpus += row['Cleaned_Resume']
```

In [158…
```
data_science_corpus=data_science_corpus.lower()
```

In [159…
```
words_data_science=data_science_corpus.split()
```

8. Find the frequencies of the important skills in Data science

In [160…
```
print('Frequency of "python"  is :', words_data_science.count("python"))
```
```
Frequency of "python"  is : 170
```

In [161…
```
print('Frequency of "sap"  is :', words_data_science.count("sap"))
```
```
Frequency of "sap"  is : 68
```

In [162…
```
print('Frequency of "analysis"  is :', words_data_science.count("analysis"))
```
```
Frequency of "analysis"  is : 76
```

In [163…
```
print('Frequency of "sql"  is :', words_data_science.count("sql"))
```
```
Frequency of "sql"  is : 71
```

In [164…
```
print('Frequency of "neural"  is :', words_data_science.count("neural"))
```
```
Frequency of "neural"  is : 47
```

In [165…
```
print('Frequency of "network"  is :', words_data_science.count("network"))
```
```
Frequency of "network"  is : 12
```

In [166…
```
print('Frequency of "networks"  is :', words_data_science.count("networks"))
```
```
Frequency of "networks"  is : 20
```

In [167…
```
print('Frequency of "pandas"  is :', words_data_science.count("pandas"))
```
```
Frequency of "pandas"  is : 23
```

In [168…
```
print('Frequency of "r"  is :', words_data_science.count("r"))
```
```
Frequency of "r"  is : 36
```

In [169…
```
print('Frequency of "excel"  is :', words_data_science.count("excel"))
```
```
Frequency of "excel"  is : 12
```

In [170…
```
print('Frequency of "anaconda"  is :', words_data_science.count("anaconda"))
```
```
Frequency of "anaconda"  is : 4
```

In [171…
```
print('Frequency of "jupyter"  is :', words_data_science.count("jupyter"))
```
```
Frequency of "jupyter"  is : 4
```

In [172…
```
print('Frequency of "education"  is :', words_data_science.count("education"
```
```
Frequency of "education"  is : 48
```

In [173…
```
print('Frequency of "experience"  is :', words_data_science.count("experienc
```
```
Frequency of "experience"  is : 52
```

In [173…