ASSIGNMENT NO. -4 :Resume Cleaning using NLP Techniques

NAME : JANHAVI KOLTE

ROLL NO : 33141

BATCH : K-9

Neccessary Imports

In [30]:

```
import numpy as np
import pandas as pd
import re
import nltk
from nltk.corpus import stopwords
import string
from wordcloud import WordCloud
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

In [44]:

```
nltk.download('wordnet')
```

```
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\hp\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

Out[44]:

```
True
```

Importing the dataset

In [22]:

```
df = pd.read_csv(r'Resume_Data.csv', encoding = 'utf-8')
df['Cleaned_Resume'] = ''
```

Exploratory Data Analysis

In [23]:

```
df.head()
```

Out[23]:

	Category	Resume	Cleaned_Resume
0	Data Science	Skills * Programming Languages: Python (pandas...	
1	Data Science	Education Details \r\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 \r\n MCA YMCAUST, Faridab...	

In [24]:

```
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
Data Science              40
Blockchain                40
ETL Developer             40
Operations Manager        40
Mechanical Engineer       40
Sales                     40
Arts                      36
Database                  33
Electrical Engineering    30
PMO                       30
Health and fitness        30
DotNet Developer          28
Business Analyst          28
Automation Testing        26
Network Security Engineer 25
Civil Engineer            24
SAP Developer             24
Advocate                  20
Name: Category, dtype: int64
```

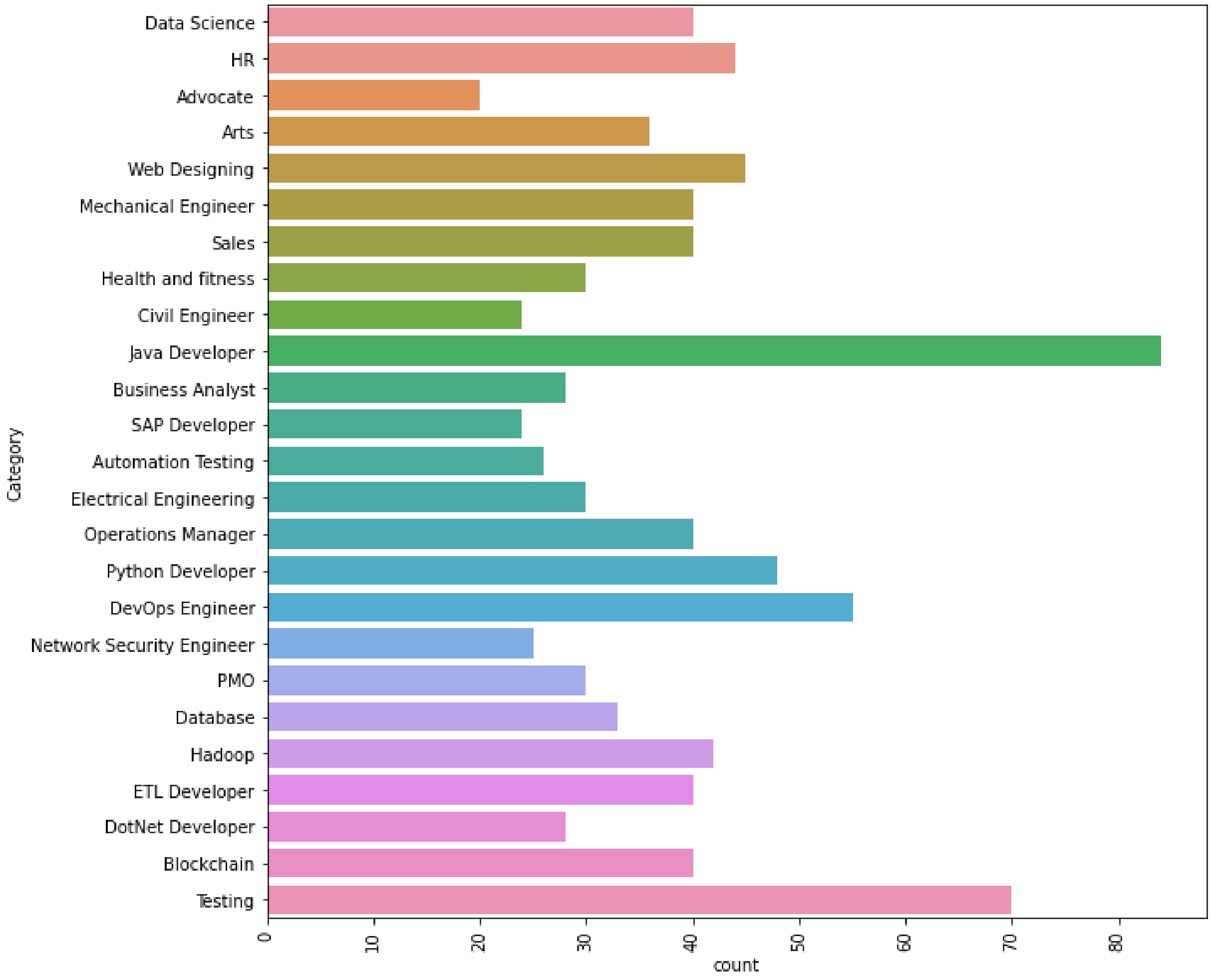
Visualizing types of people who have given the resume

In [29]:

```
plt.figure(figsize = (10, 10))                                    # Setting si
plt.xticks(rotation = 90)                                         # Rotating p
sns.countplot(y = 'Category', data = df)                          # Deciding w
```

Out[29]:

```
<AxesSubplot:xlabel='count', ylabel='Category'>
```



Data Cleaning

In [33]:

```
def Clean_Resume(resumeText):
    Removals = [                                                        # Deciding w
        'http\S+\s*',                                                   # Web URLs
        'RT|cc',                                                        # Regular ch
        '#\S+',                                                         # Hashtags
        '@\S+',                                                         # Emails
        '\s+'
    ]

    for weed in Removals: resumeText = re.sub(weed, ' ', resumeText)    # Removing w
    resumeText = re.sub('[%s]'%re.escape("""!"#$%&'_=-+()[];:,./?^*@{}|\~"""), ' ',
    resumeText = re.sub(r'[^x00-x7f]', r' ', resumeText)

    return resumeText
```

In [46]:

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

Out[46]:

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

In [37]:

```
corpus = ''
for i in range(len(df)): corpus += df['Cleaned_Resume'][i]
corpus[450:1000]
```

Out[37]:

```
'ticSearch D3 js  DC js  Plotl   kibana  matplotlib  ggplot  Tableau    Others  Regu
lar Expression  HTML  CSS  Angular 6  Logstash  Kafka  P thon Flask  Git  Docker  co
mputer vision   Open CV and understanding of Deep learning Education Details Data Sc
ience Assurance Associate Data Science Assurance Associate   Ernst    Young LLP Skill
Details JAVASCRIPT  Exprience   24 months jQuer   Exprience   24 months P thon  Expr
ience   24 monthsCompan  Details compan    Ernst   Young LLP description   Fraud Inv
estigations and Dispute Services Assurance TEC'
```

Creating the Tokenizer and Tokenizing

In [39]:

```
tokenizer = nltk.tokenize.RegexpTokenizer('\w+')
tokens = tokenizer.tokenize(corpus)                                     # Tokenizing

words = [word.lower() for word in tokens]                               # Transformi
print(len(words))
```

```
423116
```

Fetching English Stop Words

In [41]:

```
stopwords = nltk.corpus.stopwords.words('english')
```

Removing Stop words

In [43]:
```
words_new = [
    word
    for word in words
    if word not in stopwords
]
```

In [45]:
```
len(words_new)
```

Out[45]:
```
326374
```

Lemmatization

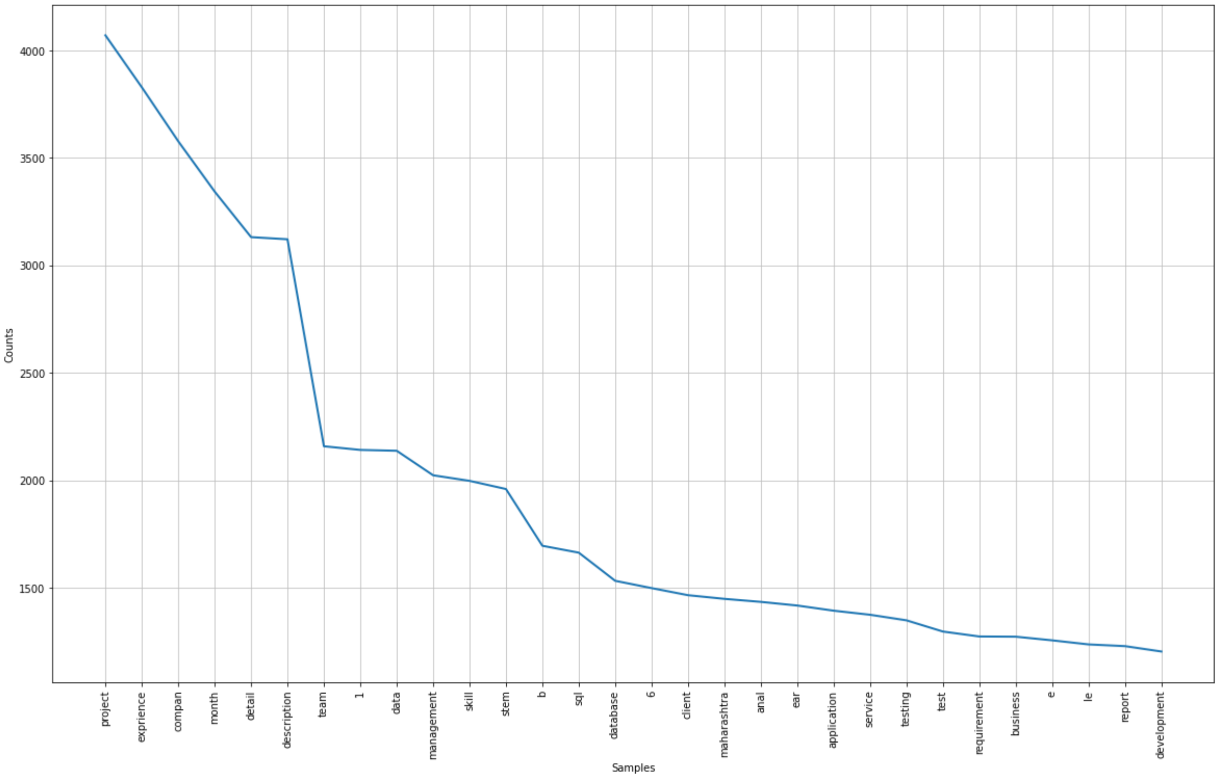
In [14]:
```
from nltk.stem import WordNetLemmatizer
wnl = WordNetLemmatizer()

lem_words = [
    wnl.lemmatize(word)
    for word in words_new
]
```

In [15]:
```
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= 311
Number of words not Lemmatized= 1521
```

In [16]:
```
freq_dist = nltk.FreqDist(lem_words)
plt.subplots(figsize=(20,12))
freq_dist.plot(30)
```



Out[16]:
```
<AxesSubplot:xlabel='Samples', ylabel='Counts'>
```

In [17]:
```
mostcommon = freq_dist.most_common(50)
mostcommon
```

Out[17]:
```
[('project', 4071),
 ('exprience', 3829),
 ('compan', 3578),
 ('month', 3344),
 ('detail', 3132),
 ('description', 3122),
 ('team', 2159),
 ('1', 2142),
 ('data', 2138),
 ('management', 2024),
 ('skill', 1998),
 ('stem', 1960),
 ('b', 1696),
 ('sql', 1664),
 ('database', 1533),
 ('6', 1499),
 ('client', 1466),
 ('maharashtra', 1449),
 ('anal', 1435),
 ('ear', 1418),
 ('application', 1394),
 ('service', 1375),
 ('testing', 1349),
 ('test', 1297),
 ('requirement', 1274),
 ('business', 1273),
 ('e', 1256),
 ('le', 1237),
 ('report', 1229),
 ('development', 1204),
 ('server', 1196),
 ('developer', 1194),
 ('customer', 1178),
 ('ltd', 1177),
 ('process', 1163),
 ('using', 1124),
 ('c', 1088),
 ('januar', 1086),
 ('java', 1076),
 ('engineering', 1055),
 ('work', 1038),
 ('pune', 1026),
 ('role', 969),
 ('ing', 925),
 ('user', 916),
 ('operation', 895),
 ('software', 886),
 ('pvt', 879),
 ('responsibility', 866),
 ('sale', 845)]
```

In [18]:
```
res=' '.join([i for i in lem_words if not i.isdigit()])
```

In [19]:
```
import os
os.system('pip install wordcloud')
```

Out[19]:
```
0
```

In [20]:
```
plt.subplots(figsize=(16,10))
wordcloud = WordCloud(
                          background_color='black',
                          max_words=200,
                          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)

anal tic
network securit
visual studio
design development
machine learning
pune maharashtra
role responsibility
requirement document
sale team
utili ation
operation manager
designed developed
project manager
html5 css3
window xp
web based
exprience month
phule pune
sql exprience
month database
month c
computer science
back end
project description
shell scripting
database exprience
service description
bootstrap php
technology description
detail januar
project detail
civil engineer
ear month
su essfull
javascript jquer
test plan
business requirement
communication skill
database sql
data warehouse
mechanical engineering
project name
organi ation
skill detail
asp net
used sql
software developer
software development
technical skill
ear monthscompan
team lead
ltd skill
core java
compan tata
technology used
description role
activit plc
india pvt
education detail
p thon
identif ing
testing exprience
completed internal
month java
deplo ment
description compan
end end
technology pvt
computer engineering
pol technic
role responsibilit
qualit check
information technolog
unit testing
exprience le
team member
b e
conducted b
good knowledge
sap bo
language c
januar januar
high school
januar bachelor
nendras technology
project management
emplo ee
programming language
development team
monthscompan detail
project title
nashik maharashtra
sale manager
oracle database
anal st
description project
management stem
anal si
b user
m office
developer skill
web application
sql server
stem testing
consultanc service
compan description
maharashtra pune
financial service
training completed
pvt ltd
mumbai maharashtra
software testing
test case
college engineering
su essful
maharashtra state
cs bootstrap
detail june
description responsibility
database administrator
thon exprience
detail compan
team si
php jquery
si e
januar ssc
control panel
c exprience
mumbai universit
requirement gathering
januar diploma
till date
limited description
exprience monthscompan
anal ing
pa ment
data science
design develop
review meeting
java developer
pune universit
engineer skill
business anal
bachelor engineering
tata consultanc
etl developer
c c la er
bny mellon
dot net
hone well
html cs
le ear
build deplo
ltd description
month javascript
well korea
stem window
universit januar
operating stem
nagpur maharashtra
anal tical
problem solving
related issue
description working
javascript exprience
anal e
deplo ed
compan detail
java developer
devops engineer
state board
shell script
januar b

In [63]: `df`

Out[63]:

	Category	Resume	Cleaned_Resume
0	Data Science	Skills * Programming Languages: Python (pandas...	Skills Programming Languages P thon pandas...
1	Data Science	Education Details \r\nMay 2013 to May 2017 B.E...	Education Details Ma 2013 to Ma 2017 B E UIT...
2	Data Science	Areas of Interest Deep Learning, Control Syste...	Areas of Interest Deep Learning Control S ste...
3	Data Science	Skills â ¢ R â ¢ Python â ¢ SAP HANA â ¢ Table...	Skills R P thon SAP HANA Table...
4	Data Science	Education Details \r\n MCA YMCAUST, Faridab...	Education Details MCA YMCAUST Faridabad Har ...
...	...	...	...
957	Testing	Computer Skills: â ¢ Proficient in MS office (...	Computer Skills Proficient in MS office ...

	Category	Resume	Cleaned_Resume
958	Testing	â Willingness to accept the challenges. â ...	Willingness to a ept the challenges P...
959	Testing	PERSONAL SKILLS â ¢ Quick learner, â ¢ Eagerne...	PERSONAL SKILLS Quick learner Eagerne...
960	Testing	COMPUTER SKILLS & SOFTWARE KNOWLEDGE MS-Power ...	COMPUTER SKILLS SOFTWARE KNOWLEDGE MS Power ...
961	Testing	Skill Set OS Windows XP/7/8/8.1/10 Database MY...	Skill Set OS Windows XP 7 8 8 1 10 Database MY...

962 rows × 3 columns

In [ ]: