

Assignment 4 - Resume Cleaning using NLP Techniques

Darshan Udikeri

TE-09

33120

K-09

Necessary Imports

In [2]:

```
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 [3]:

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

```
[nltk_data] Downloading package wordnet to /home/neil/nltk_data...
[nltk_data]   Unzipping corpora/wordnet.zip.
```

Out[3]:

True

Importing the dataset

In [4]:

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

Exploratory Data Analysis

In [5]:

```
df.head()
```

Out[5]:

	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 [6]:

```
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
Operations Manager	40
Data Science	40
Sales	40
Mechanical Engineer	40
Arts	36
Database	33
Electrical Engineering	30
Health and fitness	30
PMO	30
...	...

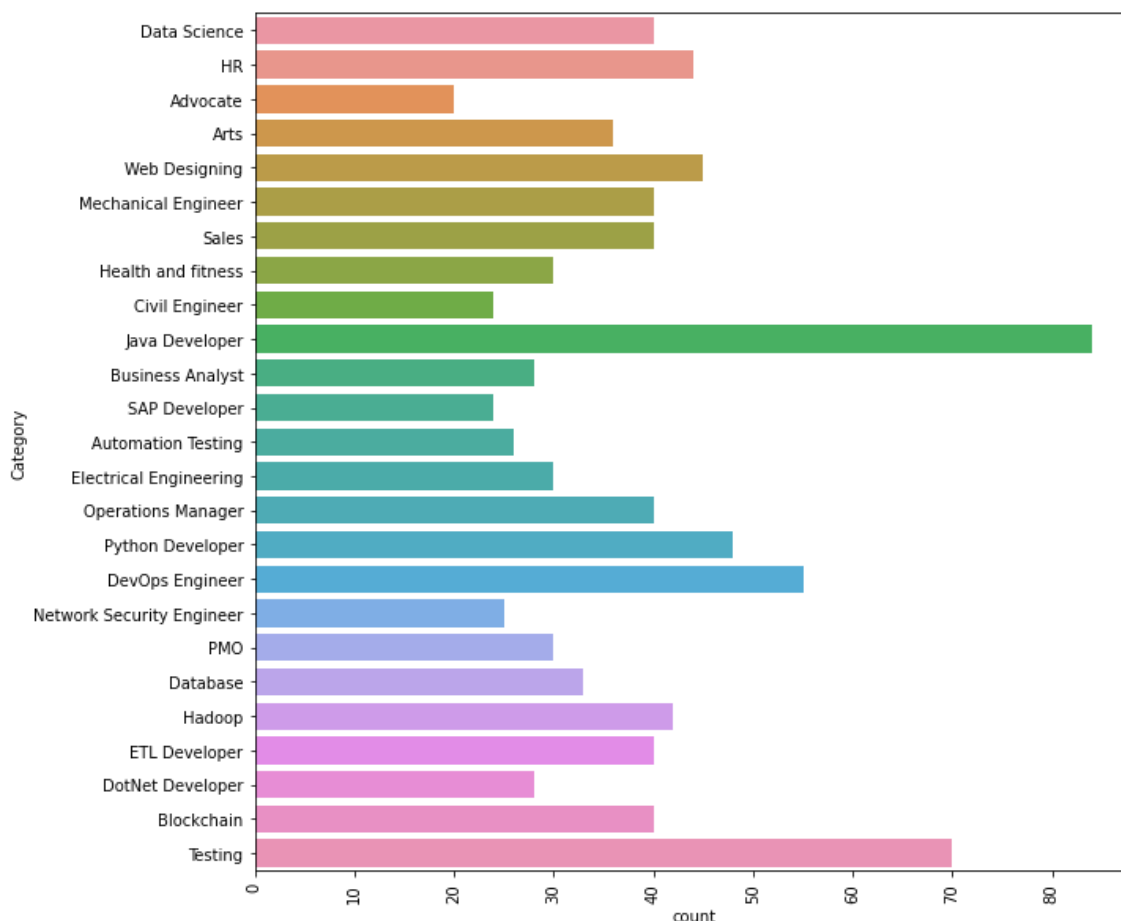
Visualizing types of people who have given the resume

In [7]:

```
plt.figure(figsize = (10, 10)) # Setting size of plot
plt.xticks(rotation = 90) # Rotating plot
sns.countplot(y = 'Category', data = df) # Deciding which axis to plot
```

Out[7]:

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



Data Cleaning

In [8]:

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

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

    return resumeText
```

In [9]:

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

Out[9]:

	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 [10]:

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

Out[10]:

```
'ticSearch D3 js DC js Plotl kibana matplotlib ggplot Tableau Ot
hers Regular Expression HTML CSS Angular 6 Logstash Kafka P thon Fl
ask Git Docker computer vision Open CV and understanding of Deep lear
ning Education Details Data Science Assurance Associate Data Science Assur
ance Associate Ernst Young LLP Skill Details JAVASCRIPT Exprience 2
4 months jQuer Exprience 24 months P thon Exprience 24 monthsCompan
Details compan Ernst Young LLP description Fraud Investigations and
Dispute Services Assurance TEC'
```

Creating the Tokenizer and Tokenizing

In [11]:

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

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

423116

Fetching English Stop Words

In [13]:

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

[nltk_data] Downloading package stopwords to /home/neil/nltk_data...

[nltk_data] Unzipping corpora/stopwords.zip.

Removing Stop words

In [14]:

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

In [15]:

```
len(words_new)
```

Out[15]:

326374

Lemmatization

In [16]:

```
from nltk.stem import WordNetLemmatizer
wnl = WordNetLemmatizer()

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

In [17]:

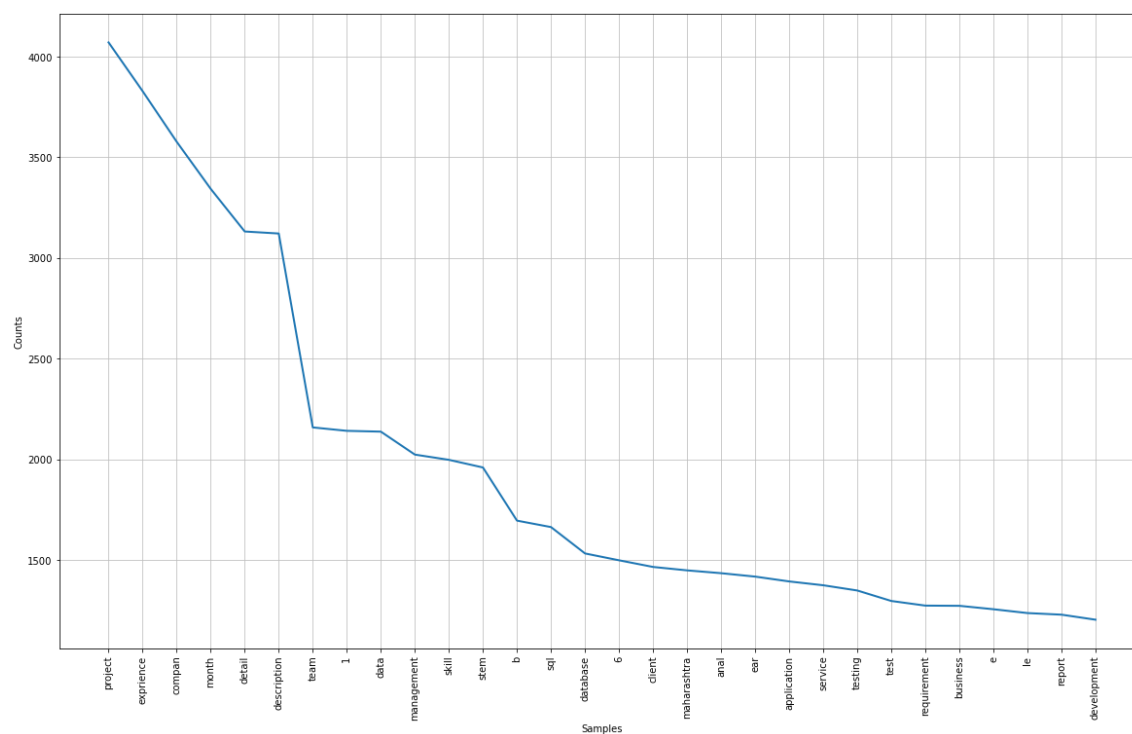
```
same=0
diff=0
for i in range(0,1832):
    if(lem_words[i]==words_new[i]):
        same=same+1
    elif(lem_words[i]!=words_new[i]):
        diff=diff+1
print('Number of words Lemmatized=', diff)
print('Number of words not Lemmatized=', same)
```

Number of words Lemmatized= 311

Number of words not Lemmatized= 1521

In [18]:

```
freq_dist = nltk.FreqDist(lem_words)
plt.subplots(figsize=(20,12))
freq_dist.plot(30)
```



Out[18]:

```
<AxesSubplot:xlabel='Samples', ylabel='Counts'>
```

In [19]:

```
mostcommon = freq_dist.most_common(50)
mostcommon
```

Out[19]:

```
[('project', 4071),
 ('expreience', 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)]
```

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

```
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()
```

[illegible]

In [23]:

df

Out[23]:

	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 ...
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 []: