Resume Cleaning using NLP Techniques

Problem Statement:

Write a program for the Information Retrieval System using appropriate NLP tools (such as NLTK, Open NLP, …) and perform following operations-

a. Text tokenization

b. Count word frequency

c. Remove stop words

d. POS tagging

Neccessary Imports

```
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
```

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

```
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data]   Unzipping corpora/wordnet.zip.
True
```

```
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
True
```

Importing the dataset

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

Exploratory Data Analysis

```
df.head()
```

	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...	

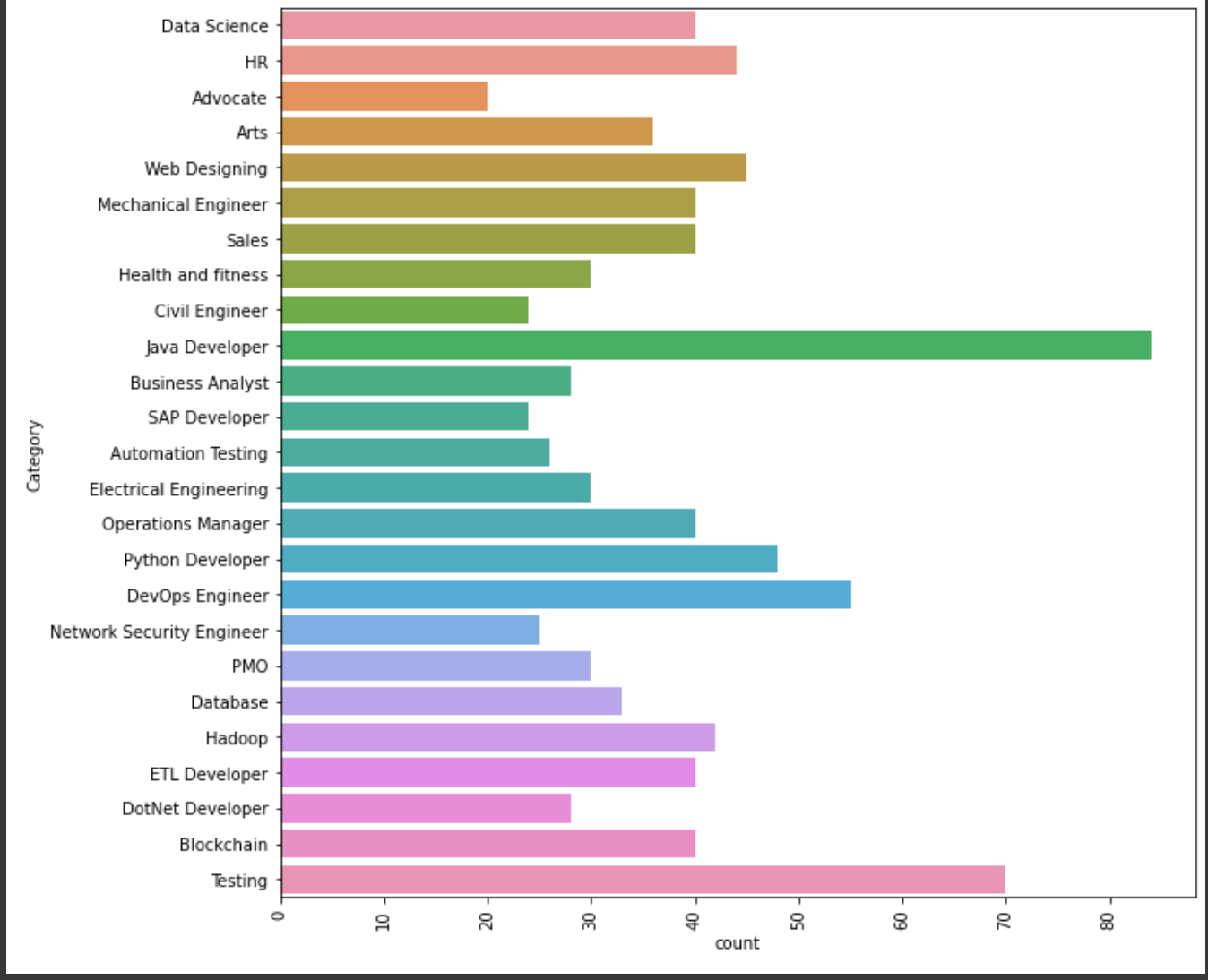
```
print("Resume Categories")
print(df['Category'].value_counts())
```

```
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
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
```

Visualizing types of people who have given the resume

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

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f96de65b810>
```



Data Cleaning

```
def Clean_Resume(resumeText):
    Removals = [                                                # Deciding weeds in resume
        'http\S+\s*',                                           # Web URLs
```

```
        'RT|cc',                                                    # Regular characters
        '#\S+',                                                     # Hashtags
        '@\S+',                                                     # Emails
        '\s+'
    ]

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

    return resumeText
```

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

	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...
	Data	Skills â¯¢ R â¯¢ Python â¯¢ SAP HANA	

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

```
'ticSearch D3 js  DC js  Plotl   kibana  matplotlib  ggplot  Tableau    Others
Regular Expression  HTML  CSS  Angular 6  Logstash  Kafka  P thon Flask  Git
Docker  computer vision   Open CV and understanding of Deep learning Education
Details Data Science Assurance Associate Data Science Assurance Associate   Er
```

Creating the Tokenizer and Tokenizing

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

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

```
423116
```

Fetching English Stop Words

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

Removing Stop words

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

```
len(words_new)
```

```
326374
```

Lemmatization

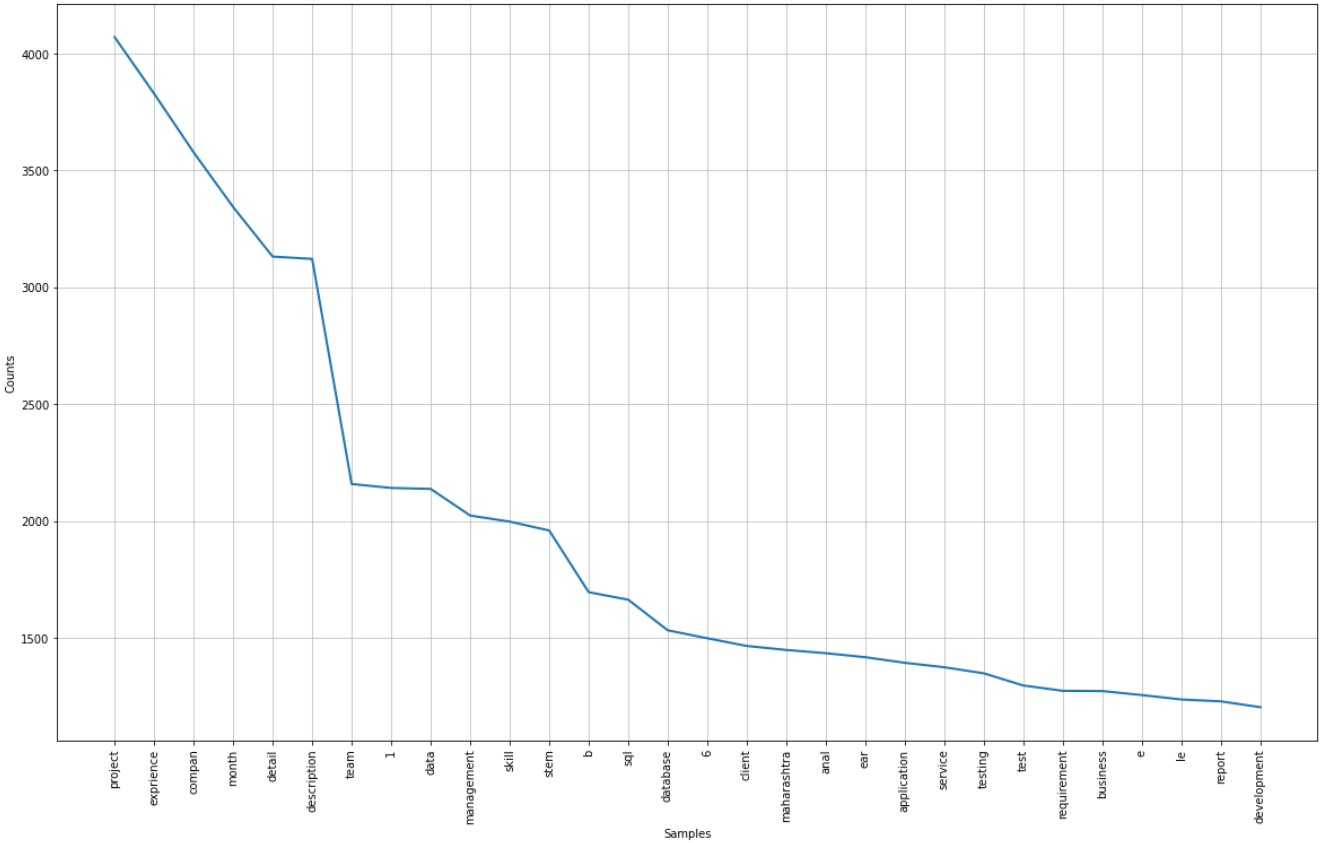
```
from nltk.stem import WordNetLemmatizer
wnl = WordNetLemmatizer()

```

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

```
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
```

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




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

```
[('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),
```

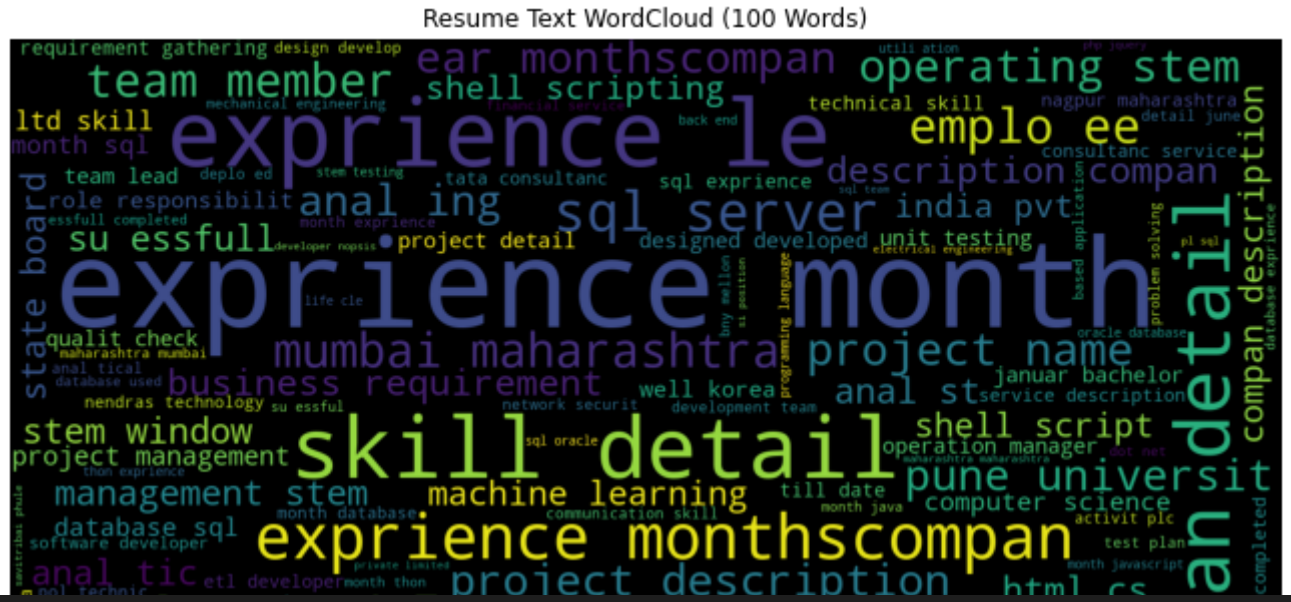
```
 ('responsibility', 866),
 ('sale', 845)]
```

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

```
import os
os.system('pip install wordcloud')
```

```
0
```

```
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)
requirement gathering design develop
ear monthscompan
operating stem
team member
shell scripting
ltd skill
exprience le
emplo ee
month sql
team lead
description compan
role responsibilit
anal ing
sql server
india pvt
su essfull
project detail
designed developed
unit testing
exprience month
state board
an detail
compan description
qualit check
mumbai maharashtra
project name
business requirement
anal st
januar bachelor
well korea
service description
stem window
project management
skill detail
shell script
operation manager
pune universit
management stem
machine learning
till date
computer science
database sql
exprience monthscompan
anal tic
project description
html cs

```
1 df
```

	Category	Resume	Cleaned_Resume
0	Data	Skills * Programming Languages: Python	Skills Programming Languages P thon
1	Data Science	2017 B.E...	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...

1