STEP 1: IMPORTING LIBRARIES

```
# INSTALLING NLTK, GENSIM AND WORDCLOUD
!pip install pandas
!pip install numpy
!pip install matplotlib
!pip install seaborn
!pip install --upgrade pip
!pip install nltk
!pip install gensim
!pip install sklearn
!pip install wordcloud
```

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud, STOPWORDS
import nltk
from nltk.stem import PorterStemmer, WordNetLemmatizer
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize, sent_tokenize
import gensim
from gensim.utils import simple_preprocess
from gensim.utils import simple_preprocess
from gensim.parsing.preprocessing import STOPWORDS
from sklearn.metrics import classification_report, confusion_matrix
```

```
Requirement already satisfied: python-dateutil>=2.7 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (2.8.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.7->matplotlib) (1.17.0)
Requirement already satisfied: seaborn in /usr/local/lib/python3.11/dist-packages (0.13.2)
Requirement already satisfied: numpy!=1.24.0,>=1.20 in /usr/local/lib/python3.11/dist-packages (from seaborn) (1.26.4)
Requirement already satisfied: pandas>=1.2 in /usr/local/lib/python3.11/dist-packages (from seaborn) (2.2.2)
Requirement already satisfied: matplotlib!=3.6.1,>=3.4 in /usr/local/lib/python3.11/dist-packages (from seaborn) (3.10.0)
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (1
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.11/dist-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (0.12
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.11/dist-packages (from matplotlib!=3.6.1,>=3.4->seaborn)
Requirement already satisfied: kiwisolver>=1.3.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib!=3.6.1,>=3.4->seaborn)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (24
Requirement already satisfied: pillow>=8 in /usr/local/lib/python3.11/dist-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (11.1.0)
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (3
Requirement already satisfied: python-dateutil>=2.7 in /usr/local/lib/python3.11/dist-packages (from matplotlib!=3.6.1,>=3.4->seaborn
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from pandas>=1.2->seaborn) (2025.1)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas>=1.2->seaborn) (2025.1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.7->matplotlib!=3.6.1,>=3
Requirement already satisfied: pip in /usr/local/lib/python3.11/dist-packages (25.0.1)
Requirement already satisfied: nltk in /usr/local/lib/python3.11/dist-packages (3.9.1)
Requirement already satisfied: click in /usr/local/lib/python3.11/dist-packages (from nltk) (8.1.8)
Requirement already satisfied: joblib in /usr/local/lib/python3.11/dist-packages (from nltk) (1.4.2)
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.11/dist-packages (from nltk) (2024.11.6)
Requirement already satisfied: tqdm in /usr/local/lib/python3.11/dist-packages (from nltk) (4.67.1)
Requirement already satisfied: gensim in /usr/local/lib/python3.11/dist-packages (4.3.3)
Requirement already satisfied: numpy<2.0,>=1.18.5 in /usr/local/lib/python3.11/dist-packages (from gensim) (1.26.4)
Requirement already satisfied: scipy<1.14.0,>=1.7.0 in /usr/local/lib/python3.11/dist-packages (from gensim) (1.13.1)
Requirement already satisfied: smart-open>=1.8.1 in /usr/local/lib/python3.11/dist-packages (from gensim) (7.1.0)
Requirement already satisfied: wrapt in /usr/local/lib/python3.11/dist-packages (from smart-open>=1.8.1->gensim) (1.17.2)
Collecting sklearn
  Using cached sklearn-0.0.post12.tar.gz (2.6 kB)
  error: subprocess-exited-with-error

  × python setup.py egg_info did not run successfully.
  | exit code: 1
```

```
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.11/dist-packages (from matplotlib->wordcloud) (4.56.0)
Requirement already satisfied: kiwisolver>=1.3.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib->wordcloud) (1.4.8)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from matplotlib->wordcloud) (24.2)
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib->wordcloud) (3.2.1)
Requirement already satisfied: python-dateutil>=2.7 in /usr/local/lib/python3.11/dist-packages (from matplotlib->wordcloud) (2.8.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.7->matplotlib->wordcloud)
```

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

STEP 2: LOADING THE DATASET

```
resume_df = pd.read_csv('resume_data.csv', encoding = 'latin-1')
resume_df
```

	resume_id	class	resume_text
0	resume_1	not_flagged	\rCustomer Service Supervisor/Tier - Isabella ...
1	resume_2	not_flagged	\rEngineer / Scientist - IBM Microelectronics ...
2	resume_3	not_flagged	\rLTS Software Engineer Computational Lithogra...
3	resume_4	not_flagged	TUTOR\rWilliston VT - Email me on Indeed: ind...
4	resume_5	flagged	\rIndependent Consultant - Self-employed\rBurl...
...	...	...	...
120	resume_121	not_flagged	\rBrattleboro VT - Email me on Indeed: indeed....
121	resume_122	not_flagged	\rResearch and Teaching Assistant - University...
122	resume_123	not_flagged	\rMedical Coder - Highly Skilled - Entry Level...
123	resume_124	flagged	\rWaterbury VT - Email me on Indeed: indeed.co...
124	resume_125	not_flagged	\rResearch and Development Scientist - Burling...

125 rows × 3 columns

Next steps: Generate code with resume_df | View recommended plots | New interactive sheet

```
resume_df = resume_df[['resume_text', 'class']]
resume_df
```

	resume_text	class
0	\rCustomer Service Supervisor/Tier - Isabella ...	not_flagged
1	\rEngineer / Scientist - IBM Microelectronics ...	not_flagged
2	\rLTS Software Engineer Computational Lithogra...	not_flagged
3	TUTOR\rWilliston VT - Email me on Indeed: ind...	not_flagged
4	\rIndependent Consultant - Self-employed\rBurl...	flagged
...	...	...
120	\rBrattleboro VT - Email me on Indeed: indeed....	not_flagged
121	\rResearch and Teaching Assistant - University...	not_flagged
122	\rMedical Coder - Highly Skilled - Entry Level...	not_flagged
123	\rWaterbury VT - Email me on Indeed: indeed.co...	flagged
124	\rResearch and Development Scientist - Burling...	not_flagged

125 rows × 2 columns

Next steps: Generate code with resume_df | View recommended plots | New interactive sheet

STEP 3: PERFORMING EXPLORATORY DATA ANALYSIS:

```
resume_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 125 entries, 0 to 124
Data columns (total 2 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   resume_text  125 non-null    object
 1   class        125 non-null    object
dtypes: object(2)
memory usage: 2.1+ KB
```

```
resume_df['class'].value_counts()
```

class	count
not_flagged	92
flagged	33

dtype: int64

```
# HERE WE OBSERVE, WE HAVE NO NULL POINTS IN OUR DATASET
resume_df['class'] = resume_df['class'].apply(lambda x:1 if x == 'flagged' else 0)
resume_df
```

```
<ipython-input-10-a97fb2daf353>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-cc
  resume_df['class'] = resume_df['class'].apply(lambda x:1 if x == 'flagged' else 0)
```

	resume_text	class
0	\rCustomer Service Supervisor/Tier - Isabella ...	0
1	\rEngineer / Scientist - IBM Microelectronics ...	0
2	\rLTS Software Engineer Computational Lithogra...	0
3	TUTOR\rWilliston VT - Email me on Indeed: ind...	0
4	\rIndependent Consultant - Self-employed\rBurl...	1
...	...	...
120	\rBrattleboro VT - Email me on Indeed: indeed....	0
121	\rResearch and Teaching Assistant - University...	0
122	\rMedical Coder - Highly Skilled - Entry Level...	0
123	\rWaterbury VT - Email me on Indeed: indeed.co...	1
124	\rResearch and Development Scientist - Burling...	0

125 rows × 2 columns

Next steps: Generate code with resume_df | View recommended plots | New interactive sheet

STEP 4: PERFORMING DATA CLEANING:

```
# PREMOVING UNNECESSARY WORDS FROM DATASET

resume_df['resume_text'] = resume_df['resume_text'].apply(lambda x: x .replace('\r', ''))

nltk.download('punkt')
nltk.download('stopwords')

from nltk.corpus import stopwords

stop_words = stopwords.words('english')

stop_words.extend(['from', 'subject', 'edu', 're', 'use', 'email', 'com'])

def preprocess(text):
```

```
    result = []
    for token in gensim.utils.simple_preprocess(text):
        if token not in gensim.parsing.preprocessing.STOPWORDS and len(token) > 2 and token not in stop_words:
            result.append(token)
    return ' '.join(result)
```

```
<ipython-input-11-b910c1193183>:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-cc
  resume_df['resume_text'] = resume_df['resume_text'].apply(lambda x: x .replace('\r', ''))
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

```
resume_df
```

	resume_text	class
0	Customer Service Supervisor/Tier - Isabella Ca...	0
1	Engineer / Scientist - IBM Microelectronics Di...	0
2	LTS Software Engineer Computational Lithograph...	0
3	TUTORWilliston VT - Email me on Indeed: indee...	0
4	Independent Consultant - Self-employedBurlingt...	1
...	...	...
120	Brattleboro VT - Email me on Indeed: indeed.co...	0
121	Research and Teaching Assistant - University o...	0
122	Medical Coder - Highly Skilled - Entry LevelSu...	0
123	Waterbury VT - Email me on Indeed: indeed.com/...	1
124	Research and Development Scientist - Burlingto...	0

125 rows × 2 columns

Next steps: Generate code with resume_df | View recommended plots | New interactive sheet

```
resume_df['cleaned'] = resume_df['resume_text'].apply(preprocess)
```

```
resume_df
```

	resume_text	class	cleaned
0	Customer Service Supervisor/Tier - Isabella Ca...	0	customer service supervisor tier isabella cata...
1	Engineer / Scientist - IBM Microelectronics Di...	0	engineer scientist ibm albert gregoritsch ecaw...
2	LTS Software Engineer Computational Lithograph...	0	lts software engineer computational lithograph...
3	TUTORWilliston VT - Email me on Indeed: indee...	0	tutorwilliston alec schwartz awork college bio...
4	Independent Consultant - Self-employedBurlingt...	1	independent consultant self alex reutter fefwo...
...	...	...	...
120	Brattleboro VT - Email me on Indeed: indeed.co...	0	brattleboro bcc skilled presenter trainer micr...
121	Research and Teaching Assistant - University o...	0	research teaching assistant university cdd gra...
122	Medical Coder - Highly Skilled - Entry LevelSu...	0	medical coder highly skilled entry levelsudbur...
123	Waterbury VT - Email me on Indeed: indeed.com/...	1	waterbury bec fcwilling relocate work employer...
124	Research and Development Scientist - Burlingto...	0	research development scientist burlington cda ...

125 rows × 3 columns

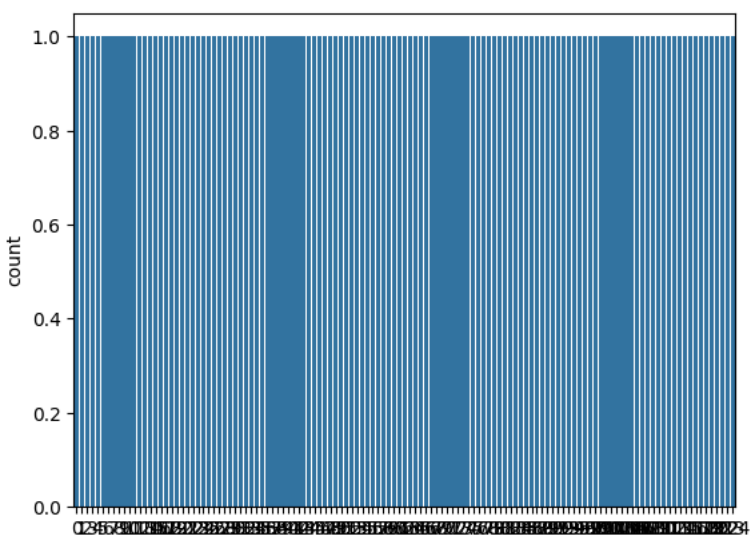
Next steps: Generate code with resume_df | View recommended plots | New interactive sheet

```
resume_df['cleaned'][0]
```

```
'customer service supervisor tier isabella catalog companysouth burlington aecf work service supervisor tierisabella catalog company sh
elburne august present customer service visual set display website maintenance supervise customer service team popular catalog company
manage day day issues resolution customer upset ensure customer satisfaction troubleshoot order shipping issues lost transit order erro
rs damages manage resolve escalated customer calls ensure customer satisfaction assist customers order placing cross selling upselling
catalog merchandise set display sample merchandise catalog library customer pick area facility website clean adding images type product
information assistant events coordinator office services assistanteileen fisher irvington february july support director architecture a
rchitecture coordinator daily activities including preparing monthly expense reports scheduling calendar maintenance arranging aspects
```

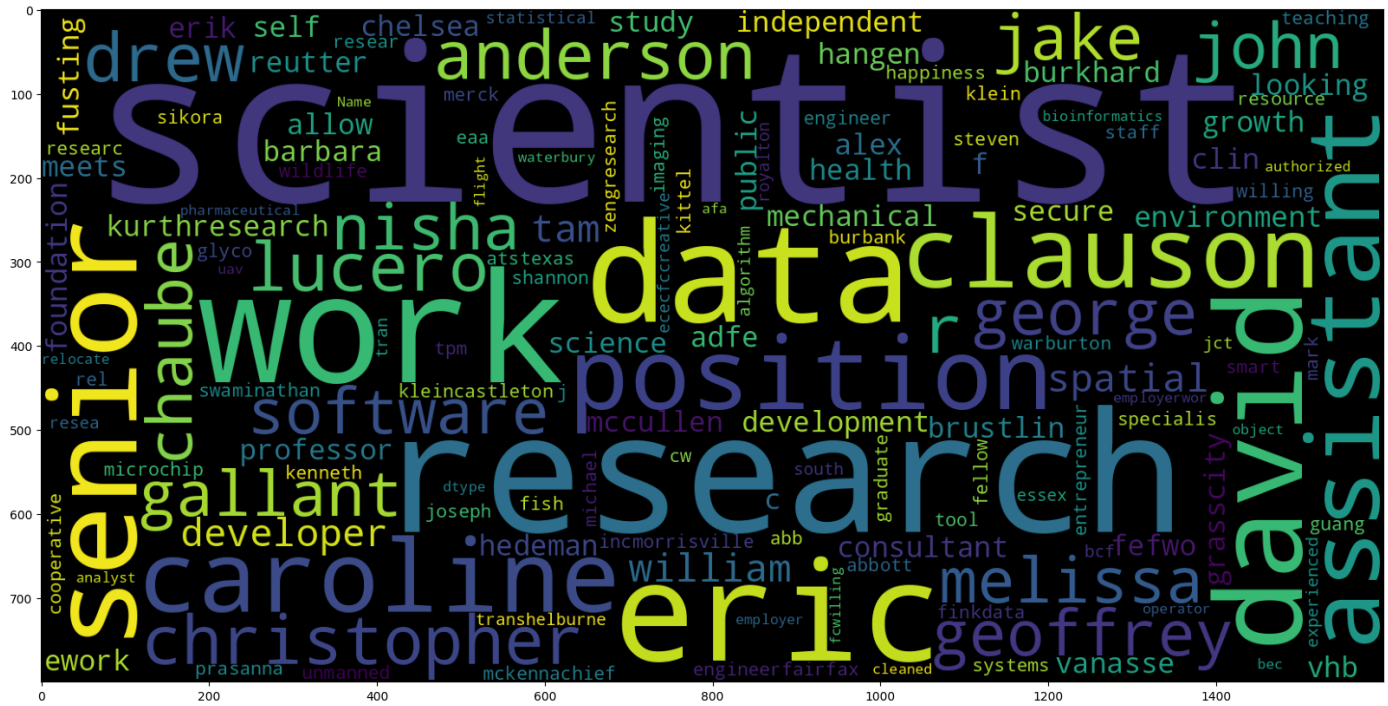
STEP 5: VISUALIZING CLEANED DATASETS

```
# PLOTTING COUNTS OF SAMPLE LABELLED AS 1 AND 0
sns.countplot(resume_df['class'], label = 'Count Plot')
plt.show()
```



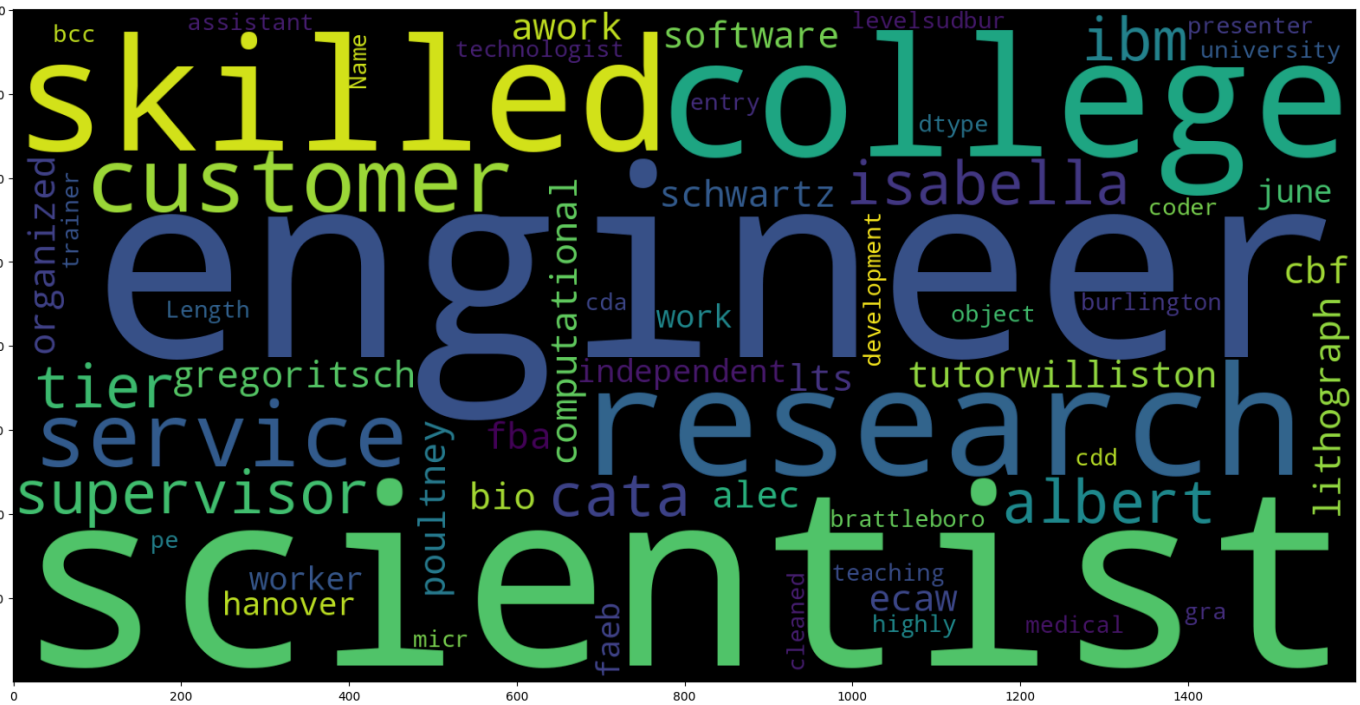

```
# PLOTTING THE WORDCLOUD:

# 1) FOR CLASS 1:

%matplotlib inline

plt.figure(figsize = (20, 20))

wc = WordCloud(max_words = 2000, width = 1600, height = 800, stopwords = stop_words).generate(str(resume_df[resume_df['class']==1].cleaned))

plt.imshow(wc)
```

```
<matplotlib.image.AxesImage at 0x7e5a2fb391d0>
```




```
#1) FOR CLASS 0:

%matplotlib inline

plt.figure(figsize = (20, 20))

wc = WordCloud(max_words = 2000, width = 1600, height = 800, stopwords = stop_words).generate(str(resume_df[resume_df['class']==0].cleaned))

plt.imshow(wc)
```

```
<matplotlib.image.AxesImage at 0x7e5a2e9ccdd0>
```



STEP 6: PREPARING THE DATA BY APPLYING COUNT VECTORIZATION

```
# CONVERTING SENTENCES INTO TOKENIZED FORMS AND THEN CONVERTING TO NUMERICAL VALUES IN ORDER FOR THE MODEL TO TRAIN:

from sklearn.feature_extraction.text import CountVectorizer

vectorizer = CountVectorizer()
countvectorizer = vectorizer.fit_transform(resume_df['cleaned'])

print(vectorizer.get_feature_names_out())
```

```
['aaalac' 'aabb' 'aac' ... 'ãæcomputer' 'ètravel' 'ô_torrent']
```

```
# PROCESSED DATA:
print(countvectorizer.toarray())
```

```
[[0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 ...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]
```

STEP 7: TRAINING A NAIVE BAYES CLASSIFER

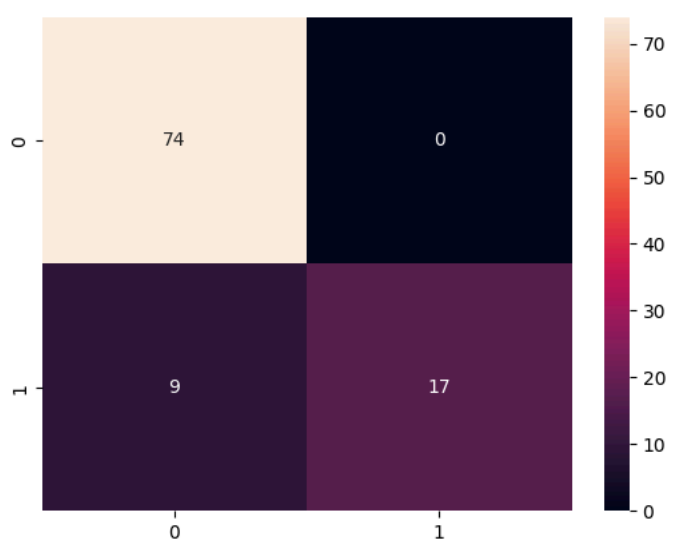
```
X = countvectorizer
y = resume_df['class']

from sklearn.model_selection import train_test_split


X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2)


from sklearn.naive_bayes import MultinomialNB
```

```
Bayes_clf = MultinomialNB(alpha = 3)
Bayes_clf.fit(X_train, y_train)  ## Training the model
```

```
MultinomialNB
MultinomialNB(alpha=3)
```

STEP 8: ASSESING THE TRAINED MODEL

```
%matplotlib inline

# PLOTTING CONFUSION MATRIX:

# 1) FOR TRAINING DATA

y_pred_train = Bayes_clf.predict(X_train)

cm = confusion_matrix(y_train, y_pred_train)

sns.heatmap(cm, annot=True)

plt.show()
```




```
%matplotlib inline

# WE CAN SEE OUR MODEL PERFORMED REALLY WELL ON TRAINING DATA: IT CLASSFIED ALL OF THE POINTS CORRECTLY

# 2) FOR TEST DATA:

y_pred_test = Bayes_clf.predict(X_test)

cm = confusion_matrix(y_test, y_pred_test)

sns.heatmap(cm, annot=True)

plt.show()
```


16
14