Notebook

May 30, 2025

[1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

[2]:
```
df=pd.read_csv(r"C:\Users\arjun\Downloads\archive\UpdatedResumeDataSet.csv")
df.head()
```

[2]:
```
       Category                                             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…
```

[4]:
```
df.shape
```

[4]:
```
(962, 2)
```

[6]:
```
df['Category'].unique()
```

[6]:
```
array(['Data Science', 'HR', 'Advocate', 'Arts', 'Web Designing',
       'Mechanical Engineer', 'Sales', 'Health and fitness',
       'Civil Engineer', 'Java Developer', 'Business Analyst',
       'SAP Developer', 'Automation Testing', 'Electrical Engineering',
       'Operations Manager', 'Python Developer', 'DevOps Engineer',
       'Network Security Engineer', 'PMO', 'Database', 'Hadoop',
       'ETL Developer', 'DotNet Developer', 'Blockchain', 'Testing'],
      dtype=object)
```

[5]:
```
df['Category'].value_counts()
```

[5]:
```
Category
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: count, dtype: int64
```

[7]:

```python
import re
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
import nltk
nltk.download('stopwords')

stop_words = set(stopwords.words('english'))
stemmer = PorterStemmer()

def preprocess(text):
    text = re.sub(r'\W', ' ', text)  # remove special chars
    text = re.sub(r'\d+', '', text)  # remove digits
    text = text.lower()  # to lowercase
    tokens = text.split()
    tokens = [stemmer.stem(word) for word in tokens if word not in stop_words]
    return ' '.join(tokens)

df['Cleaned_Resume'] = df['Resume'].apply(preprocess)
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\arjun\AppData\Roaming\nltk_data…
[nltk_data]   Unzipping corpora\stopwords.zip.
```

[8]:

```python
df.head()
```

```
[8]:        Category                                             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…

                                          Cleaned_Resume
     0  skill program languag python panda numpi scipi…
     1  educ detail may may b e uit rgpv data scientis…
     2  area interest deep learn control system design…
     3  skill â r â python â sap hana â tableau â sap …
     4  educ detail mca ymcaust faridabad haryana data…
```

```
[9]: from sklearn.feature_extraction.text import TfidfVectorizer

     tfidf = TfidfVectorizer(max_features=3000)
     X = tfidf.fit_transform(df['Cleaned_Resume']).toarray()
```

```
[10]: df.head()
```

```
[10]:        Category                                             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…

                                          Cleaned_Resume
     0  skill program languag python panda numpi scipi…
     1  educ detail may may b e uit rgpv data scientis…
     2  area interest deep learn control system design…
     3  skill â r â python â sap hana â tableau â sap …
     4  educ detail mca ymcaust faridabad haryana data…
```

```
[13]: from sklearn.preprocessing import LabelEncoder

      le = LabelEncoder()
      y = le.fit_transform(df['Category'])  # Target column
```

```
[15]: from sklearn.model_selection import train_test_split
      from sklearn.naive_bayes import MultinomialNB
      from sklearn.metrics import accuracy_score, classification_report

      X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,␣
       ↪random_state=42)
```

```
model = MultinomialNB()
model.fit(X_train, y_train)

y_pred = model.predict(X_test)
print("Accuracy:", accuracy_score(y_test, y_pred))
print(classification_report(y_test, y_pred, target_names=le.classes_))
```

```
Accuracy: 0.9896373056994818
                           precision    recall  f1-score   support

                 Advocate       1.00      1.00      1.00         3
                     Arts       1.00      1.00      1.00         6
       Automation Testing       1.00      1.00      1.00         5
               Blockchain       1.00      1.00      1.00         7
         Business Analyst       1.00      1.00      1.00         4
           Civil Engineer       1.00      1.00      1.00         9
             Data Science       1.00      1.00      1.00         5
                 Database       1.00      1.00      1.00         8
          DevOps Engineer       1.00      0.93      0.96        14
         DotNet Developer       1.00      1.00      1.00         5
            ETL Developer       1.00      1.00      1.00         7
   Electrical Engineering       1.00      1.00      1.00         6
                       HR       1.00      0.92      0.96        12
                   Hadoop       1.00      1.00      1.00         4
       Health and fitness       1.00      1.00      1.00         7
           Java Developer       0.94      1.00      0.97        15
      Mechanical Engineer       1.00      1.00      1.00         8
Network Security Engineer       1.00      1.00      1.00         3
       Operations Manager       1.00      1.00      1.00        12
                      PMO       0.88      1.00      0.93         7
         Python Developer       1.00      1.00      1.00        10
            SAP Developer       1.00      1.00      1.00         7
                    Sales       1.00      1.00      1.00         8
                  Testing       1.00      1.00      1.00        16
            Web Designing       1.00      1.00      1.00         5

                 accuracy                           0.99       193
                macro avg       0.99      0.99      0.99       193
             weighted avg       0.99      0.99      0.99       193
```

[16]:

```
def predict_resume_category(text):
    cleaned = preprocess(text)
    vectorized = tfidf.transform([cleaned])
    pred = model.predict(vectorized)
    return le.inverse_transform(pred)[0]
```

```
# Example:
new_resume = "Experience in developing machine learning models and data␣
 ↪analysis"
print(predict_resume_category(new_resume))
```

```
Data Science
```

[21]:

```
import fitz  # PyMuPDF
import re
import nltk
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.preprocessing import LabelEncoder
import pandas as pd

def extract_text_from_pdf(pdf_path):
    doc = fitz.open(pdf_path)
    text = ""
    for page in doc:
        text += page.get_text()
    return text

# STEP 6: Predict Resume Category
def predict_resume_category_from_pdf(pdf_path):
    raw_text = extract_text_from_pdf(pdf_path)
    cleaned = preprocess(raw_text)
    vectorized = tfidf.transform([cleaned])
    pred = model.predict(vectorized)
    return le.inverse_transform(pred)[0]

# === Example ===
pdf_file = r"C:\Users\arjun\Downloads\Arun Resumenew (2) (1).pdf"  # Put your␣
 ↪resume filename here
category = predict_resume_category_from_pdf(pdf_file)
print("Predicted Resume Category:", category)
```

```
Predicted Resume Category: Java Developer
```

This notebook was converted with convert.ploomber.io