AUTOMATIC RESUME CLASSIFICATION

EX.NO : 8

DATE : / /2025

DEVELOP A SYSTEM TO AUTOMATICALLY CLASSIFY RESUMES INTO JOB CATEGORIES (E.G., DATA SCIENTIST, SOFTWARE ENGINEER)

AIM:

To write a program to automatically classify resumes into job categories using Natural Language Processing (NLP) and Machine Learning techniques.

ALGORITHM:

Step 1: Start
Step 2: Import necessary libraries.
Step 3: Load and preprocess news dataset.
Step 4: Clean articles (lowercase, remove punctuation, stopwords).
Step 5: Convert text to TF■IDF vectors.
Step 6: Split into train and test sets.
Step 7: Train Logistic Regression model.
Step 8: Evaluate accuracy and classification report.
Step 9: Test with a custom input article.

PROGRAM:

```
import pandas as pd

import re

from sklearn.model_selection import train_test_split

from sklearn.feature_extraction.text import

TfidfVectorizer

from sklearn.linear_model import LogisticRegression

from sklearn.metrics import accuracy_score,

classification_report



# Clean text function

def clean_text(text):

    text = text.lower()

    text = re.sub(r'\S+@\S+', ' ', text)          #

remove emails

    text = re.sub(r'\d+', ' ', text)              #

remove numbers

    text = re.sub(r'[^a-zA-Z ]', ' ', text)       #

remove special chars

    text = re.sub(r'\s+', ' ', text).strip()      #
```

```
remove extra spaces
    return text

# Example dataset
data = {
    'resume_text': [
        "Experienced Data Scientist skilled in
Python, ML, SQL...",
        "Software Engineer with strong Java, DSA,
API development...",
        "Machine Learning Engineer with experience
in TensorFlow...",
        "Frontend Developer skilled in React, HTML,
CSS..."
    ],
    'label': ['Data Scientist','Software
Engineer','ML Engineer','Frontend Developer']
}
df = pd.DataFrame(data)
# Clean text
df['cleaned'] = df['resume_text'].apply(clean_text)
# Train-test split
X_train, X_test, y_train, y_test = train_test_split(
    df['cleaned'], df['label'], test_size=0.25,
random_state=42
)
# TF-IDF Vectorization
tfidf = TfidfVectorizer(max_features=5000)
X_train_tfidf = tfidf.fit_transform(X_train)
X_test_tfidf = tfidf.transform(X_test)
# Model training
model = LogisticRegression()
model.fit(X_train_tfidf, y_train)
```

```
# Predictions

pred = model.predict(X_test_tfidf)

print("Accuracy:", accuracy_score(y_test, pred))

print(classification_report(y_test, pred))

# Custom resume classification

custom_resume = "Strong Python, SQL, data analysis,

statistics..."

cleaned_custom = clean_text(custom_resume)

custom_vec = tfidf.transform([cleaned_custom])

result = model.predict(custom_vec)

print("Predicted Category:", result[0])
```

RESULT:

Thus, a program to automatically classify resumes into job categories using NLP and machine learning methods has been successfully executed.