Resume Categorization Project Documentation

Table of Contents

- Introduction
- Project Overview
- Data Preprocessing
- Text Classification
- Model Training and Evaluation
- Model Serialization
- Resume Prediction
- Streamlit Web Application
- Conclusion

1. Introduction

The Resume Categorization Project is a machine learning project that aims to automatically categorize resumes into different job categories based on the content of the resumes. This project involves data preprocessing, text classification, model training, and prediction.

2. Project Overview

The project is divided into several key components:

2.1 Data Preprocessing

In this step, the raw resume data is cleaned and preprocessed to prepare it for text classification. The preprocessing includes:

1.Removing URLs, hashtags, mentions, and special characters.
2.Eliminating punctuation.
3.Encoding non-ASCII characters.
4.Removing extra whitespace.

2.2 Text Classification

Text classification is the core of this project. It involves:

Encoding job categories as numerical labels using scikit-learn's LabelEncoder.
Utilizing TF-IDF (Term Frequency-Inverse Document Frequency) vectorization to convert the cleaned resume text into numerical features.
Training a machine learning classifier, in this case, a K-Nearest Neighbors (KNN) classifier using scikit-learn's OneVsRestClassifier.

2.3 Model Training and Evaluation

The model is trained on a dataset containing resumes and their corresponding job categories. Model performance is evaluated using accuracy as a metric to determine how well the classifier can categorize resumes into the correct job categories.

2.4 Model Serialization

Once the model is trained and evaluated, it is serialized and saved using the Python pickle library. This allows for easy reuse of the model without needing to retrain it each time.

2.5 Resume Prediction

After model serialization, the system is ready to predict the category of a new resume. The steps for

prediction include:

1.Loading the serialized model and TF-IDF vectorizer.
2.Cleaning the new resume data.
3.Transforming the cleaned resume text using the trained TF-IDF vectorizer.
4.Making a prediction using the loaded classifier.
5.Mapping the numerical prediction to the corresponding job category.

3. Data Preprocessing

The data preprocessing step is crucial for preparing the raw resume data for text classification. It involves removing noise and irrelevant information from the text, ensuring that only relevant features are used for classification.

4. Text Classification

Text classification is the heart of the project. It involves converting the text data into numerical features and training a machine learning classifier to predict job categories based on these features.

5. Model Training and Evaluation

The model is trained using a labeled dataset containing resumes and their corresponding job categories. Evaluation is performed to assess the accuracy of the model in categorizing resumes correctly.

6. Model Serialization

The trained model and TF-IDF vectorizer are serialized and saved to disk using the pickle library. This allows for easy loading and reuse of the model.

7. Resume Prediction

The prediction system is designed to categorize new resumes into job categories. It involves loading the serialized model and vectorizer, cleaning the new resume text, and making predictions based on the trained classifier.

8. Streamlit Web Application

The project now includes a Streamlit-based web application for resume screening. This web application provides a user-friendly interface for uploading resumes and obtaining predictions about job categories.

8.1 Streamlit Implementation

The Streamlit application is implemented using the Streamlit library. Users can upload resumes, and the application processes them using the trained model to predict the job category. The predicted category

is then displayed to the user.

8.2 Usage

To use the Streamlit web application, users can run the Python script containing the application. After launching the application, they can upload a resume, and the application will provide the predicted job category.

9. Conclusion

The Resume Categorization Project provides a practical solution for automating the categorization of resumes into job categories. It demonstrates the use of machine learning techniques, text preprocessing, and model serialization to build a functional system. This system can be extended and customized to fit specific needs in the field of human resources and recruitment.

For usage, simply load the trained model and vectorizer, clean the new resume data, and use the model to predict the job category.