Financial Statements Document Classification

Introduction

This project aims to classify financial statements documents based on their textual content using machine learning techniques. The provided code performs the following tasks:

- Reading HTML files from directories.
- Extracting text from HTML content.
- Preprocessing text data.
- Training and evaluating classification models.
- Predicting categories of new HTML files.

Approach

Data Collection and Preprocessing

The code reads HTML files from specified directories corresponding to different types of financial statements (e.g., Balance Sheets, Cash Flow, Income Statement, etc.). It then extracts text from these HTML files and assigns the labels respectively by saving the extracted text and labels into csv file.After extraction,preprocessing the text by removing non-alphanumeric characters and converting text to lowercase.

Model Selection and Training

Utilising two classification algorithms: Multinomial Naive Bayes and Logistic Regression.Trained these models using TF-IDF vectorization of text data. The TF-IDF vectorizer is used to convert text data into numerical feature vectors, and these vectors are then used to train the classification models.

Evaluation and Model Selection

The models are trained and evaluated using accuracy scores and classification reports on a test dataset. Additionally, I tried with multiple models including Support Vector Machine (SVM), Random Forest, and Gradient Boosting to find the best performing model. The best performing model is then saved for prediction.

Code Structure

The code consists of the following main components:

1. **Data Collection and Preprocessing:** Functions to read HTML files, extract text, and preprocess the text data.
2. **Model Training and Evaluation:** Utilising TF-IDF vectorization and classification algorithms to train models and evaluate their performance.
3. **Model Selection:** Explores multiple classification algorithms to find the best performing model based on accuracy scores.
4. **Prediction:** Predicts the category of new HTML files using the trained model.

Results

The code achieves the following results:

- Trained and evaluated Multinomial Naive Bayes and Logistic Regression models with an accuracy of 85.74% and 94.85%
- Explored other classification algorithms (SVM, Random Forest, Gradient Boosting) to find the best performing model.
- Saved the best performing model SVM for classification of financial docs statement with an accuracy score of 95.84%.
- Provided a function to predict the category of new HTML files using the trained model by giving 2 examples in the .ipynb file.

Conclusion

The code successfully performs text classification for financial statements using machine learning techniques. It provides a flexible framework for training, evaluating, and using classification models for categorizing financial documents based on their textual content.