Assignment: Table Classification from Financial Statements

KADAMBI KASHYAP

Introduction

This project aims to develop a robust machine learning pipeline for the classification and topic modelling of financial documents. The primary objectives are to classify documents into predefined categories and extract meaningful topics from the text using Latent Dirichlet Allocation (LDA) and Random Forest classifiers. A Streamlit application was developed to provide an interactive interface for users to upload HTML files, preprocess the text, and obtain predictions on document classes and topics.

Approach

1. Data Extraction and Preprocessing:
   - Text was extracted from HTML files using BeautifulSoup.
   - The extracted text was pre-processed through various steps including converting to lowercase, removing punctuation, newlines, URLs, numbers, and stop-words, and performing lemmatization to reduce words to their base forms.
2. Model Selection:
   - CountVectorizer was used to transform the text data into numerical format by capturing word frequencies.
   - Latent Dirichlet Allocation (LDA) was used for topic modelling to identify the latent topics in the text data.
   - Truncated Singular Value Decomposition (SVD) was also applied to perform dimensionality reduction and to extract significant topics.
   - A Random Forest Classifier was trained to classify documents into categories such as Balance Sheets, Cash Flow, Income Statement, Notes, and Others.

3. Evaluation and Tuning:
   - Models were evaluated using metrics like log-likelihood and perplexity for LDA, and accuracy scores for the Random Forest classifier.
   - Hyperparameters were tuned using grid search to find the optimal parameters for the number of topics in LDA and the configuration of the Random Forest classifier.

Model Implementation

- Preprocessing Pipeline:
  - Text preprocessing functions were defined to systematically clean and prepare the text data for modelling.
  - A combined preprocessing pipeline was implemented to apply these functions to the text data extracted from HTML files.
- Vectorization and Transformation:
  - The pre-processed text data was vectorized using CountVectorizer, transforming it into a format suitable for model input.
  - The vectorized data was then used to fit the LDA and SVD models for topic extraction.
- Model Training:
  - The Random Forest classifier was trained on the vectorized text data to predict document classes.
  - The LDA model was fitted to the vectorized data to identify and extract topics.
- Streamlit Application:
  - A user-friendly Streamlit application was developed to allow users to upload HTML files, preprocess the text, and view the predicted document class and extracted topics.
  - The application displays the predicted class from the Random Forest model and the topics from the SVD model along with their distributions.

Results

- ❖ Document Classification:
  - The Random Forest classifier demonstrated high accuracy in classifying financial documents into predefined categories.
  - The model was evaluated using cross-validation techniques to ensure its robustness and reliability.
- ❖ Topic Modelling:
  - The LDA model provided insights into the underlying topics within the documents, though high perplexity scores indicated room for improvement in topic representation.
  - The SVD model effectively reduced the dimensionality of the text data, highlighting significant terms within each topic and making it easier to interpret the results.
- ❖ Streamlit Application:
  - The Streamlit application successfully integrated all components, providing an interactive and easy-to-use interface for document classification and topic modelling.
  - Users can upload HTML files and receive immediate feedback on document classification and topic extraction, enhancing the utility of the developed models.

Conclusion

This project demonstrates the successful application of machine learning techniques for the classification and topic modelling of financial documents. The developed Streamlit application offers a practical tool for financial analysts and researchers to classify documents and extract meaningful topics. Future work could focus on optimizing the number of topics in the LDA model and exploring more advanced preprocessing techniques to further improve model performance. The high accuracy of the Random Forest classifier and the valuable insights from topic modelling underscore the effectiveness of the chosen approach.