Enhancing Text Classification in Information Retrieval: A Comprehensive Approach with TF-IDF, Naive Bayes, Word Embeddings, LSA, and SVM

About

This project focuses on leveraging machine learning techniques for text classification within the domain of Information Retrieval. The objective is to preprocess a dataset of training documents, apply TF-IDF vectorization, implement a Naive Bayes classifier, integrate word embeddings with Latent Semantic Analysis (LSA), and apply SVM for further analysis. The performance will be assessed on a distinct test dataset.The dataset for this project is provided here.

Project Overview

Your task is to implement the following components:

1. **Document Preprocessing:**
   - Read and preprocess the training documents.
   - Convert documents into a word list, tokenize, remove punctuation, and perform stemming.
2. **TF-IDF Vectorization:**
   - Utilize TF-IDF vectorization for each document after preprocessing.
3. **Naive Bayes Classification:**
   - Implement the Naive Bayes classifier using TF-IDF vectors.
   - Train the classifier on the training dataset.
4. **Word Embeddings and LSA with SVM Classification:**
   - Choose one of the following word embedding techniques: Word2Vec, GloVe, or FastText.
   - Apply the selected word embedding technique to represent words in a continuous vector space.
   - Apply Latent Semantic Analysis (LSA) to the word embedding-based document vectors to capture latent semantic structures.
   - Use Support Vector Machine (SVM) for classification with the LSA-transformed word embedding vectors as input features.
   - Train the SVM classifier on the training dataset.
5. **Using all Word Embeddings(Optional)**
   - Explore using Word2Vec, GloVe, and FastText embeddings separately with SVM for classification.
   - Train and evaluate SVM classifiers for each word embedding technique on the test dataset.
   - Compare the results obtained from Word2Vec, GloVe, and FastText embeddings in your report, highlighting any differences or similarities in their performance.
6. **Evaluation on Test Dataset:**
   - Evaluate the trained Naive Bayes classifier and SVM classifier on the test dataset.
   - Compare the results obtained from both classifiers, emphasizing any differences or similarities in their performance.
   - Report key classification metrics, including accuracy, precision, recall, and F1-score.

Deliverables

- Functions or methods for handling document preprocessing, applying TF-IDF vectorization, implementing the Naive Bayes classifier with TF-IDF vectors, selecting and applying one of Word2Vec, GloVe, or FastText embeddings, applying LSA to the word embedding-based document vectors, and using SVM for text classification.
- A comprehensive report summarizing key findings, challenges faced, and insights gained during the project, with a particular emphasis on the application of different techniques to information retrieval and text classification.

References

Document and reference the sources, libraries, and tools used in the project. Utilize the NLTK library for text processing and explore relevant literature on Information Retrieval techniques.