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

About

This project focuses on utilizing machine learning techniques for text classification within the domain of Information Retrieval. The primary objective is to preprocess a dataset of training documents, then implement a Naive Bayes classifier. Advanced approaches involve applying word embeddings followed by a Support Vector Machine (SVM) classifier and applying Latent Semantic Analysis (LSA) followed by Support Vector Machine (SVM). The performance will be evaluated on a separate test dataset 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. **Naive Bayes Classification:**
   - Implement the Naive Bayes classifier.
   - Train the classifier on the training dataset.
3. **Word Embeddings 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.
   - Utilize a function to transform the word embeddings into document embeddings. You are free to choose any sensible function that best captures the semantic meaning of the entire document.
   - Use Support Vector Machine (SVM) for classification with the document embeddings as input features.
   - Train the SVM classifier on the training dataset.
4. **LSA with SVM Classification:**
   - Apply Latent Semantic Analysis (LSA) to the documents 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 classifiers 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, implementing the Naive Bayes classifier, 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.