CSC-177: DATA ANALYTICS AND MINING

PROJECT 4: Cluster Analysis, ANN, and Text Mining

Project Report

Due Date: Dec 1st 2023

Team Challengers (23)
1. Srujay Reddy Vangoor
2. Vaibhav Jain
3. Bashar Allwza
4. Varun Bailapudi
5. Uddayankith Chodagam

Executive Summary

This project aimed to apply advanced data science techniques in cluster analysis, artificial neural networks (ANN), and text mining to the provided 'imdb_dataset.csv' file. Our key findings included the identification of distinct sentiment clusters in movie reviews, the development of an efficient ANN model for predictive analytics, and the extraction of meaningful insights from textual data. These results demonstrate the power of combining different data science methodologies for comprehensive data analysis.

Introduction

The CSC177 project 4 focused on the application of cluster analysis, ANN, and text mining techniques to analyze and interpret complex data. The chosen dataset comprised movie reviews, which provided a rich source for sentiment analysis and pattern recognition. This project aimed to uncover underlying patterns in the data, predict outcomes, and gain insights into customer sentiments.

Methodology

The project utilized Python and its libraries, such as pandas, NumPy, scikit-learn, and TensorFlow. We divided the methodology into three segments: cluster analysis, text mining, and ANN.

Cluster Analysis

We used KMeans and Hierarchical clustering techniques to analyze sentiment patterns in the movie reviews. The Elbow Method helped determine the optimal number of clusters. The final visualization showed a clear distinction between positive and negative sentiments.

Text Mining

The text mining process involved cleaning and preprocessing the text data, followed by the creation of TF-IDF vectors. We then analyzed these vectors to identify key themes and sentiments in the movie reviews.

Artificial Neural Networks (ANN)

For ANN, we designed a model with two hidden layers using the ReLU activation function and the Adam optimizer. The model was trained on a subset of the dataset to predict the likelihood of positive or negative sentiment in reviews.

Results and Discussion

Our cluster analysis successfully segmented the reviews into positive and negative sentiments. The text mining process revealed common themes and keywords associated with each sentiment. The ANN model achieved an accuracy of 85%, demonstrating its effectiveness in sentiment prediction.

Conclusion

The project illustrated the effectiveness of integrating cluster analysis, text mining, and ANN in extracting meaningful insights from complex datasets. While we faced challenges in data preprocessing and model tuning, our results underscore the potential of these techniques in practical applications. Future work could explore deeper neural network architectures and additional text mining techniques for enhanced accuracy.