Latent Semantic Indexing using Singular Value Decomposition

İpek Güneş Aymergen

***Abstract*—This project uses Singular Value Decomposition (SVD) to create Latent Semantic Indexing (LSI) from scratch. Information retrieval systems and document analysis frequently use Latent Structure Indexing (LSI), a potent indexing and retrieval technique. LSI makes it possible to extract latent semantic structures from massive document sets by utilizing SVD to create a semantic space. The objective of this study is to identify the underlying semantic links between terms and texts by applying LSI to a dataset of customer complaints about a corporation.**

I. Introduction

A popular mathematical method in information retrieval and natural language processing is called latent semantic indexing (LSI). It makes use of Singular Value Decomposition (SVD) to lower the term-by-document matrix's dimensionality, which makes it possible to uncover latent semantic structures in textual data. The goal of this project is to create LSI from scratch using SVD and apply it to a dataset of customer complaints. In the process, we hope to show how useful LSI is for revealing the latent semantic links between terms and documents in the context of analyzing customer complaints.

II. Dataset Description

Customer complaints against "Comcast Corporation" make up the dataset used for this study. It contains details about the complaint, including the author, the post date, the satisfaction rating, and the complaint's author. Only complaints from 2009 onward are included in the preprocessed dataset, and missing values are eliminated. To prepare the text data for additional analysis, text preparation techniques including tokenization, stopword removal, stemming, and lemmatization are also used.

III. Implementation Details

Preprocessing the text data includes tokenization, stopword removal, stemming, and lemmatization before LSI is implemented. Using the CountVectorizer module from the scikit-learn library, a term-by-document matrix is produced once the text data has been preprocessed. The matrices U, Sigma, and VT are then obtained by normalizing and decomposing this matrix using Singular Value Decomposition (SVD). NumPy arrays are used in the SVD implementation to calculate the decomposition from scratch.

IV. Experimental Setup

Quantitative measurements such as Mean Squared Error (MSE) and Frobenius Norm (FN) are employed in the experimental setting to evaluate the performance of the developed LSI model. The reconstruction errors are calculated for different values of the low-rank approximation's rank (k). The optimal value of k, which is obtained from the least MSE and FN errors, indicates the rank that is most appropriate for the approximation.

V. Results

The experiments show that the developed LSI model approximates the original term-by-document matrix with good performance. Visualizing the reconstruction mistakes across a range of k values is done through tables and charts. Based on the least MSE and FN errors, the ideal value of k is determined, offering information on the efficient dimensionality reduction that the LSI model is able to accomplish.

VI. Discussion

The effectiveness of the established LSI model is analyzed and the experiment results are interpreted in the discussion. Understanding the underlying links between terms and documents in the dataset is made easier by gaining insight into the latent semantic structures that the model captures. Future research directions and possible areas for further improvement are also suggested.

VII. Conclusion

To sum up, this project successfully builds Singular Value Decomposition (SVD) from scratch for Latent Semantic Indexing (LSI) and applies it to a dataset of customer complaints. The findings show how well LSI works to extract latent semantic structures from textual data, revealing important information about the connections between terms and texts. By demonstrating the practical uses of LSI, this study advances the fields of document analysis and information retrieval.