Experiment 6
Latent Semantic Indexing

I. EXPERIMENTATION DONE

In this lab we will be performing Latent Semantic Indexing on the 20 Newsgroup dataset

II. APPROACH

A. Setting up the environment

We downloaded the 20 Newsgroup dataset and imported different python libraries for our task such as numpy, pandas, natural language toolkit (nltk), re(for regular expressions) and sklearn (scikit-learn learn) for classification models.

B. Data analysis

The dataset of the 20 Newsgroup dataset contains two main columns as data and target. The data column contains the actual article of the newspaper and the target column indicates the class to which the given article belongs.

C. Data Cleaning, Tokenization and Vectorization

Before performing latent semantic analysis on these articles we have to perform to cleaning task to extract better semantics from our data. For this purpose first of all we lowercase the whole text followed by stop word removal and punctuation removal. Then we observed that our dataset contained some special characters like @, /, brackets etc so we removed those symbols using regex expressions. Finally after cleaning our dataset we tokenize the text data and create Term-Document Matrix using Sklearn's TF-IDF vectorizer.

D. Singular Value Decomposition

The we use svd to reduce the dimension of our text data. This will create a low rank approximation for Term-Document matrix of rank k which will be significantly smaller than the the original TF-IDF matrix. We use sklearn's TruncatedSVD function to obtain singular value decomposition of our TF-IDF matrix.

E. Getting optimal value for k

In order to determine the optimal number of k we have to minimize the Frobenius norm of the original matrix and the approximated matrix of rank k.

III. Discussion on the Results

We were able to reduce the dimensionality of our Term-Document matrix using SVD which gives us three smaller matrices S, U and V. To obtain k rank approximation of A we select first k singular values and truncate the 3 matrix accordingly.

IV. INTERPRETATION

Latent semantic indexing (LSI) is a text indexing and retrieval technique that use another mathematical technique known as singular value decomposition (SVD). It discovers patterns in the relationships between terms and their semantics in an unstructured collection of text.

V. CONCLUSION

Text data suffers from their high dimensionality. Latent semantic analysis in its core uses SVD to create a matrix of lower dimensionality. It further helps us to recognize patterns and relationships between words on the basis of semantics they carry.

VI. GITHUB LINK

Click here to visit the Lab6 GitHub repository.