Progress Summary
Of
Latent Semantic Analysis

About Project

1. Project Title

Latent Semantic Analysis

2. Introduction

- Latent semantic analysis (LSA) is a technique in natural language processing, in particular distributional semantics, of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. LSA assumes that words that are close in meaning will occur in similar pieces of text (the distributional hypothesis).
- A matrix containing word counts per document (rows represent unique words and columns represent each document) is constructed from a large piece of text and a mathematical technique called singular value decomposition (SVD) is used to reduce the number of rows while preserving the similarity structure among columns. Documents are then compared by taking the cosine of the angle between the two vectors (or the dot product between the normalizations of the two vectors) formed by any two columns. Values close to 1 represent very similar documents while values close to 0 represent very dissimilar documents.

Progress of a Project

3. Data Source

- Name - All the News
    - Link: https://www.kaggle.com/snapcrack/all-the-news
- About Dataset - There are 42571 articles in this dataset but we will be needing the first 200 which is feasible for the Topic Modeling

4. Pre-processing

- Remove URLs from the text.
- Remove the following characters: | : , ; & ! ? /.
- Lowercase and stemming to reduce word inflections.

- Remove user mentions from the text.
- Remove stop words from the text.

5. Feature Engineering

- Bag of Words
  - It is a technique of text modelling. We can say that it is a method of feature extraction with text data.
- TF-IDF
  - TF-IDF is a statistical measure that evaluates how relevant a word is to a document in a collection of documents.
- Single Value Decomposition
  - It is a factorization of a real or complex matrix that generalizes the eigen decomposition of a square normal matrix to any matrix.

6. Work Done

- Kaggle news extraction and pre-processing
- News visualization and term-doc matrix

7. Work To be Done

- LSA Implementation
- SVD Model
- K-means Clustering
- Logistic Regression
- Comparison and Conclusion

8. Team Member

Ravi Satvik	202018008
Omkar Chavan	202018037
Dev Patel	202018055
Smit Gandhi	202018057