# Latent Semantic Indexing

Rushil Gupta

3rd October 2024

# Introduction

This assignment is on using Latent Semantic Indexing (LSI) to find latent representations of documents and terms, and to use these representations to perform information retrieval.

# Methodology

## The Dataset

About 10000 books were scraped from Ashoka's library. This is a limitation since we do not have the actual text of the books and do not know much about them. Our current model assumes that the titles can fully encapsulate the information of the book.

## Term-Document Matrix

The foundation of LSI is the term-document matrix, $A$, where each row represents a term, each column represents a document, and each entry $a_{ij}$ represents some measure of the importance of term $i$ in document $j$.

Note that $A \in \mathbb{R}^{11783 \times 8460}$. In this assignment, I have explored 4 different ways to construct this matrix:

1. **Occurrences**: $a_{ij}$ is the raw count of term $i$ in document $j$.

2. **Normalized**: $a_{ij} = \frac{f_{ij}}{\sum_k f_{kj}}$, where $f_{ij}$ is the frequency of term $i$ in document $j$.

3. **Word Length**: $a_{ij} = \frac{f_{ij}}{\sum_k f_{kj}} \cdot \log(|i|)$, where $|i|$ is the length of term $i$. This was a modification I wanted to explore with the following rationale: a longer word is generally more informative than a shorter word, and would usually appear less number of times in a document, so it should be given more weight.

4. **TF-IDF**: $a_{ij} = \frac{f_{ij}}{\sum_k f_{kj}} \cdot \log(\frac{N}{n_i})$, where $N$ is the total number of documents and $n_i$ is the number of documents containing term $i$ [1].

## Singular Value Decomposition

Now, by applying SVD to the matrix, we get the following result:

$$A = U\Sigma V^T$$

Where $U$ and $V$ are orthogonal matrices and $\Sigma$ is a diagonal matrix containing the singular values of $A$. Now, to reduce the dimensionality of the matrix (which allows us to do a couple of things: reduce noise, reduce the explaination's dimensionality, and also increase stability by increasing precision), we truncate the matrices to $k$ dimensions:

$$A_k = U_k \Sigma_k V_k^T$$

Where $k$ is the number of singular values we want to keep, and $A_k$ is the rank-$k$ approximation of $A$.

## Term and Document Spaces

Now, having the matrices $U$ and $V$, we can change the basis of the term and document matrices into the latent semantic space. All we mean by this is:

- Each column of $U$ represents a unique concept in the space. Each row represents the importance of a term to that concept.
- Each row of $V^T$ represents a unique concept in the space. Each column represents the importance of a document to that concept.
- $\Sigma$ scales the importance of each concept.

Now, to understand how we can convert documents and terms into this latent semantic space, we need can do the following:

- For each document $d_j$, we know that the $j^{th}$ column of $\Sigma V^T$ represents the document in the latent semantic space.
- For each term $t_i$, we know that the $i^{th}$ row of $\Sigma U^T$ represents the term in the latent semantic space.

## Query Processing

To process a query $q$, we first represent it as a vector in the term space (with the same preprocessing as the term-document matrix), then project it into the $k$-dimensional LSI space:

$$q_k = q^T U_k \Sigma_k^{-1}$$

Now, we can compute the similarity between the query and a document $d$ in the LSI space using cosine similarity:

$$\text{similarity}(q_k, d_k) = \frac{q_k \cdot d_k}{\|q_k\| \|d_k\|}$$

Documents are then ranked based on their similarity to the query vector in the LSI space, and the top $N$ results are returned.

# Results

As mentioned before, 4 different types of matrices were used to construct the term-document matrix. Moreover, 3 different ranks were used to see the effect of dimensionality reduction. The following 3 queries were used: "Agriculture", "Artificial Intelligence", and "Policy". The results are shown in Table 1.

## PCA Visualization

Since LSI allows us to plot the data in a lower-dimensional space, I also tried to visualize the data using PCA. The following plots show the first two principal components of the documents in the latent semantic space (the terms were too many and too condensed to be plotted).

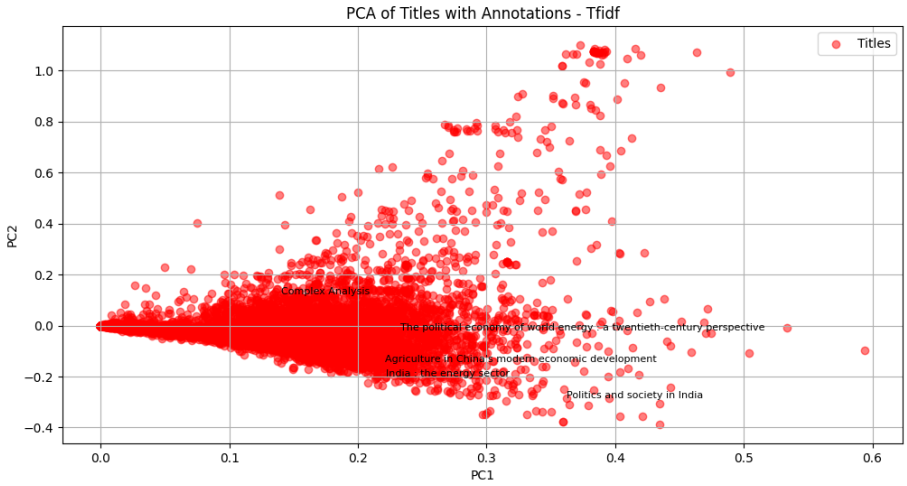


Figure 1: PCA of Titles

Again, we do not see any distinct clusters. This is likely because for each document, we have only one title, and thus very little information about the document itself. Therefore, the principal components are not able to distinguish the documents.

# Conclusion

In this assignment, I explored Latent Semantic Indexing and performed information retrieval on a set of 10k books. The model was constructed with the following steps:

1. Construct a term-document matrix $A$ in 4 different ways.
2. Perform SVD on $A$ and compute the singular values.
3. Reduce the dimensionality of $U$, $\Sigma$, and $V$ using the top $k$ singular values.
4. Convert documents and terms into the latent semantic space.
5. Perform information retrieval on the documents in the latent semantic space.
6. Visualize the results using PCA.

# Appendix

| Query | Matrix Type | Rank | Top Result |
|---|---|---|---|
| Agriculture | Occurrences | 100 | Surplus manpower in agriculture and economic development with special reference to India |
| | | 500 | Ms Militancy |
| | | 2500 | Algal biotechnology and its importance in agri. |
| | Normalized | 100 | Agriculture in China's modern economic dev. |
| | | 500 | Agriculture in China's modern economic dev. |
| | | 2500 | Algal biotechnology and its importance in agri. |
| | Word Length | 100 | Agriculture in China's modern economic dev. |
| | | 500 | The economic structure of backward agriculture |
| | | 2500 | Algal biotechnology and its importance in agri. |
| | TF-IDF | 100 | The River Pollution Dilemma in Victorian England: Nuisance Law versus Economic Efficiency |
| | | 500 | Agriculture in China's modern economic dev. |
| | | 2500 | Essays on the commercialization of Indian agri. |
| Artificial Intelligence | Occurrences | 100 | Confidence : Emotional Intelligence |
| | | 500 | Confidence : Emotional Intelligence |
| | | 2500 | Confidence : Emotional Intelligence |
| | Normalized | 100 | Confidence : Emotional Intelligence |
| | | 500 | Confidence : Emotional Intelligence |
| | | 2500 | Inside intelligence |
| | Word Length | 100 | Confidence : Emotional Intelligence |
| | | 500 | Confidence : Emotional Intelligence |
| | | 2500 | Inside intelligence |
| | TF-IDF | 100 | Introduction to Artificial Intelligence |
| | | 500 | Confidence : Emotional Intelligence |
| | | 2500 | Introduction to Artificial Intelligence |
| Policy | Occurrences | 100 | Rethinking Monetary -Fiscal policy coordination |
| | | 500 | Rethinking Monetary -Fiscal policy coordination |
| | | 2500 | China's Foreign Policy |
| | Normalized | 100 | Rethinking Monetary -Fiscal policy coordination |
| | | 500 | Rethinking Monetary -Fiscal policy coordination |
| | | 2500 | China's Foreign Policy |
| | Word Length | 100 | Rethinking Monetary -Fiscal policy coordination |
| | | 500 | Rethinking Monetary -Fiscal policy coordination |
| | | 2500 | China's Foreign Policy |
| | TF-IDF | 100 | China's Foreign Policy |
| | | 500 | China's Foreign Policy |
| | | 2500 | A theory of wage policy |

Table 1: Top results for different queries, matrix types, and ranks

# References

[1] Wikipedia. (2024). Tf-idf. https://en.wikipedia.org/wiki/Tf%E2%80%93idf