Bangla Topic Modeling using Non-negative Matrix Factorization

Created by

Rony Majumder

Roll : 1325

Supervised by

Mohd. Zulfiquar Hafiz
Professor,IITDU

Project Overview

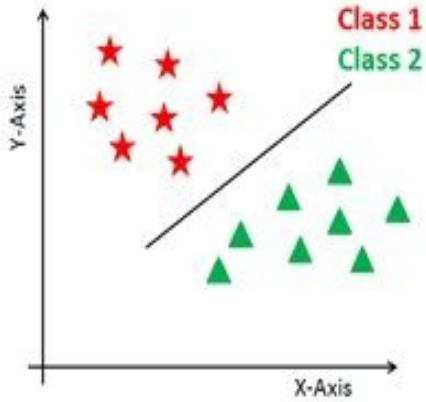
Topic modeling is a computational technique used in natural language processing (NLP) to discover latent themes or topics within a collection of documents.

This project aims to implement Non-Negative Matrix Factorization (NMF) for topic modelling, a technique used to extract latent topics from a collection of textual documents. By utilizing NMF, the project will analyze a given corpus, factorize it into non-negative matrices representing document-topic and topic-word distributions, and subsequently identify and label the underlying topics. The project will enable efficient exploration and understanding of large text datasets of **Bangla text** , providing valuable insights into the prominent themes and trends present within the documents.

Visual Project Overview

TF-IDF

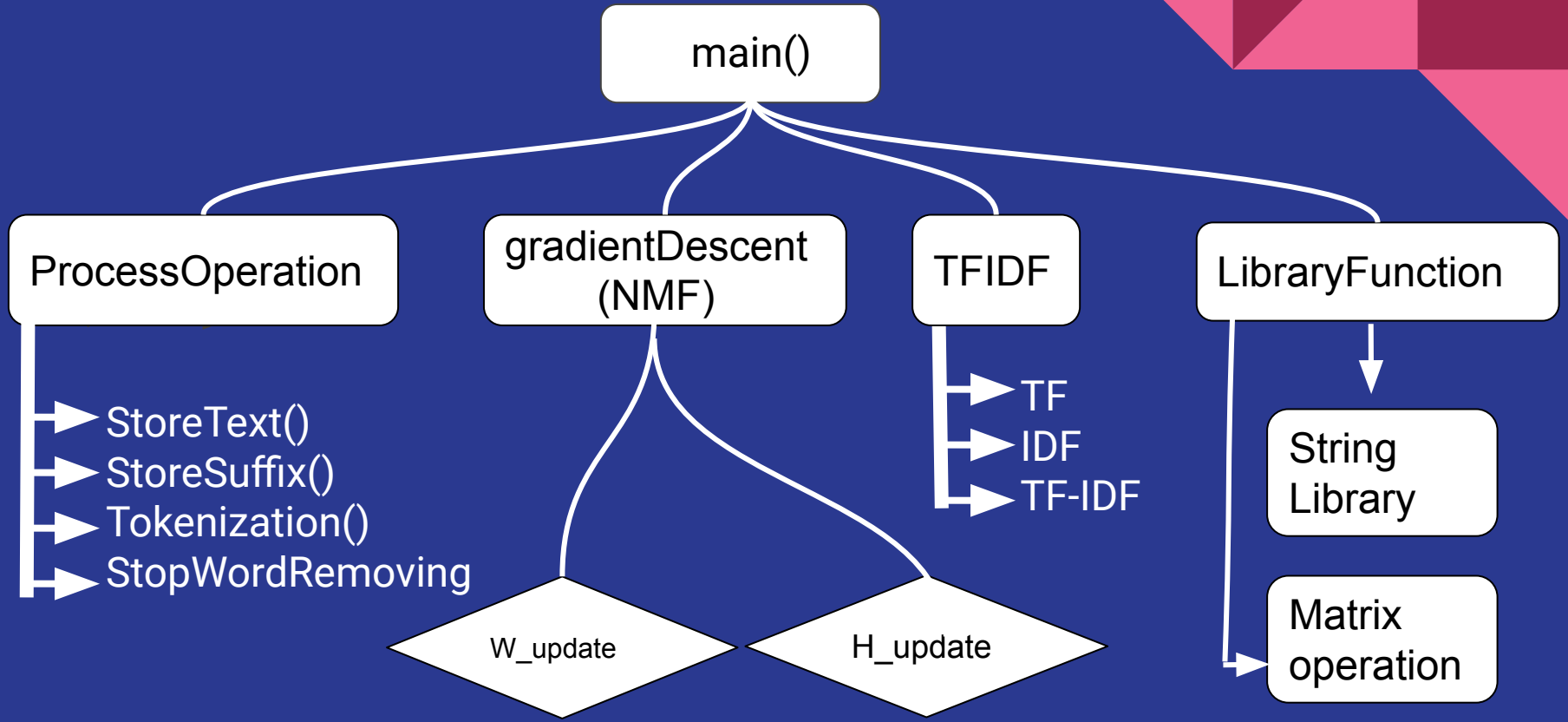
Text Classification



Topic Modeling

	Doc-1	Doc-2	Doc-3	Doc-4
Topic-1	0.425	0.032	0.015	0.553
Topic-2	0.052	0.512	0.623	0.121

Flow of project



Non negative Matrix Factorization is being implemented in this project by using gradient descent

NMF Algorithms

- **Multiplicative Update Rule** for **W** and **H** matrices:

$$H_{i,j} = H_{i,j} \frac{(W^T A)_{i,j}}{(W^T WH)_{i,j} + \epsilon}$$

$$W_{i,j} = W_{i,j} \frac{(AH^T)_{i,j}}{(WHH^T)_{i,j} + \epsilon}$$

TF-IDF dimension = m x n where row = number of sentence, col = cluster

$$TF(t,d) = \frac{number\ of\ times\ t\ appears\ in\ d}{total\ number\ of\ terms\ in\ d}$$

$$IDF(t) = log \frac{N}{1 + df}$$

$$TF - IDF(t,d) = TF(t,d) * IDF(t)$$

Project outcomes and references

Main.cpp	138
gradientDescent.cpp	306
MatrixOperation.cpp	251
Printmatrix.cpp	58
processOperation.cpp	238
StringOperation.h	194
TF-IDF.cpp	120
Others	40
total—----------	1350(+)

Project Outcomes:

- participated in "Bengali Grammatical Error Detection 2023" organized by Bengali.Ai [https://www.kaggle.com/competitions/bengali-ged/leaderboard]
- Learnt about natural language processing.
- Working with multiple source files

References

[1] https://en.wikipedia.org/wiki/Non-negative_matrix_factorization

[2] Lee, Daniel, and H. Sebastian Seung. "Algorithms for non-negative matrix factorization." Advances I n neural information processing systems 13 (2000)

[3] https://github.com/rony31416/SPL-1