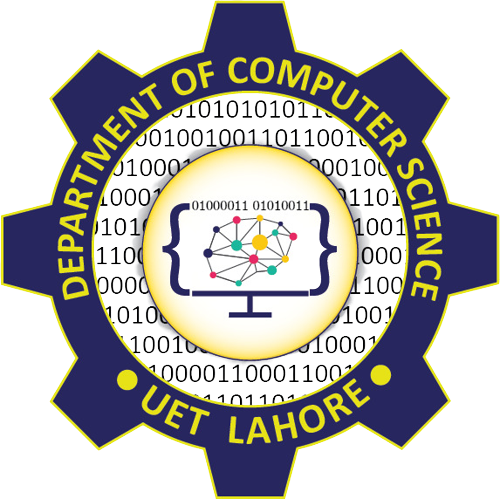
Set-Theoretic IR Model

Generalized Vector Model



Session: 2021 – 2025

# Submitted By:

Huzaifa Mumtaz 2021-CS-164

# Supervised By:

Dr. Khaldoon Syed Khurshid

Department of Computer Science

University of Engineering and Technology Lahore Pakistan

[1. Introduction 3](#_Toc25347)

[1.1. Introduction 3](#_Toc9578)

[1.2. Purpose of Assignment 3](#_Toc20227)

[2. Background 4](#_Toc31677)

[3. Methodology 5](#_Toc17711)

[3.1. Document Indexing 5](#_Toc21863)

[3.2. Search Features 5](#_Toc3265)

[3.3. Cosine Similarity 5](#_Toc26016)

[3.4. User Interface 5](#_Toc31083)

[4. Results/Output 6](#_Toc20801)

[5. Conclusion 7](#_Toc1231)

# Introduction

## Introduction

The Generalized Vector Model (GVM) is a widely used Information Retrieval (IR) model that represents documents and queries as weighted vectors in a multidimensional space. It enhances traditional vector models by accommodating term dependencies, resulting in more accurate document retrieval. This project implements a GVM-based document search system in C++, allowing efficient retrieval and ranking of documents stored in a specified directory.

## Purpose of Assignment

The purpose of this assignment is to provide hands-on experience with the implementation of the Generalized Vector Model (GVM), a foundational concept in Information Retrieval (IR). Through this assignment, students learn to represent documents and queries as vectors, apply term weighting schemes, and use cosine similarity to rank documents based on relevance. The assignment emphasizes practical skills, such as designing a user-friendly document search system, processing real-world datasets, and handling challenges like tokenization and file extension management. By completing this task, students bridge theoretical IR concepts with practical applications, preparing them for more advanced topics in information retrieval and search engine development.

# Background

Information Retrieval (IR) systems are essential for managing and retrieving relevant information from large datasets. The Generalized Vector Model (GVM) is an extension of the traditional Vector Space Model (VSM) used in IR. It represents documents and queries as vectors in a multidimensional space, where each dimension corresponds to a unique term. By leveraging term weighting schemes and similarity measures, such as cosine similarity, GVM enhances the precision and ranking of search results.

The model addresses the limitations of Boolean retrieval by allowing partial matches and ranking documents based on their relevance to a given query. This approach is widely applied in search engines, digital libraries, and content management systems. Implementing the GVM provides valuable insights into how modern IR systems efficiently manage and process unstructured data, making it an essential topic for understanding foundational IR techniques.

# Methodology

## Document Indexing

* All files in the specified directory were read and tokenized into individual terms.
* Term frequencies were calculated for each document and normalized using term frequency (TF) weights.
* The documents were stored as vectors in a term-document matrix.

## Search Features

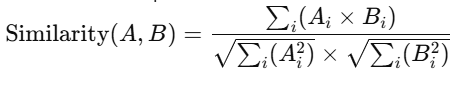
1. **By Document Name**

* Users can search for a document by name, with the system ignoring file extensions for more flexible matching

1. **By Keyword**

* The system allows users to input a query, tokenizes it, and converts it into a vector.
* Cosine similarity is calculated between the query vector and document vectors, ranking documents by relevance.

## Cosine Similarity

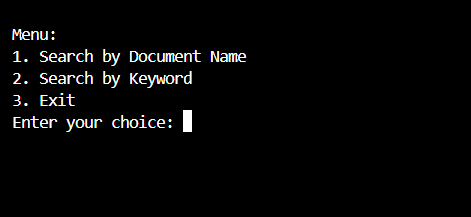
* The similarity between vectors is computed as:
* This measure ensures that documents with terms most similar to the query are ranked higher.

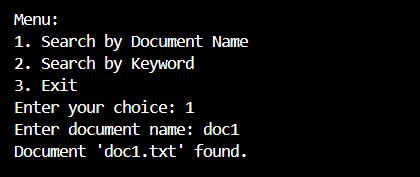
## User Interface

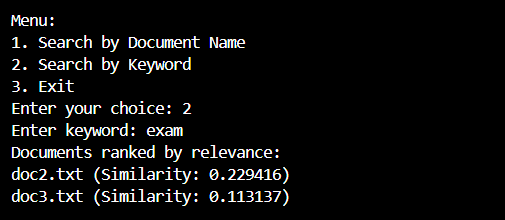
1. A menu-based interface was implemented with the following options:

* Search by document name.
* Search by keywords.
* Exit the program.

# Results/Output







# Conclusion

The implementation of the Generalized Vector Model (GVM) demonstrates the effectiveness of vector-based approaches in Information Retrieval (IR). By representing documents and queries as weighted vectors and leveraging cosine similarity for ranking, the system provides accurate and efficient document retrieval. The project highlights the importance of term weighting and normalization in improving relevance and ranking results. Additionally, integrating user-friendly features such as document name searches and keyword-based queries ensures a practical and interactive search experience. This assignment bridges theoretical IR concepts with real-world applications, offering a solid foundation for exploring advanced retrieval techniques and their applications in modern search engines and information management systems.