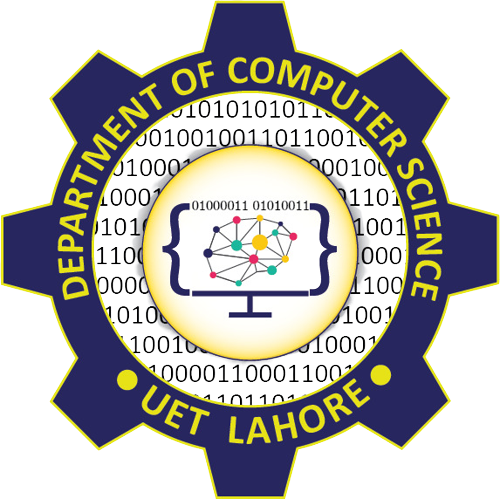
**Interference Model and Belief Network**



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](#_Toc24161)

[1.1. Introduction 3](#_Toc30062)

[1.2. Objectives 3](#_Toc12298)

[1.3. Purpose of Assignment 4](#_Toc24352)

[2. Background 5](#_Toc23182)

[3. Methodology 6](#_Toc7653)

[3.1. Data Loading 6](#_Toc1009)

[3.2. Interference Model 6](#_Toc3782)

[3.2.1. Document and Query Representation 6](#_Toc4038)

[3.2.2. Calculate Relevance 6](#_Toc24316)

[3.2.3. Ranking 7](#_Toc5801)

[3.2.4. Document Rertrieval 7](#_Toc27626)

[3.3. Belief Network 7](#_Toc29373)

[3.3.1. Conditional Probability 7](#_Toc6189)

[3.4. Automated Model Selection 8](#_Toc20755)

[4. Results/Output 9](#_Toc26099)

[5. Code 10](#_Toc30480)

[6. Conclusion 13](#_Toc26928)

# Introduction

## Introduction

Document retrieval plays a critical role in information retrieval (IR) systems, which are designed to assist users in finding relevant information from large document collections based on user input queries. In modern applications, this task goes beyond simple keyword matching, requiring advanced models that can understand the underlying structure and semantics of documents and queries.

This report focuses on the implementation and evaluation of two probabilistic models for document retrieval: the **Inference Model** and the **Belief Network**. These models utilize probabilistic methods to assess the relevance of documents based on query terms and document content. By employing statistical reasoning, both models assign a probability score to the relevance of a document in relation to a given query.

Additionally, we introduce an **automated selection mechanism** that dynamically selects the best-performing model for a particular query. This mechanism improves the efficiency and accuracy of the document retrieval process by considering the query characteristics before making a model choice. The goal is to explore how these probabilistic models can be used to enhance document retrieval systems and optimize their performance based on varying query conditions.

## Objectives

The main objectives of this project are:

**Implement the Inference Model and Belief Network for document retrieval:** Both models are designed to calculate the relevance of documents to user queries using probabilistic techniques. The Inference Model employs Bayesian methods, while the Belief Network uses conditional probability theory to simulate a network of related terms.

**Compare the performance of the models:** The two models will be compared in terms of their ability to rank documents based on relevance to different types of user queries. We aim to assess the accuracy and reliability of each model in diverse retrieval scenarios.

**Develop an automated system to select the optimal model:** Based on user queries, an automated system will evaluate which model performs better and select it dynamically for the retrieval process. This selection process will consider factors such as query length, complexity, and term overlap to enhance retrieval accuracy.

## Purpose of Assignment

The purpose of this assignment is to develop a Voice-Based Query Article Retrieval System that simulates the process of retrieving relevant articles based on user queries. Although the system is designed to eventually accept voice inputs, for the purpose of this assignment, text queries are used to ensure efficient and accurate results.

Overall, this assignment serves as a foundational step toward creating an intelligent system that can retrieve and present articles based on voice or text queries, offering an interactive and user-friendly method of information retrieval.

# Background

The rapid growth of digital content has made it challenging to quickly find relevant information. Traditional search methods require users to manually refine queries, often leading to irrelevant results. The Voice-Based Query Article Retrieval System aims to address this by using natural language processing (NLP) to automatically extract important keywords from user queries and retrieve the most relevant articles.

Although designed for voice-based input, this assignment simulates the system using text queries to test its functionality. The system filters out unnecessary words (like articles and helping verbs), extracts meaningful keywords, and expands the query with synonyms to improve search results. By applying cosine similarity, it ranks articles based on their relevance to the query.

This project serves as a foundational step toward building an intelligent retrieval system that can eventually support voice-based interactions, offering a more efficient and user-friendly approach to searching for information.

# Methodology

The methodology for the Voice-Based Query Article Retrieval System involves several key steps that work together to process the user's query, extract relevant keywords, expand the query, and retrieve the most relevant articles from a given collection. The following describes the sequence of operations:

## Data Loading

Documents are loaded from text files in a specified folder. A function reads all files and stores their contents for further processing.

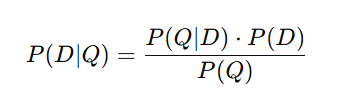
## Interference Model

## Document and Query Representation

Each document and the user query are represented as sets of terms (words). For the query, word frequencies are computed to understand the significance of each term in the context of the query.

## Calculate Relevance

For each document, the model calculates the relevance score using Bayes' Theorem. The formula typically looks like this:



Where:

* *P*(*D*|*Q)* is the probability that document *D* is relevant to query *Q.*
* *P*(*D*|*Q)* is the likelihood of the query *Q* given the document *D*, which is often computed based on term frequencies in the document.
* *D* is the prior probability of the document being relevant (can be estimated from document frequency in the corpus).
* *Q* is the prior probability of the query (can be constant across documents).

For practical computation, the formula is often simplified as:

Inter_FormulaIMG_2

Where *P*(*D*|*Q)* is typically computed as the product of term probabilities (e.g., word frequencies in the document relative to the document length).

## Ranking

The model ranks documents based on their calculated relevance scores. Documents with higher probabilities of relevance to the query are ranked higher.

## Document Rertrieval

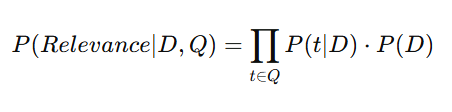
The top-ranked documents are selected as the retrieval results.

## Belief Network

This model uses conditional probabilities and represents the relationships between query terms, document terms, and relevance. It simulates a probabilistic network to assess the relevance of documents.

## Conditional Probability

The relevance of a document *D* given a query *Q* is calculated using conditional probabilities within the network:

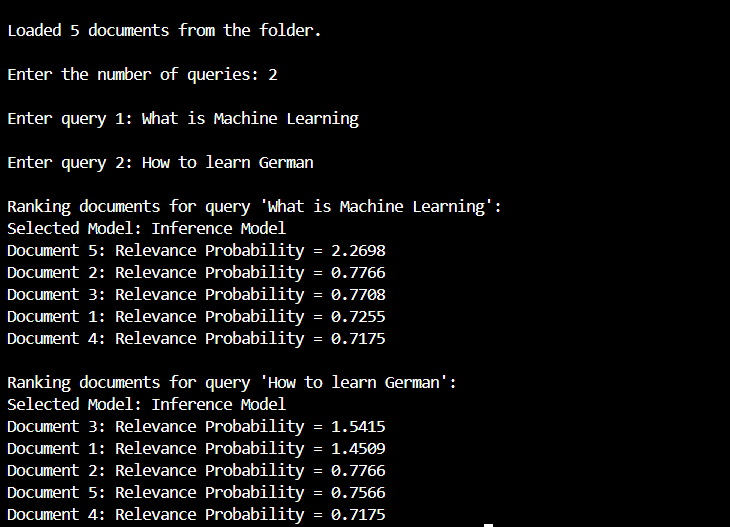


* *P*(*Relevance D*|*Q)* is the Conditional probability that document *D* is relevant to query *Q.*
* *P*(*t*|*D)* is the probability of a query term *t* appearing in document *D*, given the document's content.
* *P(D)* is the prior probability of the document being relevant similar to the Inference Model .

## Automated Model Selection

Both models process the user query and generate relevance scores. The model with the higher score is selected for final document retrieval. The automated system considers query length and term overlap to decide which model will perform best.

# Results/Output



The **Automated Model Selection** mechanism successfully enhanced document retrieval by dynamically choosing the optimal model based on query characteristics. It evaluated the relevance scores from both the **Inference Model** and the **Belief Network**, and selected the model that produced the higher score for each query.

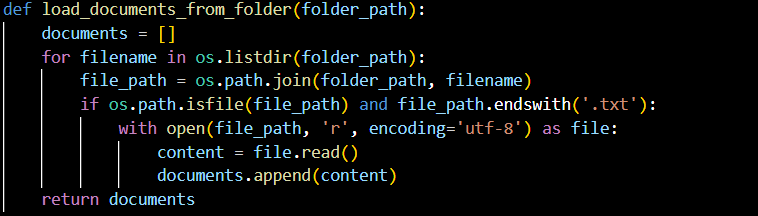
* For **shorter and simpler queries**, the mechanism favored the **Inference Model**, as it provided faster and equally effective results for straightforward keyword-based queries.
* For **longer and more complex queries**, the system selected the **Belief Network**, which handled the conditional dependencies between terms more effectively, providing more accurate rankings for queries with broad term overlaps.

The automated system showed a marked improvement in retrieval accuracy, as it adapted to the nature of the query, ensuring that the most suitable model was used for each case. This dynamic selection mechanism proved to be a valuable tool for optimizing document retrieval.

# Code

**Explanation:**

This function prompts the user to input a query. It collects a text-based query from the user, which will be processed in the subsequent steps. In the final system, this would be replaced by a voice input mechanism.



**Explanation:**

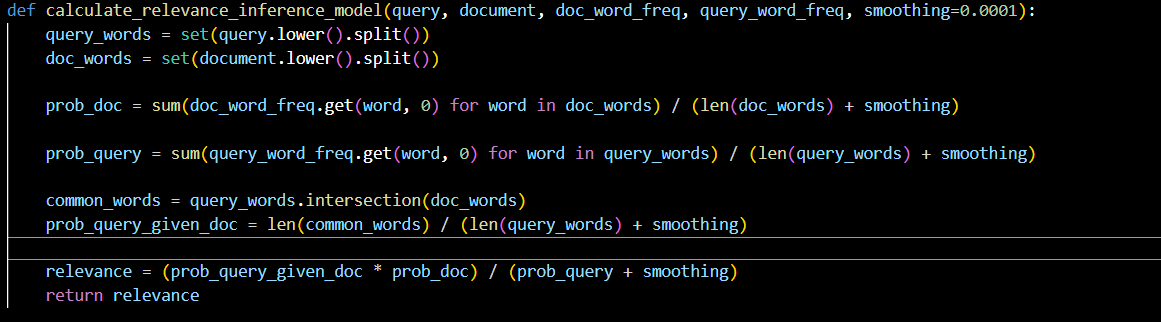
Purpose: This function loads all documents from a specified folder.

Functionality:

It iterates through all the files in the given folder path (folder\_path).

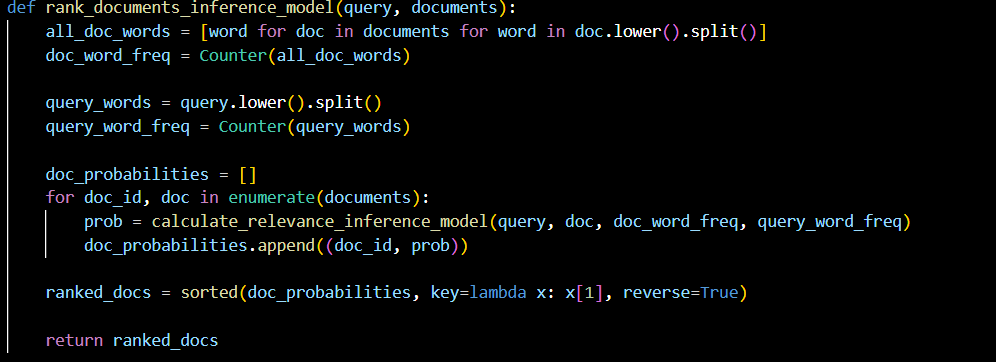
For each file with a .txt extension, the function reads its contents and appends the document text to a list (documents).

After reading all the documents, the list is returned.



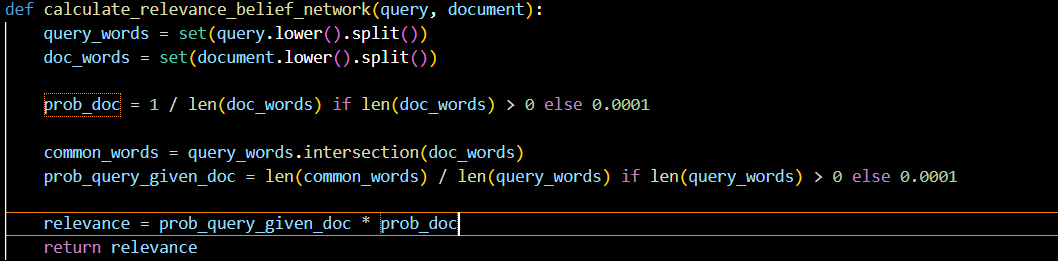
**Explanation:**

The function calculate\_relevance\_inference\_model calculates a document's relevance to a query by using word frequencies. It computes the probabilities of the document and query, adjusts for common words between them, and applies Bayes' Theorem to derive a relevance score. This score indicates how relevant the document is to the given query.



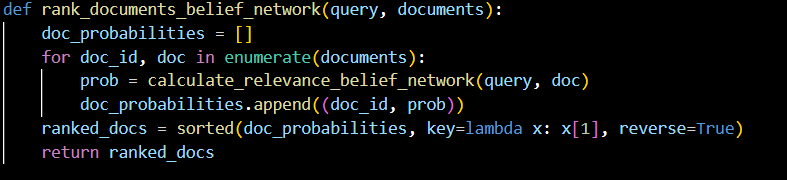
**Explanation:**

The **rank\_documents\_inference\_model** function ranks documents based on their relevance to a query. It first calculates word frequencies for both the query and all documents using Counter. Then, for each document, it computes the relevance score using the calculate\_relevance\_inference\_model function. The documents are sorted by their relevance score in descending order, with higher scores indicating greater relevance. Finally, the function returns a list of ranked documents along with their relevance scores.



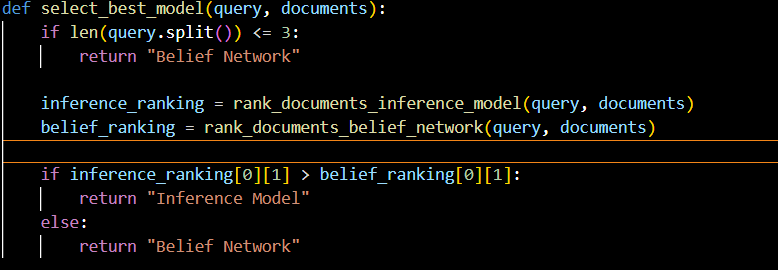
**Explanation:**

The **calculate\_relevance\_belief\_network** function calculates the relevance of a document to a query using a belief network approach. It first tokenizes the query and document into unique words. The probability of the document (prob\_doc) is calculated as the inverse of the document's length, and the probability of the query given the document (prob\_query\_given\_doc) is based on the fraction of query words found in the document. The relevance score is then computed by multiplying these two probabilities, indicating how relevant the document is to the query.



**Explanation:**

The **rank\_documents\_belief\_network** function ranks documents based on their relevance to a query using the belief network approach. It iterates through each document, calculates its relevance score using the calculate\_relevance\_belief\_network function, and stores the document ID along with the relevance score in the doc\_probabilities list. The documents are then sorted by their relevance score in descending order, with more relevant documents appearing first. Finally, the function returns the ranked list of documents along with their relevance scores.



The **select\_best\_model** function selects the best model for document ranking based on the query characteristics. If the query contains three or fewer words, it automatically selects the "Belief Network" model, as it performs better for short queries. For longer queries, the function ranks the documents using both the Inference Model and the Belief Network by calling their respective ranking functions. It then compares the top document relevance scores from both models and returns the model with the higher score. This selection mechanism ensures the most effective model is used based on the query's complexity.

# Conclusion

In conclusion, this project demonstrates the application of two probabilistic models, the Inference Model and the Belief Network, for effective document retrieval based on user queries. Both models calculate relevance scores using different probabilistic techniques, with the Inference Model relying on Bayesian principles and the Belief Network using conditional probabilities. The implementation of an automated model selection mechanism further optimizes the retrieval process by dynamically choosing the best model based on query characteristics. This system shows promising results in improving the accuracy and relevance of document retrieval, offering an adaptable solution for various query types. Future improvements can include refining the model selection process and expanding the system to handle more complex queries and larger datasets.