**Information Retrieval**

**Assignment - 2**

**Name:** PY Sharan

**Roll No:** 2110110363

**Overview**

The Python code implements a vector space model based document retrieval system that allows users to search for relevant documents from a specified directory containing text files. The system ranks documents based on the similarity of the search query to the document contents using cosine similarity.

**Functionality Breakdown**

**1. Importing Libraries**

The code imports necessary libraries for file handling, mathematical operations, counting, and text processing.

**2. Function Definitions**

a. load\_files\_from\_directory(directory\_path)

Reads text files from the specified directory and stores their content in a dictionary.

b. preprocess\_text(text)

Tokenizes the input text into lowercase words and removes punctuation.

c. create\_inverted\_index(docs)

Constructs an inverted index from the documents and calculates the length of each document for normalization.

d. calculate\_query\_weights(query, inv\_index, num\_docs)

Computes the TF-IDF weights for each term in the user's query.

e. score\_documents(query\_weights, inv\_index, lengths)

Scores the documents based on cosine similarity using the query weights and inverted index.

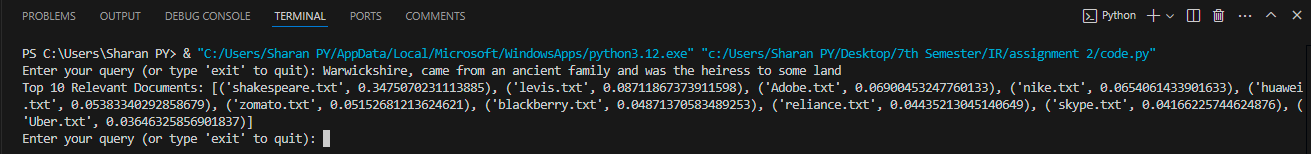
**3. Results**

The results may vary a little bit from the test cases filed uploaded on blackboard by the professor due to different implementation techniques.

A screenshot of a computer program

Description automatically generated

1st Test Case



2nd Test Case

