**DIGITAL ASSIGNMENT – 2**

INFORMATION RETRIEVAL AND ORGANIZATION

VECTOR BASED RETRIEVAL MODEL

NAME: V VARUN ANTO (20MIA1124)

M MADHAVAN (20MIA1113)

**DOCUMENTATION:**

**Application Description:**

The Maha Shivratri search engine is built using a Vector Space Model (VSM) for information retrieval.The model extracts and lemmatizes the input documents, calculates the term frequency (TF) and  the inverse document frequency (IDF) of each vocabulary word, and calculates the cosine similarity  and the Jaccard similarity between the query and each document of the data set. Returns the top 2  documents with the highest output similarity scores.

**Dataset:**

The dataset used in this project contains a four document with information about Maha Shivratri.

**Preprocessing:**

The following preprocessing steps are performed on the input document and query.

**Tokenization:**

The input text is split into individual words or tokens.

**Lemmatization:**

The tokens are converted to their base form using the WordNetLemmatizer from the nltk library.

**Vocabulary Terms:**

Vocabulary term are identified where it used to find unique terms from the documents and query to avoid duplication of words.

**Term Frequency (TF):**

TF is calculated as the frequency of a term in a document divided by the total number of terms in the document.

**Inverse Document Frequency (IDF):**

IDF is calculated as the logarithm of the total number of documents in the dataset divided by the number of documents containing the term.

**Cosine Similarity:**

Cosine similarity measures the cosine of the angle between two vectors, where each vector represents a document or query in a dataset. It is calculated as the scalar product of the document vector and the request vector divided by the product of their magnitudes.

**Jaccard Similarity:**

Jaccard similarity measures the intersection of two sets divided by the union of the sets. In this case, it measures the overlap of the query and document in terms of their tokens.

**Packages used:**

nltk, word\_tokenize, WordNetLemmatizer / nltk package is imported for tokenization and lemmatization

**Function defined:**

1.cosine\_similarities function.

2.jaccad \_similarities function.