Team Name: Group-6

Members: Chenchu Aravind 20MIA1126

Shiva Sindhu Perla 20MIA1104

Information Retrieval and Organization

Digital Assignment-2

Vector space model-Documentation

Github: Link

Application Description:

The vector space model is a mathematical model used in information retrieval (IR) to represent documents and queries as vectors in a high-dimensional space. In this model, each document or query is represented as a vector of weights, with each weight corresponding to a term in a vocabulary. In this project a search engine for “Maha Sivaratri” is built using Vector Space Model using Python.

Dataset:

For this project few text about “Maha Sivaratri” is fed into the model as .txt documents.

Pre-processing:

- Special characters are removed using the function “remove_special_characters”
- Digits are removed using the function “remove_function”.
- Next the document is tokenized. Where the text in the document is breaked into smaller parts called tokens
- The running time for the retrieval is 34Seconds

Packages:

- Nltk
  - Stopwords
  - Punctuation
- Re (Regular Expression)
- Sys
- Collections
- Math
- Glob

Functions :

- Intialize_document_frequencies()

    - Returns the number of times a term appears in the document
- Intialize_lengths()
    - Computes the length for each document
- Term_frequency()
    - Returns the term frequency of term in document id
- inverse_document_frequency
    - Return the inverse document frequency.
- Similarity()
    - Returns the cosine similaritybetween the query and document id

Output:

The model takes the query from the user and retrieves the document based on the cosine similarity score in descending order.