INFORMATION RETRIEVAL AND ORGANIZATION
DIGITAL ASSIGNMENT – 2

VECTOR SPACE RETRIEVAL MODEL

SRIKAR ALLADI- 19MIA1103
RIYAN IMMACULATE-19MIA1024

DESCRIPTION

A popular information retrieval paradigm called the Vector Space Model (VSM) portrays documents as vectors in a high-dimensional space where each dimension is associated with a word from the lexicon. The VSM is predicated on the idea that

The distribution of a document's terms can be used to infer its meaning, and distributions of terms in documents with related content will be comparable.
To begin using the VSM, a group of documents must first be tokenized, stemmed, and stop words removed. The next step is to build a term-document matrix, where each row corresponds to a term and each column to a document. Each term's frequency in each document, or a variation of it, is contained in the matrix (e.g., term frequency-inverse document frequency, TF-IDF).

In the same space as the documents, the query is also pre-processed and represented as a vector. Finally, using a cosine similarity metric, a similarity score is calculated between each document vector and the query vector. The top-ranked papers are returned as the search results after documents are rated according to how similar they are to the query.
The VSM has various benefits, including its ease of use, potency, and capacity for handling enormous document collections. It does, however, have certain drawbacks, such as the "bag of words" assumption that ignores word order and context and the term scarcity issue, when numerous terms are found in a small number of documents. More complex models that take into account the semantic links between words and documents, including probabilistic models or neural models, can be used to overcome these restrictions.

.

DATASET

Every year Maha Shivratri is celebrated with a lot of pomp and grandeur. It is considered to be a very special time of
the year since millions of people celebrate this momentous occasion with a lot of fervour and glee.
Lord Shiva devotees celebrate this occasion with a lot of grandness. It is accompanied by folk dances, songs,
prayers, chants, mantras etc. This year, the beautiful occasion of Maha Shivratri will be celebrated on February 18.
People keep a fast on this Maha shivratri, stay awake at night and pray to the lord for blessings, happiness, hope and

prosperity. This festival holds a lot of significance and is considered to be one of the most important festivals in
India.
The festival of Maha Shivratri will be celebrated on February 18 and is a very auspicious festival. This Hindu
festival celebrates the power of Lord Shiva. Lord Shiva protects his devotees from negative and evil spirits. He is the
epitome of powerful and auspicious energy

PREPROCESSING

The steps involved in preprocessing are -

1. Tokenization: The first step is to split the text into individual words or tokens. This can be done by using whitespace as a delimiter, but more sophisticated methods such as regular expressions and natural language processing tools can also be used to handle punctuation, abbreviations, and other variations in text.
2. Lowercasing: Convert all the tokens to lowercase so that the same word with different cases is not treated as different words.
3. Stop word Removal: Words such as "and", "the", "is", etc. occur frequently in text but do not carry much meaning. These words are called stop words and are typically removed to reduce noise in the data.
4. Stemming: Words such as "running", "runs", and "ran" all have the same root word "run". Stemming is the process of reducing words to their root form so that variations of the same word are treated as a single term.
5. Lemmatization: Unlike stemming, lemmatization involves reducing words to their base or dictionary form, rather than just stripping off suffixes. For example, "am", "is", and "are" all get converted to "be".
6. Term Frequency (TF) and Inverse Document Frequency (IDF): Finally, after pre-processing, the term frequency (TF) and inverse document frequency (IDF) are calculated for each term in the corpus. TF measures the number of times a term appears in a document, while IDF measures the rarity of the term across the entire corpus. These values are used to weigh the importance of each term in a document and across the collection.

PACKAGES USED

1) **Math**:

The Python math module provides functions that are useful in number theory as well as in **representation theory**, a related field. These functions allow you to calculate a range of important values, including the following:

- The **factorials** of a number
- The **greatest common divisor** of two numbers
- The sum of **iterables**

2) **nltk**:

NLTK is **a standard python library that provides a set of diverse algorithms for NLP**. It is one of the most used libraries for NLP and Computational Linguistics

3) **punkt**:

This tokenizer divides a text into a list of sentences by using an unsupervised algorithm to build a model for abbreviation words, collocations, and words that start sentences. It must be trained on a large collection of plaintext in the target language before it can be used.

4) **Wordnet :**

WordNet is a lexical database of English. Using synsets, **helps find conceptual relationships between words such as hypernyms, hyponyms, synonyms, antonyms etc**. The lexical entry for a single morphological form of a sense-disambiguated word

5) Omw -1.4:

Open Multilingual Wordnet. OMW stands for "on my way." It's an quick and informal way to let someone know to expect your arrival.

6) Word_tokenize:

word_tokenize is a function in Python that splits a given sentence into words using the NLTK library.

7) **WordNetLemmatizer:**

Wordnet Lemmatizer with NLTK. Wordnet is an large, freely and publicly available lexical database for the English language aiming to **establish structured semantic relationships between words**. It offers lemmatization capabilities as well and is one of the earliest and most commonly used lemmatizers.

Formulas

Calculating term frequency :

TF:

TF of a term or word is the number of times the term appears in a document compared to the total number of words in the document

$$TF = \frac{\text{number of times the term appears in the document}}{\text{total number of terms in the document}}$$

IDF:

IDF of a term reflects the proportion of documents in the corpus that contain the term. Words unique to a small percentage of documents (e.g., technical jargon terms) receive higher importance values than words common across all documents (e.g., a, the, and).

$$IDF = log(\frac{\text{number of the documents in the corpus}}{\text{number of documents in the corpus contain the term}})$$

Cosine similarities:

Cosine similarity is a measure of similarity, often used to measure document similarity in text analysis. We use the below formula to compute the cosine similarity.

```
Similarity = (A.B) / (||A||.||B||)
```

Jaccard Similarity :

Jaccard Similarity is a common proximity measurement used to compute the similarity between two objects, such as two text documents.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$