Vidyavardhini's College of Engineering and Technology

Department of Artificial Intelligence & Data Science

AY:2024-25

Class:	**BE**	**Semester:**	**VII**
Course Code:	**CSDOL7011**	**Course Name:**	**Natural Language Processing**

Name of Student:	**Parth Raut**
Roll No.:	**40**
Experiment No.:	**8**
Title of the Experiment:	**Text Similarity Recognition for Chosen Documents**
Date of Performance:	
Date of Submission:	

Evaluation

Performance Indicator	Max. Marks	Marks Obtained
Performance	5	
Understanding	5	
Journal work and timely submission	10	
Total	20	

Performance Indicator	Exceed Expectations (EE)	Meet Expectations (ME)	Below Expectations (BE)
Performance	4-5	2-3	1
Understanding	4-5	2-3	1
Journal work and timely submission	8-10	5-8	1-4

Checked by

Name of Faculty :

Signature :

Date :

Experiment 8

Aim: Implement Text Similarity Recognizer for the chosen text documents.

Objective: Understand the importance of Implementing Text Similarity Recognizer for the chosen text documents.

Theory:

1. **Preprocess the Text Data:**
    - Tokenization
    - Stopwords removal
    - Lemmatization or stemming
2. **Feature Extraction:**
    - TF-IDF Vectorization
    - Word Embeddings (e.g., Word2Vec, GloVe)
    - Sentence Embeddings (e.g., Sentence-BERT)
3. **Compute Similarity:**
    - Cosine Similarity
    - Euclidean Distance
4. **Evaluate Similarity:**
    - Compare and interpret the similarity scores.

Code:

In [1]:
```
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
```

In [2]:
```
sentenceOne = 'My house is empty today'
sentenceTwo = 'Nobody is at my home'
documents = [sentenceOne, sentenceTwo]
```

In [3]:
```
tfidf = TfidfVectorizer ()
sparseMatrix = tfidf.fit_transform (documents)
```

In [4]:
```
docTermMatrix = sparseMatrix.todense ()
```

In [5]:
```
df = pd.DataFrame (
    docTermMatrix,
    columns = tfidf.get_feature_names_out (),
    index = ['sentenceOne', 'sentenceTwo']
)
```

In [7]:
```
simScore = cosine_similarity (df, df)[0, 1]
match_keys = df.isin ([0]).sum (axis = 0)
match_words = match_keys[match_keys.values == 0].keys ()
```

In [8]:
```
print (f'Cosine Similarity : {round (simScore, 2)}')
print (f'Matching Words : {list (match_words)}')
```

```
Cosine Similarity : 0.25
Matching Words : ['is', 'my']
```

Conclusion : Implementing a Text Similarity Recognizer involves several crucial steps: preprocessing the text data, extracting meaningful features, computing similarity measures, and interpreting the results. In this example, we utilized TF-IDF Vectorization for feature extraction and Cosine Similarity for computing similarity between text documents.