Vidyavardhini's College of Engineering and Technology

Department of Artificial Intelligence & Data Science

AY: 2025-26

Class:	**BE-AI&DS**	**Semester:**	**VII**
Course Code:	**CSDOL7011**	**Course Name:**	**NLP Lab**

Name of Student:	BARI ANKIT VINOD
Roll No. :	61
Experiment No.:	**3**
Title of the Experiment:	**Generating Word Embeddings using Word2Vec for Text Similarity Analysis**
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 :

Aim: To train a Word2Vec model on a corpus and generate vector representations (embeddings) of words for analyzing semantic similarity.

Objective: To generate dense word vector representations using Word2Vec for capturing semantic similarity.

Tools Required:

- Python (Jupyter Notebook or Google Colab)
- gensim library
- nltk (for corpus and preprocessing)
- matplotlib and sklearn.manifold (for visualization)

Procedure:

1. Import necessary libraries:
    a. gensim.models.Word2Vec, nltk, matplotlib, sklearn.manifold.TSNE, re

2. Prepare the corpus:
    a. Use a sample text corpus (e.g., NLTK's Gutenberg corpus or a custom text file).
    b. Preprocess text: lowercase, remove special characters, tokenize, remove stop words.

3. Train Word2Vec:
    a. Use gensim.models.Word2Vec with parameters:
        i. vector_size=100
        ii. window=5
        iii. min_count=2

iv. workers=4
v. sg=0 (CBOW) or sg=1 (Skip-gram)

4. Explore embeddings:
    a. Find most similar words to a given word using .most_similar().
    b. Check similarity scores between two words using .similarity().

5. Visualize word vectors:

Use t-SNE to reduce dimensions and plot selected word vectors in 2D.

Description of the Experiment:

In this experiment, students generate distributed representations of words—known as embeddings—using the Word2Vec model. These embeddings capture semantic relationships between words, allowing us to measure how similar two words are in meaning. This method overcomes the limitations of sparse representations like BoW.

Detailed Description of the NLP Technique:

Word Embeddings:

Word embeddings are dense vector representations of words in a continuous vector space, where semantically similar words are mapped closer together.

Word2Vec:

Word2Vec is a popular algorithm introduced by Google that uses shallow neural networks to learn word representations from large text corpora. It has two architectures:

1. CBOW (Continuous Bag of Words):
    a. Predicts the current word using surrounding context words.
    b. Faster and better for large datasets.

2. Skip-Gram:

a. Predicts surrounding words given the current word.
b. Performs better for smaller datasets and rare words.

3. Properties of Word2Vec embeddings:
    a. Captures semantic relationships (e.g., king – man + woman ≈ queen).
    b. Enables similarity comparison between words using cosine similarity.
    c. Reduces dimensionality while retaining semantic meaning.

4. t-SNE Visualization:
    a. t-Distributed Stochastic Neighbor Embedding (t-SNE) is used to reduce high-dimensional vectors into 2D for visualization.
    b. It helps visualize word clusters and relationships learned by Word2Vec.

Conclusion:

The results from the Word2Vec model demonstrate that it effectively captures the semantic and syntactic relationships between words. Words with similar meanings or used in similar contexts appear closer together in the vector space. The embeddings successfully reflect meaningful analogies such as king – man + woman ≈ queen, showing the model's ability to understand linguistic relationships.

The t-SNE visualization further confirms these findings by displaying clear clusters of related words. Words belonging to similar categories or contexts are grouped together, indicating that Word2Vec embeddings preserve the semantic structure of language even after dimensionality reduction.

Overall, the results highlight the strength of Word2Vec in generating dense and meaningful word representations and demonstrate that visualization techniques like t-SNE are valuable tools for interpreting and validating embedding models.

CSDOL7011: Natural Language Processing Lab