Exp 5 : Implement the word2Vec model python program

Code :

import numpy as np

from collections import defaultdict

import random

corpus = "the quick brown fox jumps over the lazy dog"

window\_size = 2

embedding\_dim = 10

learning\_rate = 0.01

epochs = 1000

def preprocess(corpus):

tokens = corpus.lower().split()

vocab = set(tokens)

word2idx = {word: idx for idx, word in enumerate(vocab)}

idx2word = {idx: word for word, idx in word2idx.items()}

return tokens, word2idx, idx2word

tokens, word2idx, idx2word = preprocess(corpus)

vocab\_size = len(word2idx)

def generate\_training\_data(tokens, word2idx, window\_size):

training\_data = []

for idx, center\_word in enumerate(tokens):

for w in range(-window\_size, window\_size + 1):

context\_idx = idx + w

if context\_idx < 0 or context\_idx >= len(tokens) or context\_idx == idx:

continue

context\_word = tokens[context\_idx]

training\_data.append((word2idx[center\_word], word2idx[context\_word]))

return training\_data

training\_data = generate\_training\_data(tokens, word2idx, window\_size)

W1 = np.random.rand(vocab\_size, embedding\_dim)

W2 = np.random.rand(embedding\_dim, vocab\_size)

def one\_hot\_vector(word\_idx, vocab\_size):

vec = np.zeros(vocab\_size)

vec[word\_idx] = 1

return vec

def softmax(x):

e\_x = np.exp(x - np.max(x))

return e\_x / e\_x.sum()

for epoch in range(epochs):

loss = 0

for center\_idx, context\_idx in training\_data:

x = one\_hot\_vector(center\_idx, vocab\_size)

h = np.dot(W1.T, x)

u = np.dot(W2.T, h)

y\_pred = softmax(u)

y\_true = one\_hot\_vector(context\_idx, vocab\_size)

error = y\_pred - y\_true

dW2 = np.outer(h, error)

dW1 = np.outer(x, np.dot(W2, error))

W1 -= learning\_rate \* dW1

W2 -= learning\_rate \* dW2

loss += -np.log(y\_pred[context\_idx] + 1e-9)

if epoch % 100 == 0:

print(f"Epoch {epoch}, Loss: {loss:.4f}")

def get\_embedding(word):

idx = word2idx[word]

return W1[idx]

print("\nWord Embedding for 'fox':\n", get\_embedding("fox"))

Output :

Epoch 0, Loss: 66.1045

Epoch 100, Loss: 48.5471

Epoch 200, Loss: 43.8555

Epoch 300, Loss: 42.9267

Epoch 400, Loss: 42.6848

Epoch 500, Loss: 42.5993

Epoch 600, Loss: 42.5628

Epoch 700, Loss: 42.5450

Epoch 800, Loss: 42.5352

Epoch 900, Loss: 42.5289

Word Embedding for 'fox':

[ 0.54204138 2.21468158 1.14436996 -0.55893533 -0.23553762 0.82091689

0.04969511 -0.18002016 0.39034785 0.70524861]