README
Project Overview

This project performs document embedding and clustering using two different methods:

1. **Doc2Vec** - Generates document embeddings using three different configurations and clusters them using K-Means.
2. **Word2Vec with Bag-of-Bins -** Clusters words into bins and uses word frequencies to generate document vectors, followed by clustering using K-Means.

The effectiveness of the clustering is evaluated using **Silhouette Scores (cosine distance)**, and results are visualized through **bar charts and PCA scatter plots**.

Prerequisites

Ensure you have the following dependencies installed:

```
pip install pandas numpy nltk gensim scikit-learn matplotlib
```

Files

- lab8.py: The main Python script for embedding generation, clustering, and evaluation.
- reddit_posts.csv: The dataset used for training embeddings and clustering (must be placed in the same directory).

Running the Code

Run the script using : python lab8.py

Ensure that reddit_posts.csv is in the same directory before execution.