Team Members: Jaival, Mayank, Pratham

Course: DSCI-560: Data Science Practicum
Assignment: Lab 8 - Representing Document Concepts with Embeddings
Github Link - https://github.com/jaivalupadhyay/dsci-560.git

1. Objective

This lab explores **Doc2Vec** and **Word2Vec Bin-based embeddings** for representing Reddit posts as numerical vectors. We perform clustering on these vectors and compare their quality using **silhouette scores** and **PCA visualizations**.

The goal is to evaluate how different embedding methods affect clustering performance and determine the most effective configuration.

2. Dataset and Preprocessing

- The dataset consists of **Reddit posts** stored in reddit_posts.csv.
- Preprocessing steps:
    - Convert text to lowercase
    - Remove punctuation
    - Tokenize using nltk.word_tokenize

Each post is then transformed into a numerical vector using **Doc2Vec** and **Word2Vec Bin-based embeddings**.

3. Implementation Details

3.1 Doc2Vec Embeddings & Clustering

- 9We experiment with **three vector sizes: 50, 100, 200**.
- Each document is converted into a vector using Doc2Vec.
- Clustering is performed using **K-Means (k=3)**.
- **Silhouette scores** (cosine metric) are calculated to evaluate clustering quality.

3.2 Word2Vec Bin Embeddings & Clustering

- **Words** from all posts are clustered into **10 bins** using Word2Vec (with vector sizes 50, 100, 200).
- Each document is then represented as a **normalized frequency vector** of word bins.

- Clustering and silhouette scoring are performed similarly to Doc2Vec.

4. Results & Discussion

4.1 Silhouette Scores

The following silhouette scores were obtained:

Model	50-D	100-D	200-D
Doc2Vec	-0.35	-0.34	-0.37
Word2Vec Bin	0.18	0.10	0.10

- **Doc2Vec:** Scores are **negative**, suggesting poor clustering.
- **Word2Vec Bin:** Scores are **positive**, indicating better-defined clusters, but still relatively low.

4.2 PCA Scatter Plots

We generated **2D PCA projections** for dimension = 100. Observations:

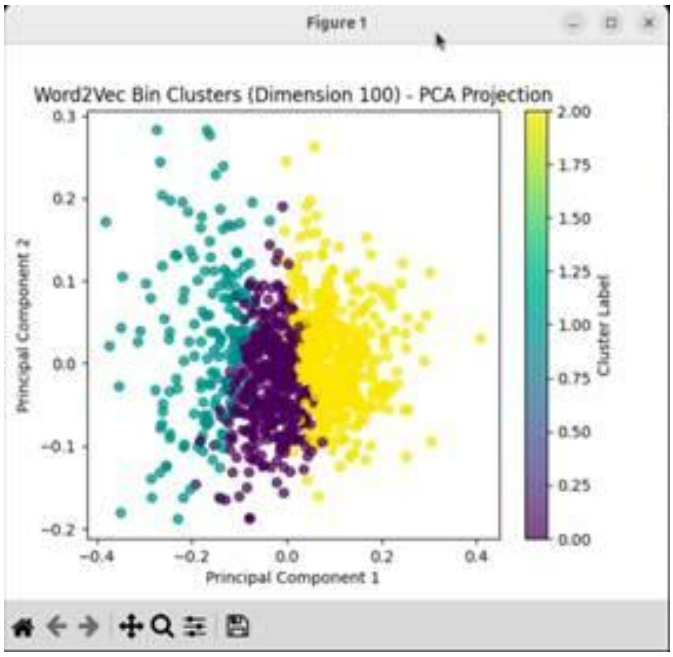
- Doc2Vec: Clusters are more spread out with overlap.
- Word2Vec Bin: Clusters are slightly more distinct but still mixed.

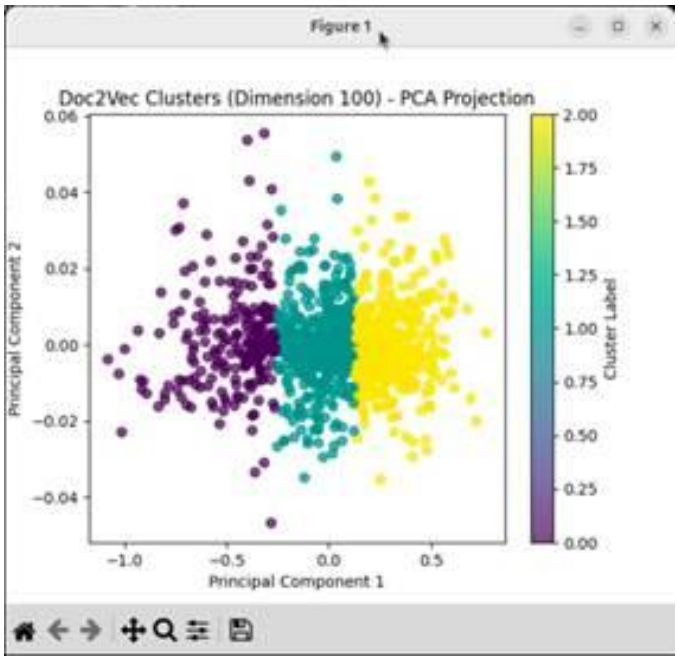
5. Conclusion

- **Word2Vec Bin** performed **better** than **Doc2Vec**, showing positive silhouette scores.
- **Doc2Vec clusters overlap significantly**, leading to poor performance.
- **Future improvements** can involve tuning hyperparameters, increasing data, or testing different distance metrics.

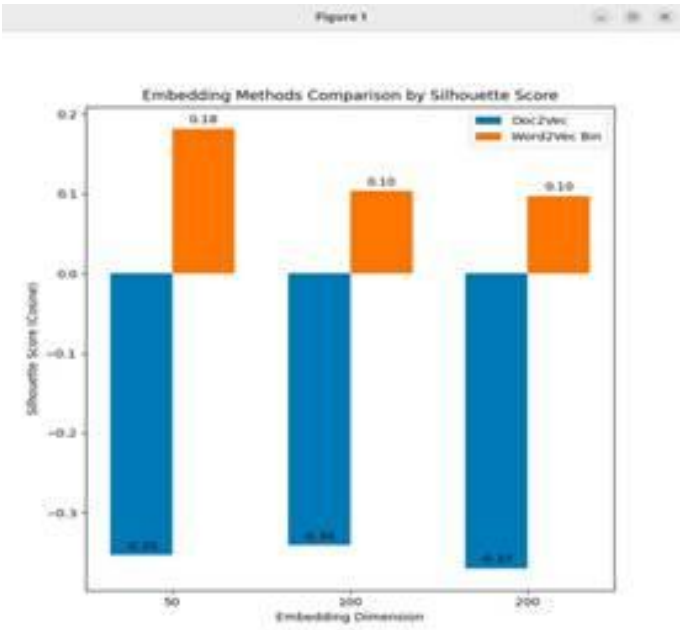
6. Output

1. PCA Scatter Plots (Doc2Vec & Word2Vec Bin)





2. Silhouette Score Comparison Bar Chart



Code Explanation

1. Imports and Basic Setup

- It imports necessary libraries like pandas, numpy, nltk, gensim, sklearn, and matplotlib.
- Downloads NLTK's tokenizer package for text processing.

2. Load and Preprocess Data

- Loads **Reddit posts dataset (reddit_posts.csv)**.
- Cleans and tokenizes the text using:
    - Converts text to lowercase.
    - Removes punctuation using regex.
    - Tokenizes words using word_tokenize.

3. Doc2Vec Embeddings & Clustering

- **Creates Doc2Vec embeddings with 3 configurations:**

    - Vector sizes of **50, 100, and 200**.
- Converts each post into TaggedDocument format for training.
- Trains **three Doc2Vec models** and infers document vectors.
- Clusters documents using **KMeans (3 clusters)**.
- Evaluates clustering quality using **Silhouette Score (cosine distance)**.

Output:

- Prints silhouette scores for each Doc2Vec configuration.

4. Word2Vec + Bag-of-Bins Clustering

- Trains **Word2Vec models** (50, 100, 200 dimensions).
- Extracts word embeddings from Word2Vec.
- Uses **KMeans (10 bins)** to group words into clusters (bins).
- Converts documents into **word-bin frequency vectors**.
- Clusters document vectors using **KMeans (3 clusters)**.
- Computes **Silhouette Score** for Word2Vec-based clusters.

Output:

- Prints silhouette scores for Word2Vec-based embeddings.

5. Visualization: Silhouette Score Bar Chart

- **Compares Doc2Vec vs. Word2Vec silhouette scores** using a **bar chart**.

6. PCA Scatter Plot for Clusters

- Uses **PCA (Principal Component Analysis)** to reduce embeddings to **2D**.
- **Visualizes clusters** for **Doc2Vec and Word2Vec** embeddings.

Output:

- **Two scatter plots** showing clustering of documents.