**Institute of Business Administration**

**Introduction to Text Analytics**

**Assignment 03 – Assessment**

**Name: ZUHA AQIB ID: 26106**

Report each experiment’s detail and scores for k = 5, 9, and 13. You are required to perform ten experiments for each ‘k’ (number of clusters). Please set random seed value to your ERP ID for each K-Means clustering experiment.

\*The first two entries in the table are provided for reference only. Hence, the scores do not interpret anything and have been entered randomly. Replace these entries while submitting.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **k (Number of clusters)** | **Vectorizer Type and Details** | **vector\_size** | **window** | **Epochs Count** | **CBoW/Skipgram**  **OR**  **DM/DBoW** | **Silhouette Score** | **WSS**  **Score** |
| **5** | Word2Vec | 300 | 10 | 150 | CBoW | 0.568 | 10.25 |
| Doc2Vec | 100 | 15 | 100 | DBoW | 0.471 | 11.56 |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| **9** |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| **13** |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

**Analysis & Interpretation:**

* **Identify which embedding technique resulted in the best clustering.**
* **Discuss how different choices of hyperparameters impacted the results.**
* **Compare the performance of word2vec and doc2vec embeddings with those used in previous assignment (Assignment 02)**

|  |
| --- |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |