**Institute of Business Administration**

**Introduction to Text Analytics**

**Assignment 02 – K-Means Clustering Assessment**

**Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ID: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

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 four 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** | **Stemming (Yes/No)** | **Lemmatization (Yes/No)** | **N-Grams Utilized** | **Stop words (Yes/No)** | **Silhouette Score** | **WSS**  **Score** |
| **5** | CountVectorizer (with term presence) | Yes | Yes | Unigrams | Yes | 0.568 | 10.25 |
| TFIDF | No | No | Bigrams | No | 0.471 | 11.56 |
| LSA/SVD Embeddings (n\_elements = 10) | Yes | Yes | Unigrams | Yes | 0.660 | 9.568 |
| TruncatedSVD (n\_components = 10) | No | No | Bigrams | No | 0.652 | 8.562 |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| **9** |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| **13** |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

**Analysis & Interpretation:**

**o Identify which embedding technique resulted in the best clustering.**

**o Discuss how preprocessing choices impacted the results.**

**o Provide sample headlines from different clusters to analyze coherence.**

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