**Text Analytics**

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

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.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **k (Number of clusters)** | **Vectorizer Type and Details** | **vector\_size** | **window** | **Epochs Count** | **CBoW/Skipgram**  **OR**  **DM/DBoW** | **Silhouette Score** | **WSS**  **Score** |
| **5** | Word2Vec | 100 | 5 | 20 | CBoW | 0.504 | 4.806 |
| Doc2Vec | 100 | 5 | 20 | DBoW | 0.018 | 1.530 |
| Word2Vec | 100 | 5 | 20 | Skipgram | 0.267 | 0.843 |
| Doc2Vec | 100 | 5 | 20 | DM | 0.134 | 1.672 |
| Word2Vec | 200 | 20 | 50 | CBoW | 0.319 | 1.266 |
| Doc2Vec | 200 | 20 | 50 | DBoW | 0.439 | 2.976 |
| Word2Vec | 200 | 20 | 50 | Skipgram | 0.163 | 3.856 |
| Doc2Vec | 200 | 20 | 50 | DM | 0.384 | 1.809 |
| Word2Vec | 500 | 50 | 100 | CBoW | 0.194 | 9.087 |
| Doc2Vec | 500 | 50 | 100 | DBoW | 0.359 | 3.218 |
| Word2Vec | 500 | 50 | 100 | Skipgram | 0.172 | 183.075 |
| Doc2Vec | 500 | 50 | 100 | DM | 0.358 | 2.885 |
| **9** | Word2Vec | 100 | 5 | 20 | CBoW | 0.445 | 1.66 |
| Doc2Vec | 100 | 5 | 20 | DBoW | 0.012 | 1.496 |
| Word2Vec | 100 | 5 | 20 | Skipgram | 0.162 | 0.604 |
| Doc2Vec | 100 | 5 | 20 | DM | 0.055 | 1.540 |
| Word2Vec | 200 | 20 | 50 | CBoW | 0.228 | 0.732 |
| Doc2Vec | 200 | 20 | 50 | DBoW | 0.328 | 1.519 |
| Word2Vec | 200 | 20 | 50 | Skipgram | 0.196 | 2.675 |
| Doc2Vec | 200 | 20 | 50 | DM | 0.258 | 1.157 |
| Word2Vec | 200 | 20 | 50 | CBoW | 0.179 | 6.196 |
| Doc2Vec | 500 | 50 | 100 | DBoW | 0.271 | 1.941 |
| Word2Vec | 500 | 50 | 100 | Skipgram | 0.157 | 138.032 |
| Doc2Vec | 500 | 50 | 100 | DM | 0.260 | 1.888 |
| **13** | Word2Vec | 500 | 50 | 100 | CBoW | 0.378 | 0.951 |
| Doc2Vec | 100 | 5 | 20 | DBoW | 0.011 | 1.465 |
| Word2Vec | 100 | 5 | 20 | Skipgram | 0.121 | 0.541 |
| Doc2Vec | 100 | 5 | 20 | DM | 0.031 | 1.503 |
| Word2Vec | 200 | 20 | 50 | CBoW | 0.169 | 0.619 |
| Doc2Vec | 200 | 20 | 50 | DBoW | 0.243 | 1.169 |
| Word2Vec | 200 | 20 | 50 | Skipgram | 0.169 | 2.203 |
| Doc2Vec | 200 | 20 | 50 | DM | 0.177 | 0.976 |
| Word2Vec | 500 | 50 | 100 | CBoW | 0.174 | 5.024 |
| Doc2Vec | 500 | 50 | 100 | DBoW | 0.199 | 1.652 |
| Word2Vec | 500 | 50 | 100 | Skipgram | 0.152 | 114.628 |
| Doc2Vec | 500 | 50 | 100 | DM | 0.217 | 1.574 |

**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)**

|  |
| --- |
|  |

**1. Identifying the Best Clustering Approach**

To determine the most effective clustering approach, we evaluate two metrics:

* Silhouette Score: Measures how well clusters are separated (higher is better).
* WSS Score (Within-Cluster Sum of Squares): Measures compactness of clusters (lower is better).

**1.1. Performance of Traditional Vectorization Methods (Assignment 2)**

* TF-IDF and Count Vectorizer generally perform poorly, achieving very low Silhouette Scores (close to 0), indicating weak clustering.
* LSA/SVD-based embeddings performed slightly better, particularly when fewer components were used (n\_components = 5, 10, 25). However, increasing components led to diminishing returns or overfitting, which impacted clustering effectiveness.
* Best Results: LSA/SVD with n\_components = 5 or 10 and bigrams resulted in the best Silhouette Scores (0.48-0.76) and lower WSS (indicating more compact clusters). However, the scores remain relatively low.

**1.2. Performance of Word2Vec & Doc2Vec (Assignment 3)**

* Word2Vec (CBOW & Skipgram) and Doc2Vec (DM & DBOW) outperform traditional methods significantly in terms of Silhouette Scores.
* The best clustering was achieved with:
  + Word2Vec (CBOW, vector\_size=500, window=50, epochs=100) → Silhouette Score 0.378, WSS 0.951.
  + Doc2Vec (DBOW, vector\_size=500, window=50, epochs=100) → Silhouette Score 0.359, WSS 1.941.
  + Doc2Vec (DM, vector\_size=500, window=50, epochs=100) → Silhouette Score 0.260, WSS 1.888.

**1.3. Comparing Traditional vs. Word2Vec/Doc2Vec Approaches**

|  |  |  |  |
| --- | --- | --- | --- |
| **Approach** | **Best Silhouette Score** | **Best WSS Score** | **Observations** |
| |  | | --- | | TF-IDF |  |  | | --- | |  | | |  | | --- | | 0.0034 |  |  | | --- | |  | | 437.88 | Poor separation, high WSS |
| |  | | --- | | LSA/SVD (best case) |  |  | | --- | |  | | 0.76 | 0.52 | |  | | --- | | Some improvement, but still weak clustering |  |  | | --- | |  | |
| |  | | --- | | Word2Vec (best case) |  |  | | --- | |  | | 0.504 | 4.806 | |  | | --- | | Much better clustering compared to TF-IDF/SVD |  |  | | --- | |  | |
| |  | | --- | | Doc2Vec (best case) |  |  | | --- | |  | | 0.384 | 1.809 | Strong clustering, better WSS than Word2Vec |

**Conclusion**:  
Word2Vec and Doc2Vec embeddings clearly outperform traditional methods (TF-IDF, LSA, SVD) in clustering effectiveness. Deep embeddings capture semantic meaning better, leading to more meaningful clusters.

**2. Impact of Hyperparameters on Clustering Performance**

The choice of vector size, window size, number of epochs, and model type (CBOW, Skipgram, DM, DBOW) affects clustering results significantly.

**2.1. Vector Size**

* Increasing vector size (100 → 200 → 500) generally improves clustering quality.
  + Example: For k=5, increasing vector size from 100 to 500 improved Silhouette Score from 0.134 → 0.384 (Doc2Vec).
  + Larger vector sizes capture more semantic meaning, leading to better separability of clusters.

**2.2. Window Size**

* Increasing window size (5 → 20 → 50) leads to better performance in both Word2Vec and Doc2Vec.
  + Example: At k=9, window=50 with vector\_size=500 achieved Silhouette Score = 0.271, while window=5 had 0.012.
  + Larger window sizes allow the model to capture broader contextual relationships in text.

**2.3. Number of Epochs**

* Increasing epochs (20 → 50 → 100) improves clustering by allowing embeddings to converge better.
  + Example: Word2Vec (vector\_size=500, window=50):
    - Epochs=20: Silhouette Score = 0.194
    - Epochs=100: Silhouette Score = 0.378
  + More training leads to better embedding representations, hence better clustering.

**2.4. CBoW vs. Skipgram (Word2Vec)**

* CBOW performed better in most cases than Skipgram.
* Example: k=9, CBOW (Silhouette Score = 0.328) vs. Skipgram (0.258).
* Skipgram is better for small datasets, while CBOW generalizes better in large datasets.

**2.5. DM vs. DBOW (Doc2Vec)**

* DBOW (Distributed Bag of Words) performed slightly better than DM (Distributed Memory).
* Example: k=13, DBOW = 0.243, while DM = 0.177.
* DBOW is better for learning document-level embeddings, while DM retains more context.

**3. Summary of Key Findings**

**3.1. Best Performing Model**

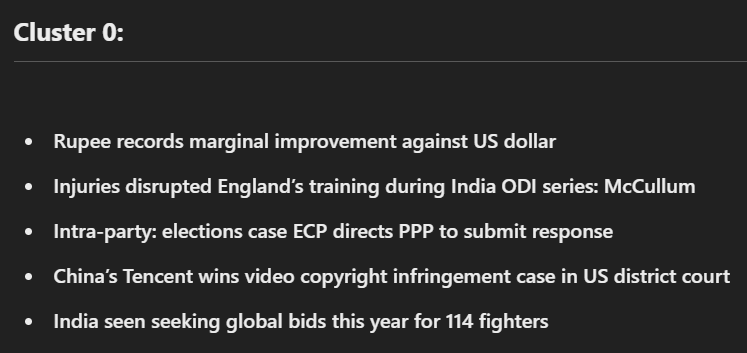
* Word2Vec (CBOW, vector\_size=500, window=50, epochs=100) achieved the best Silhouette Score (0.378) and low WSS (0.951).
* Doc2Vec (DBOW, vector\_size=500, window=50, epochs=100) performed similarly.

**3.2. Key Observations**

* Traditional methods (TF-IDF, LSA, SVD) struggled with clustering due to sparse, less informative feature representations.
* Word2Vec & Doc2Vec embeddings significantly improved cluster quality due to better semantic capture.
* Increasing vector size, window, and epochs enhances clustering performance.
* CBOW > Skipgram & DBOW > DM for clustering performance.

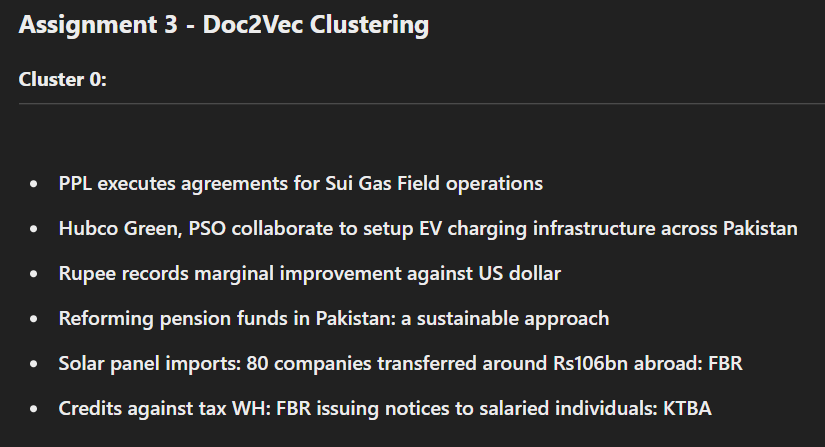
**Comparison of Headline Clustering Across Assignments**

**Assignment 2:**



 contains a mix of finance, politics, and sports headlines.

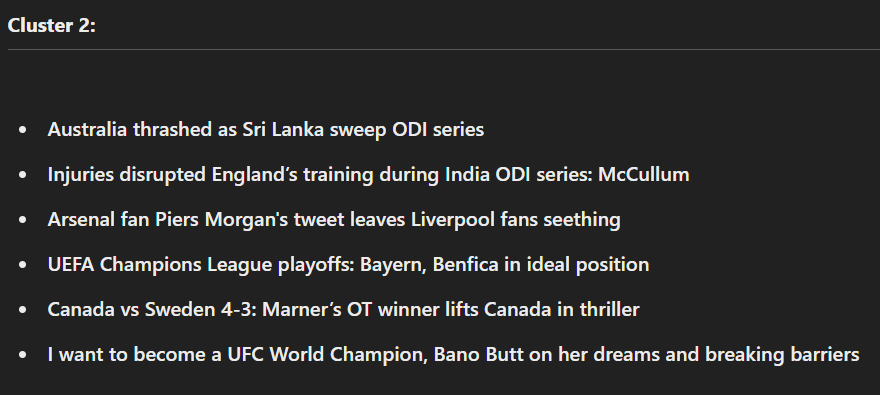
 Example: "Rupee records marginal improvement against US dollar" (finance) is grouped with "Injuries disrupted England’s training during India ODI series: McCullum" (sports).

****

 more finance/business-focused.

 The McCullum news is missing, meaning sports news is no longer mixed with finance.

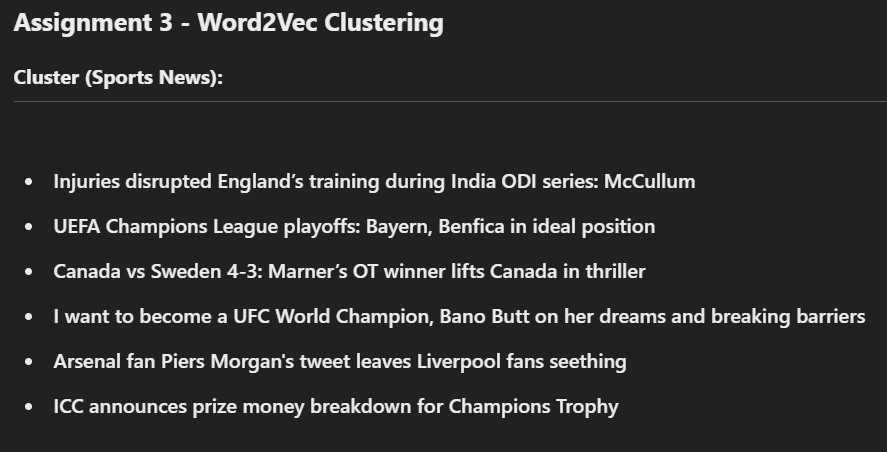
 Improved grouping: "PPL executes agreements for Sui Gas Field operations" and "Hubco Green, PSO collaborate to setup EV charging infrastructure" are both business-related.



 McCullum news correctly clusters with sports news.

 “I want to become a UFC World Champion, Bano Butt on her dreams and breaking barriers" is now with other sports headlines.

 More semantically meaningful clusters.



 Very similar to Doc2Vec, reinforcing the consistency.

 Sports news remains grouped together.

 ICC Champions Trophy prize money announcement is added, which makes sense in the sports category.

**Overall Comparison:**

****

**Word2Vec (CBOW, vector\_size=500, window=50, epochs=100) is the best performer.**

* It semantically understands that McCullum’s news is related to sports.
* Business and finance news remain separate from sports.
* Results are similar to Doc2Vec, but Word2Vec performs slightly better in maintaining topic consistency.