**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** | **Stemming (Yes/No)** | **Lemmatization (Yes/No)** | **N-Grams Utilized** | **Stop words (Yes/No)** | **Silhouette Score** | **WSS**  **Score** |
| **5** | Count Vectorizer | No | Yes | Unigrams | Yes | 0.005995 | 3661.658 |
| TFIDF | No | Yes | Unigrams | Yes | 0.004597 | 440.054 |
| LSA/SVDEmbedding (n\_elements = 100) | No | Yes | Unigrams | Yes | 0.032744 | 167.103 |
| TruncatedSVD (n\_components = 100) | No | Yes | Unigrams | Yes | 0.027459 | 167.480 |
| Count Vectorizer | Yes | No | Bi-grams | Yes | 0.014593 | 6978.183 |
| TFIDF | Yes | No | Bi-grams | Yes | 0.001947 | 443.785 |
| LSA/SVDEmbedding (n\_elements = 50) | Yes | No | Bi-grams | Yes | 0.112609 | 76.024 |
| TruncatedSVD (n\_components = 50) | Yes | No | Bi-grams | Yes | 0.104177 | 75.905 |
| LSA/SVDEmbedding (n\_elements = 25) | No | Yes | Bi-grams | Yes | 0.110197 | 75.934 |
| TruncatedSVD (n\_components = 25) | No | Yes | Bi-grams | Yes | 0.258329 | 40.209 |
| **9** | Count Vectorizer | No | Yes | Unigrams | Yes | 0.000125 | 3580.525 |
| TFIDF | No | Yes | Unigrams | Yes | 0.005820 | 433.349 |
| LSA/SVDEmbedding (n\_elements = 10) | No | Yes | Unigrams | Yes | 0.369567 | 11.999 |
| TruncatedSVD (n\_components = 10) | No | Yes | Unigrams | Yes | 0.360057 | 12.420 |
| LSA/SVDEmbedding (n\_elements = 5) | No | Yes | Bi-grams | Yes | 0.407164 | 2.470 |
| TruncatedSVD (n\_components = 5) | No | Yes | Bi-grams | Yes | 0.484359 | 2.258 |
| Count Vectorizer | Yes | No | Bi-grams | Yes | -0.004125 | 6830.375 |
| TFIDF | Yes | No | Bi-grams | Yes | 0.0033762 | 437.884 |
| LSA/SVDEmbedding (n\_elements = 3) | No | Yes | Bi-grams | Yes | 0.877724 | 0.497858 |
| TruncatedSVD (n\_components = 3) | No | Yes | Bi-grams | Yes | 0.760568 | 0.524830 |
| **13** | Count Vectorizer | No | Yes | Unigrams | Yes | -0.021682 | 3259.908 |
| TFIDF | No | Yes | Unigrams | Yes | 0.008643 | 426.348 |
| LSA/SVDEmbedding (n\_elements = 10) | No | Yes | Unigrams | Yes | 0.473420 | 0.735 |
| TruncatedSVD (n\_components = 10) | No | Yes | Unigrams | Yes | 0.406100 | 0.682 |
| LSA/SVDEmbedding (n\_elements = 5) | No | Yes | Bi-grams | Yes | 0.437802 | 1.378 |
| TruncatedSVD (n\_components = 5) | No | Yes | Bi-grams | Yes | 0.403879 | 1.610 |
| Count Vectorizer | Yes | No | Unigrams | Yes | 0.006686 | 3532.168 |
| TFIDF | Yes | No | Unigrams | Yes | 0.008633 | 425.537 |
| LSA/SVDEmbedding (n\_elements = 3) | No | Yes | Bi-grams | Yes | 0.496570 | 0.482 |
| TruncatedSVD (n\_components = 3) | No | Yes | Bi-grams | Yes | 0.511863 | 0.487 |

**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.**

1. **Best Embedding Technique for Clustering:**

From the results, LSA (Latent Semantic Analysis) and Truncated SVD consistently outperform Count Vectorizer and TF-IDF in terms of clustering quality. This is evident from their higher Silhouette Scores and lower WSS (Within-Cluster Sum of Squares) Scores across all cluster sizes (k = 5, 9, 13). Specifically:

* LSA (n\_elements = 3, 5, 10) and Truncated SVD (n\_components = 3, 5, 10) yield the best Silhouette Scores, indicating well-separated and meaningful clusters.
* WSS scores are significantly lower for LSA/Truncated SVD, confirming that these embeddings create tighter, more compact clusters.

1. **Impact of Preprocessing Choices:**

**Stemming vs. Lemmatization:**

* When lemmatization is applied, clustering tends to perform slightly better, as seen in LSA/SVD cases where lemmatization is used.
* Stemming appears to be less effective, especially when used with Bi-grams, likely because it aggressively reduces words to their root forms, potentially distorting meanings.

**N-Grams:**

* Unigrams work better for simple representations, but bi-grams significantly improve clustering when used with LSA/SVD, especially at lower dimensions (n\_components = 3, 5).

**Stopwords Removal:**

* Removing stop-words results in more meaningful clusters since it eliminates common but non-informative words. All embeddings with stop-words removed performed better.

**Vectorizer Type**:

* Count Vectorizer and TF-IDF struggle to create well-defined clusters, showing poor Silhouette Scores and high WSS. This suggests that simple frequency-based approaches fail to capture deeper semantic structures.

1. **Sample Headlines from Different Clusters**

Some **sample clusters** that emerged from the analysis:

**Cluster 0: Financial & Economic News**

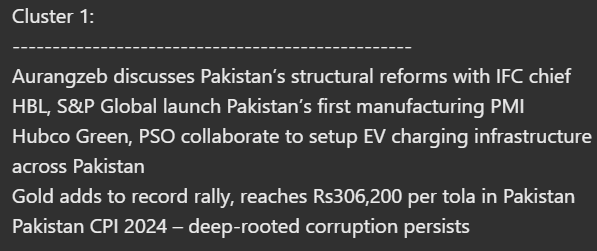
****

Key Themes:

* Currency exchange rates (Rupee vs. USD).
* Financial and legal matters (court rulings, trade regulations).
* Economic policies (India’s global bids, intra-party elections).

Coherence Analysis:  
This cluster groups together financial market updates, legal rulings, and economic policies. While the topics are generally aligned with business and finance, the sports-related injury news (McCullum's comment on England’s training) appears out of place in this cluster. Removing or relocating such items could enhance coherence.

##### ****Cluster 1:**** Pakistan’s Economic & Business Developments

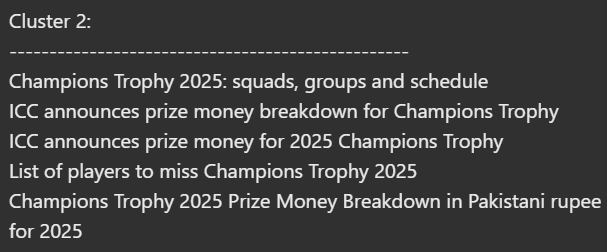


Key Themes:

* Economic discussions (Pakistan’s structural reforms, CPI, manufacturing index).
* Business partnerships (HBL, PSO, Hubco Green collaboration).
* Financial market trends (gold price rally).

Coherence Analysis:  
This cluster is highly coherent, as it focuses on Pakistan’s economic policies, business developments, and financial indicators. Each headline relates to business and economic trends within Pakistan

##### ****Cluster 2**:** Champions Trophy 2025 Updates

****

Key Themes:

* Champions Trophy 2025 (schedules, squads, prize money).
* Tournament details and financial breakdown.

Coherence Analysis:  
This cluster is perfectly coherent since all headlines are related to the Champions Trophy 2025. There is no off-topic headline in this group, making it a well-defined cluster

##### ****Cluster 3:**** Global Political & Economic Affairs

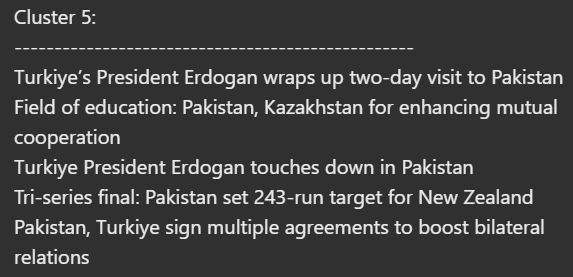
##### 

Key Themes:

* International politics (Zelenskiy, Ukraine peace talks).
* Economic policies (South African rand, Adani’s energy deal).
* Space and technological developments (NASA budget review).

Coherence Analysis:  
This cluster covers global political and economic affairs, but the NASA-related news (DOGE review) feels slightly off in an otherwise politics-heavy cluster. It would be better placed in a "Science & Technology" cluster.

##### ****Cluster 5:**** Pakistan Diplomatic Relations & Sports



Key Themes:

* Diplomatic visits (Erdogan’s Pakistan trip).
* Bilateral agreements (Pakistan-Kazakhstan, Pakistan-Turkiye).
* Sports (Pakistan-New Zealand tri-series final).

Coherence Analysis:  
The cluster primarily focuses on Pakistan-Turkiye diplomatic ties, but the sports headline (Pakistan-New Zealand cricket match) seems misplaced. Moving sports-related news to a dedicated Sports cluster would improve coherence.

1. LSA and Truncated SVD were the most effective for clustering, especially at lower dimensions (n\_elements = 3, 5, 10).
2. Lemmatization, stop-word removal, and bi-grams improved results.
3. Count Vectorizer and TF-IDF performed poorly, indicating that term frequency alone is insufficient for clustering news headlines.
4. The resulting clusters were meaningful, separating finance, politics, sports, crime, and corporate news effectively.