Institute of Business Administration
Introduction to Text Analytics
Assignment 02 – K-Means Clustering Assessment

Name: Yahya Ahmed Khan **ID: 24442**

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	Max Features
5	cv	No	Yes	(1, 2)	No	-0.024	6617.85	5000
	cv	Yes	No	(1, 3)	Yes	0.0867	2805.20	1000
	cv	Yes	No	(1, 2)	No	0.0094	6658.83	5000
	lsa	No	Yes	(1, 2)	Yes	0.1813	76.43	5000
	lsa	Yes	No	(1, 3)	Yes	0.0803	119.92	1000
	lsa	Yes	No	(1, 2)	Yes	0.1197	76.91	5000
	tfidf	No	Yes	(1, 2)	No	0.0029	443.52	5000
	tfidf	Yes	No	(1, 3)	No	0.0103	432.40	1000
	tfidf	Yes	No	(1, 2)	No	0.003	443.41	5000
	tfidf	No	Yes	(1, 3)	Yes	0.0032	443.47	5000
9	cv	Yes	No	(1, 2)	No	-0.0028	6508.32	5000
	cv	Yes	No	(1, 3)	Yes	0.0616	2688.85	1000
	cv	No	Yes	(1, 2)	Yes	-0.0121	6553.21	5000
	lsa	Yes	No	(1, 2)	Yes	0.1311	70.43	5000
	lsa	No	Yes	(1, 3)	No	0.1036	110.50	1000
	lsa	Yes	No	(1, 2)	No	0.1283	69.89	5000
	tfidf	Yes	No	(1, 2)	Yes	0.005	437.15	5000
	tfidf	No	Yes	(1, 3)	No	0.0129	423.76	1000
	tfidf	Yes	No	(1, 2)	No	0.0043	437.38	5000
	tfidf	Yes	No	(1, 3)	No	0.0043	437.47	5000
13	cv	No	Yes	(1, 2)	No	-0.0009	6389.03	5000
	cv	No	Yes	(1, 3)	No	0.0676	2608.27	1000
	cv	No	Yes	(1, 2)	No	-0.003	6462.49	5000
	lsa	No	Yes	(1, 2)	Yes	0.1985	64.27	5000

	lsa	Yes	No	(1, 3)	Yes	0.1142	100.01	1000
	lsa	Yes	No	(1, 2)	No	0.0874	63.16	5000
	tfidf	No	Yes	(1, 2)	Yes	0.0052	431.97	5000
	tfidf	Yes	No	(1, 3)	Yes	0.0179	414.29	1000
	tfidf	Yes	No	(1, 2)	Yes	0.0064	431.23	5000
	tfidf	Yes	No	(1, 3)	Yes	0.0053	432.00	5000

Analysis & Interpretation:

- **Identify which embedding technique resulted in the best clustering.**
- **Discuss how preprocessing choices impacted the results.**
- **Provide sample headlines from different clusters to analyze coherence.**

Analysis & Interpretation

The results clearly show that **Latent Semantic Analysis (LSA)** outperformed other embedding techniques in clustering. The best **Silhouette Score** (0.1985) was achieved using LSA with **bi-grams (1,2), stopword removal, and lemmatization** at **k=13**, while it also had the **lowest WSS score**. This indicates that LSA formed more distinct and compact clusters compared to **CountVectorizer (cv) and TF-IDF**, which struggled to separate clusters effectively. Many cv and TF-IDF models had near-zero or negative **Silhouette Scores**, suggesting that the clusters overlapped and were not well-defined.

Preprocessing choices had a significant impact on clustering quality. **N-gram selection played a crucial role**, with **bi-grams (1,2) performing better than tri-grams (1,3)** across most models. This suggests that capturing two-word phrases helped in distinguishing topics while avoiding the excess noise that tri-grams can introduce. **Stopword removal proved to be highly beneficial**, particularly for LSA, as it helped eliminate common but non-informative words, leading to clearer topic differentiation. **Lemmatization also performed better than stemming**, likely because it preserved the natural structure of words rather than reducing them to unnatural root forms, which can sometimes lose meaning.

Another key factor was the **max features setting**. While reducing the number of features to **1000** occasionally improved WSS scores, the best clustering results consistently came from using **5000 features**, especially when paired with LSA. This suggests that a richer vocabulary helped in forming meaningful clusters, rather than limiting the model to a smaller set of words.

Overall, if the goal is to achieve **optimal clustering**, LSA with **lemmatization, stopword removal, and bi-grams** is the best approach. These findings highlight how **embedding techniques and preprocessing choices directly impact clustering performance**, and careful tuning of these parameters can make a significant difference in real-world applications

Cluster Examples:

Running with parameters: model=lsa, ngram=(1, 2), stop_words=english, n_comp=50, max_features=5000, n_clusters=13, lemmatization=True

WSS (Within-Cluster Sum of Squares): 63.6783

Silhouette Score: 0.1036

Cluster 0:

- Justice Sarfraz Dogar sworn in as acting chief justice of IHC

- Seven new judges Join Supreme Court with Oath-Taking Ceremony

- Former SC judge Sheikh Azmat Saeed’s funeral to be held today

- In a first, new SC judges to take oath outdoors

- Law ministry notifies appointments of six SC judges, four high court chief justices

Cluster 1:

- ECC endorses purchase of $582mn capital shares in BRICS’s New Development Bank

- Rapper Gillie Da Kid claims 17-Year-Old Noah Scurry killed his son before being shot dead

- Druski roasted by NBA fans for bold 2025 All-Star game stat predictions and lack of defense

- Ralph Macchio reveals the key to his 38-year marriage and Cobra Kai legacy

- "Not my King': Anti-monarchy protesters chant at King Charles during visit to Middlesbrough

Cluster 2:

- Stocks extend overnight losses for want of triggers

- Gold price soars by Rs2,500 per tola

- Non-banking microfinance sector: SECP announces series of initiatives to empower women, strengthen consumer protection

- Govt hikes RLNG prices by up to 1.86pc

- PSX witnesses bearish trend, loses 360 points

Running with parameters: model=tfidf, ngram=(1, 3), stop_words=english, n_comp=50, max_features=5000, n_clusters=13, lemmatization=False

WSS (Within-Cluster Sum of Squares): 432.0577

Silhouette Score: 0.0051

Cluster 0:

- ECC endorses purchase of $582mn capital shares in BRICS’s New Development Bank

- Britain announces new sanctions against Putin allies

- Conspiracy theorist Gabbard confirmed as new US spy chief

- New power projects: Govt removes FSA guarantee requirement

- Soulja Boy drags Marlon Wayans’ dead mother in ongoing beef

Cluster 1:

- Hamas urges Arab summit, OIC meeting to reject Trump’s plan for Palestinian displacement

- Palestinian population 'must remain in its land’: Vatican

- PlayStation state of play returns with new games, trailers, and updates

- Pakistan condemns Netanyahu’s remarks on Palestinian state in Saudi Arabia

- Polio certificate must for Saudi-bound passengers: PIA

Cluster 2:

- Pakistan Air Force fighter Jets to kick off ICC Champions Trophy 2025 in Style

- ‘England can be dangerous in Champions Trophy despite India loss’, says Butler

- Lahore set to host 9th Faiz Festival 2025 from today

- Champions Trophy 2025 Prize Money Breakdown in Pakistani rupee for 2025

- Rupee records marginal improvement against US dollar

Running with parameters: model=cv, ngram=(1, 3), stop_words=None, n_comp=50, max_features=5000, n_clusters=13, lemmatization=True

WSS (Within-Cluster Sum of Squares): 7293.3849

Silhouette Score: -0.0550

Cluster 0:

- Karachi administration revises timings for movement of heavy vehicles

- Pakistan Refinery says will shut down plant for ‘approximately 6 days’

- Term of incumbent Ogra chairman set to expire but search for replacement not in sight

- PM thanks President Erdogan for visiting Pakistan

- Druski roasted by NBA fans for bold 2025 All-Star game stat predictions and lack of defense

Cluster 1:

- TikTok’s Back in the Game! The Viral App Finally Returns to U.S. App Stores!

Cluster 2:

- Wang's London visit marks revival of UK ties

- Senate panel advances nomination of Kash Patel as FBI director pick

- Lizzo teases ‘End of an era’ with cryptic Instagram post months after sexual abuse scandal

- Taylor Swift's bodyguard Drew becomes viral sensation for protecting the star

- Basketball star Jahki Howard caught sliding into DMs of trans influencer