Institute of Business Administration
Introduction to Text Analytics
Assignment 03 – 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 a random seed value to your ERP ID for each K-Means clustering experiment.

	Vectorizer Type and Details	vector_size	window	Epochs Count	CBoW/Skipgram OR DM/DBoW	Silhouette Score	WSS Score
5	Word2Vec, FastText	300	na	na	Skip-gram	0.038	57.485
	Word2Vec	100	5	20	Skip-gram	0.336	0.753
	Word2Vec	50	5	20	Skip-gram	0.276	1.172
	Word2Vec	100	3	20	Skip-gram	0.101	0.320
	Word2Vec	50	3	20	CBOW	0.032	0.420
	Doc2Vec	50	5	20	DBoW	0.022	0.821
	Doc2Vec	50	3	20	DBoW	0.022	0.821
	Word2Vec	50	5	10	CBOW	0.021	0.381
	Doc2Vec	50	5	20	DM	0.020	0.730
	Doc2Vec	50	3	20	DM	0.019	0.733
9	Word2Vec, FastText	300	na	na	Skip-gram	0.035	54.053
	Word2Vec	100	5	20	Skip-gram	0.221	0.501
	Word2Vec	50	5	20	Skip-gram	0.160	0.900
	Word2Vec	50	3	20	Skip-gram	0.056	0.573
	Word2Vec	100	3	20	CBOW	0.027	0.206
	Word2Vec	50	5	20	CBOW	0.026	0.415
	Doc2Vec	50	3	20	DM	0.019	0.706
	Doc2Vec	50	5	20	DM	0.019	0.705
	Doc2Vec	50	3	10	DBoW	0.018	0.690
	Doc2Vec	50	5	10	DBoW	0.018	0.690
13	Word2Vec, FastText	300	na	na	Skip-gram	0.037	51.823
	Word2Vec	100	5	20	Skip-gram	0.132	0.452
	Word2Vec	50	5	20	Skip-gram	0.114	0.841
	Word2Vec	50	3	20	Skip-gram	0.041	0.552
	Word2Vec	50	5	20	CBOW	0.029	0.400
	Word2Vec	100	3	20	CBOW	0.024	0.200
	Doc2Vec	50	3	10	DM	0.020	0.662

	Word2Vec	100	5	10	CBOW	0.020	0.184
	Word2Vec	50	3	10	CBOW	0.020	0.353
	Doc2Vec	50	5	10	DM	0.020	0.661

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)**

The best clustering results came from **Word2Vec Skip-gram with a vector size of 100, window size of 5, and 20 epochs**, achieving the highest **Silhouette Score (0.336)**. This means the clusters were well-separated, making it the most effective setup. On the other hand, **Word2Vec CBOW (100, window 5, 10 epochs) had the lowest WSS score (0.184)**, indicating compact clusters but with less separation.

Larger vector sizes didn't always help, **100 dimensions worked better than 50, but 300 added noise** and reduced clustering quality. A **larger window (5) improved performance** by capturing broader context, while **Skip-gram consistently outperformed CBOW in separation**. **Doc2Vec (both DM and DBoW) performed the worst**, with low Silhouette Scores around **0.018-0.022**. Increasing the **number of epochs to 20** helped refine the embeddings and improve clustering.

Overall, **Word2Vec Skip-gram (100, window 5, 20 epochs) was the best setup**, providing a good balance of separation and compactness.

Compared the previous assignment, the WSS scores are very different, mainly because the WSS is not scaled and depends on how large the vectors are for the document embeddings. In this case, I normalized the vectors so that they all are of length 1. However, the silhouette scores are comparable across models, so those showed a significant improvement from previous results, going as high as 0.3.