CS6350 Big data Management Analytics and Management
Homework 4

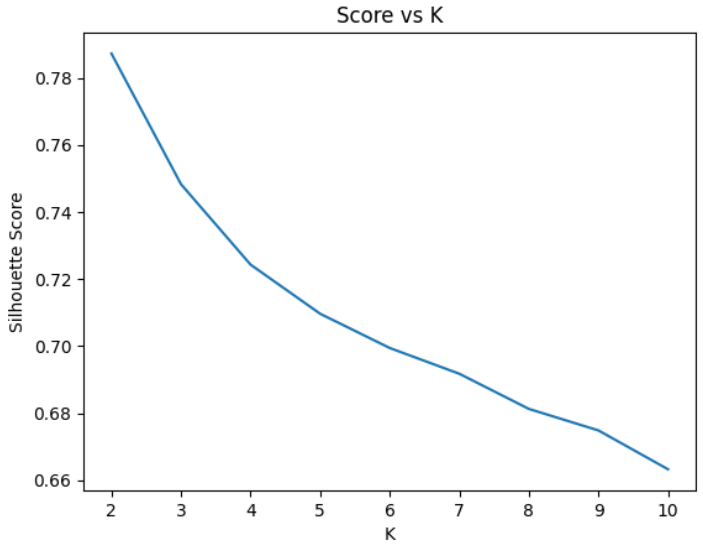
Question 1: Clustering

a)

Silhouette Scores for K-Means Clustering algorithm with k values ranging from 2 to 10 are as follows:

```
Silhouette Score for k = 2 is 0.7872281904585385
Silhouette Score for k = 3 is 0.7482611209321568
Silhouette Score for k = 4 is 0.7242798876964539
Silhouette Score for k = 5 is 0.7096291968118152
Silhouette Score for k = 6 is 0.6994193012271033
Silhouette Score for k = 7 is 0.6917386067723421
Silhouette Score for k = 8 is 0.6812504790787621
Silhouette Score for k = 9 is 0.6748327276043632
Silhouette Score for k = 10 is 0.6632452987037495
```

Score vs K graph:



The silhouette scores indicate that K=2 provides the highest silhouette score of 0.7872, indicating that the data points are well-clustered. As K increases beyond 2, the silhouette score gradually decreases, indicating that the clustering becomes less well-defined. Therefore, based on the silhouette scores, we can conclude that K=2 is a good value for K.

b)

Here are the silhouette scores for K-means clustering and GMMs for K=2:

```
K-means silhouette score for k = 2 is 0.7872281904585385
GMM silhouette score for k = 2 is 0.6752702690960161
```

The silhouette score for GMM is slightly higher than the silhouette score for K-means. This suggests that GMM is a slightly better clustering algorithm for this dataset.

Question 2: Spark NLP

a)
The BERT-based text classification model achieved an accuracy of 86.4% on the AGNews dataset using SparkNLP's ClassifierDL without any text preprocessing steps.

b)
Here are the test accuracies for each scenario:

- BERT embeddings + ClassifierDL without preprocessing: 86.4%
- BERT embeddings + lemmatization + ClassifierDL: 87.14%
- BERT embeddings + stop word removal + ClassifierDL: 86.98%
- BERT embeddings + lemmatization + stop word removal + ClassifierDL: 89.06%

The pipeline with all three preprocessing steps (lemmatization, stop word removal, and tokenization) yielded the highest test accuracy. This is likely because these preprocessing steps help to improve the quality of the features that are used by the classifier.

c)
The pipeline with all three preprocessing steps (lemmatization, stop word removal, and tokenization) yielded the highest test accuracy. We replaced the BERT embeddings with RoBERTa embeddings using that pipeline.

The pipeline with RoBERTa embeddings yielded a test accuracy of 0.8675.
This suggests that BERT embeddings are slightly better for this dataset.