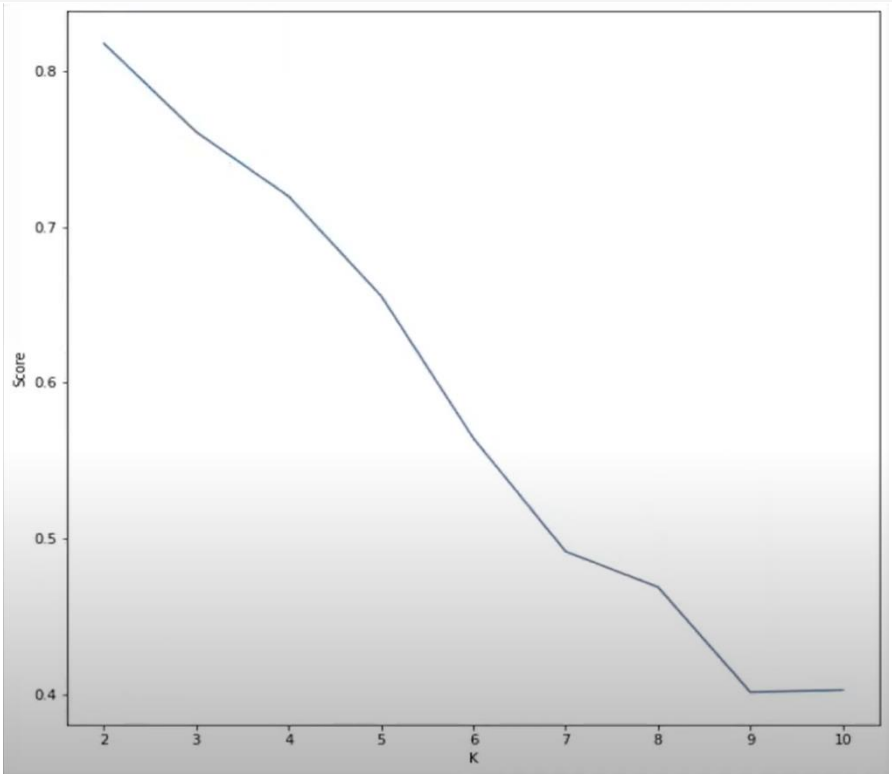
CS6350
Big data Management Analytics and Management
Spring 2023
Homework 4
Submission Deadline: April 30th, 2023, 11:59pm

Question 1: Clustering

a) Perform K-means clustering using Spark MLib library on the given dataset using values of K ranging between 1 to 10 and generate the corresponding silhouette scores for each K.

Then plot the Score vs K graph as shown below. Note: The diagram below is an example only and might not be same as your answer. Analyze the graph to determine a good value for K. Give a brief reasoning for your answer.

b) For the value of K that you chose in part (a), compare the performance of K-means clustering with that of Gaussian Mixture Model. Use silhouette scores to determine which is the better model and why.



Question 2: Spark NLP

In this task you are required to import SparkNLP library to implement a text classification model. Use AGNews dataset (check reference 1 about how you can download it) to train for 5 epochs and compare the performance of these different models:

a) Use BERT embeddings with a generic annotator model in SparkNLP called ClassifierDL, without any text preprocessing steps and find the test accuracy for it.

b) Add preprocessing steps, specifically lemmatization and stop word removal, to the pipeline in (a) and compare its impact on the overall performance of the model. Report the test accuracies when each step is implemented individually and when they are used together. Identify the pipeline that yields the highest test accuracy and give a brief explanation of why it performs the best.
c) Lastly, select the best pipeline from (a) and (b) and use RoBerta embeddings instead of BERT embeddings. Report which embedding gives the best results and why.

You can use Google Colab to do this task. Use the following links for reference:

1) https://github.com/JohnSnowLabs/spark-nlp-workshop/blob/master/tutorials/Certification_Trainings/Public/5.1_Text_classification_examples_in_SparkML_SparkNLP.ipynb
2) https://towardsdatascience.com/text-classification-in-spark-nlp-with-bert-and-universal-sentence-encoders-e644d618ca32
3) https://nlp.johnsnowlabs.com/docs/en/quickstart
4) https://github.com/JohnSnowLabs/spark-nlp-workshop/tree/master/tutorials/Certification_Trainings/Public

What to submit: All the code files and a separate pdf answering all the questions.