Lab - 7 Assignment

References:

1. Kafka Documentation - https://kafka.apache.org/documentation/
2. Nice blog on Kafka - https://sookocheff.com/post/kafka/kafka-in-a-nutshell/
3. Spark Streaming API - https://spark.apache.org/docs/latest/streaming-programming-guide.html
4. Spark Streaming + Kafka API (Receiver based and Dstream) – https://spark.apache.org/docs/2.0.0-preview/streaming-kafka-integration.html
5. Structured Streaming API - https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html
6. Structured Streaming + Kafka API - https://spark.apache.org/docs/latest/structured-streaming-kafka-integration.html
7. Structured Streaming paper - https://cs.stanford.edu/~matei/papers/2018/sigmod_structured_streaming.pdf
8. How to run Kafka on GCP - https://www.learningjournal.guru/courses/kafka/kafka-foundation-training/kafka-in-gcp/

Assignment:

The aim of this assignment is to use the iris model trained in lab5 for making real-time predictions.

1. Write a producer.py file that reads the iris.csv line by line and writes each row into a particular topic in Kafka.
2. Write a subscriber.py file that uses spark streaming (can be receiver-based, dstream or structured) for producing real-time predictions on these rows by utilizing the model trained in lab5 and calculates the accuracy (the real-time predictions, true labels and accuracy all should get printed on console).

Note that for task 2, you also need to figure out how to save a trained model and load it back for predictions.

Submission:

1. Create a PDF report that contains screenshots demonstrating
   a. The rows written by producer.py are received by a consumer (basically producer.py works fine)
   b. The console output generated by subscriber.py
2. The python files - producer.py and subscriber.py

Zip the pdf along with the python files and submit the zip file on Moodle. Also attach screenshots both the tasks.

Note:

The jar files for Structured Streaming are provided below:

1. gs://bdl2022/lab7/jar_files/commons-pool2-2.6.2.jar
2. gs://bdl2022/lab7/jar_files/kafka-clients-2.6.0.jar
3. gs://bdl2022/lab7/jar_files/lz4-java-1.7.1.jar
4. gs://bdl2022/lab7/jar_files/scala-library-2.12.10.jar
5. gs://bdl2022/lab7/jar_files/slf4j-api-1.7.30.jar
6. gs://bdl2022/lab7/jar_files/snappy-java-1.1.7.3.jar
7. gs://bdl2022/lab7/jar_files/spark-sql-kafka-0-10_2.12-3.1.1.jar
8. gs://bdl2022/lab7/jar_files/spark-tags_2.12-3.1.1.jar
9. gs://bdl2022/lab7/jar_files/spark-token-provider-kafka-0-10_2.12-3.0.0-preview2.jar
10. gs://bdl2022/lab7/jar_files/unused-1.0.0.jar
11. gs://bdl2022/lab7/jar_files/zstd-jni-1.4.4-7.jar

The jar files required for Receiver-based API or DStream-based API can be figured out by looking at the documentation.