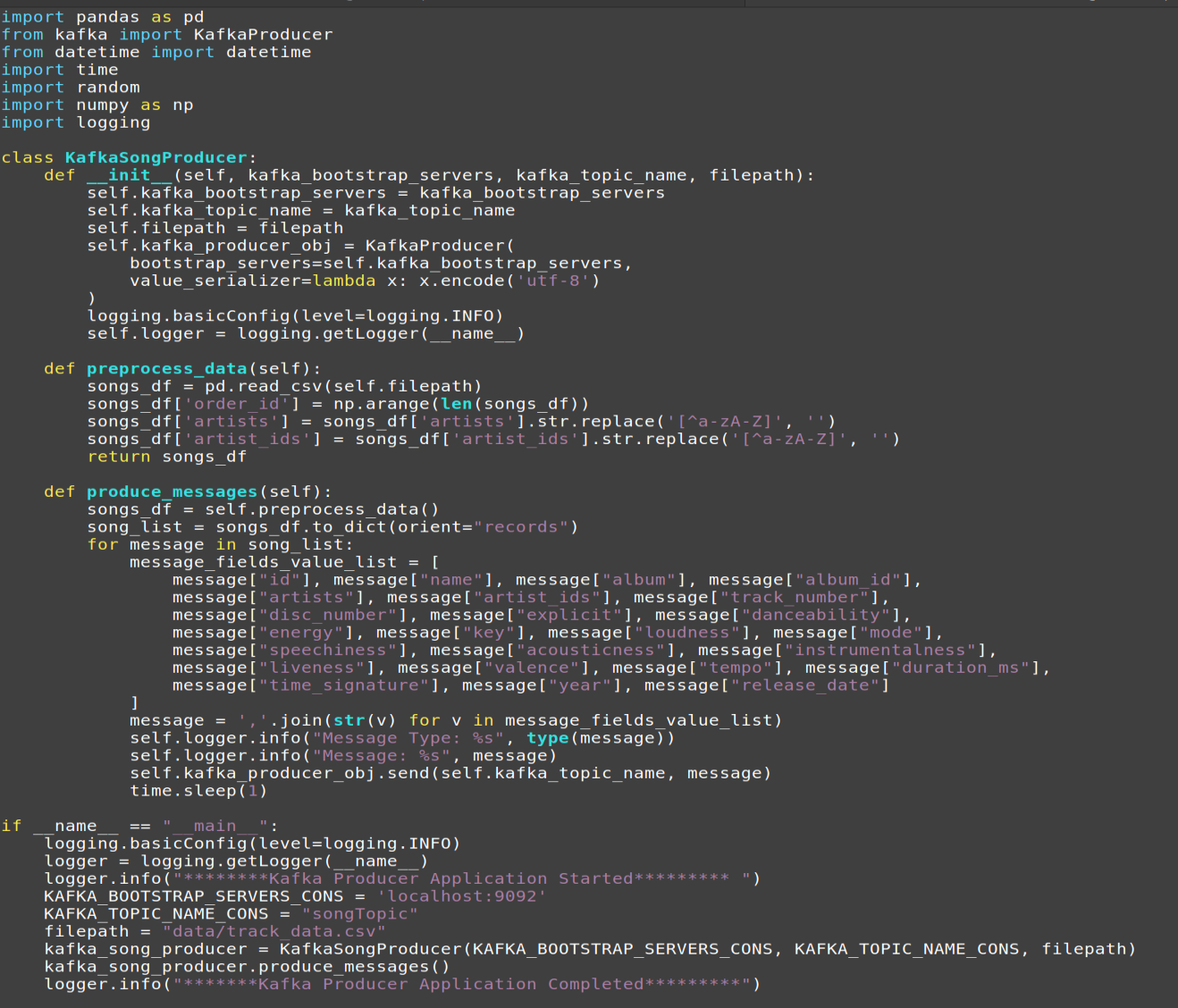
CS624 Final Project

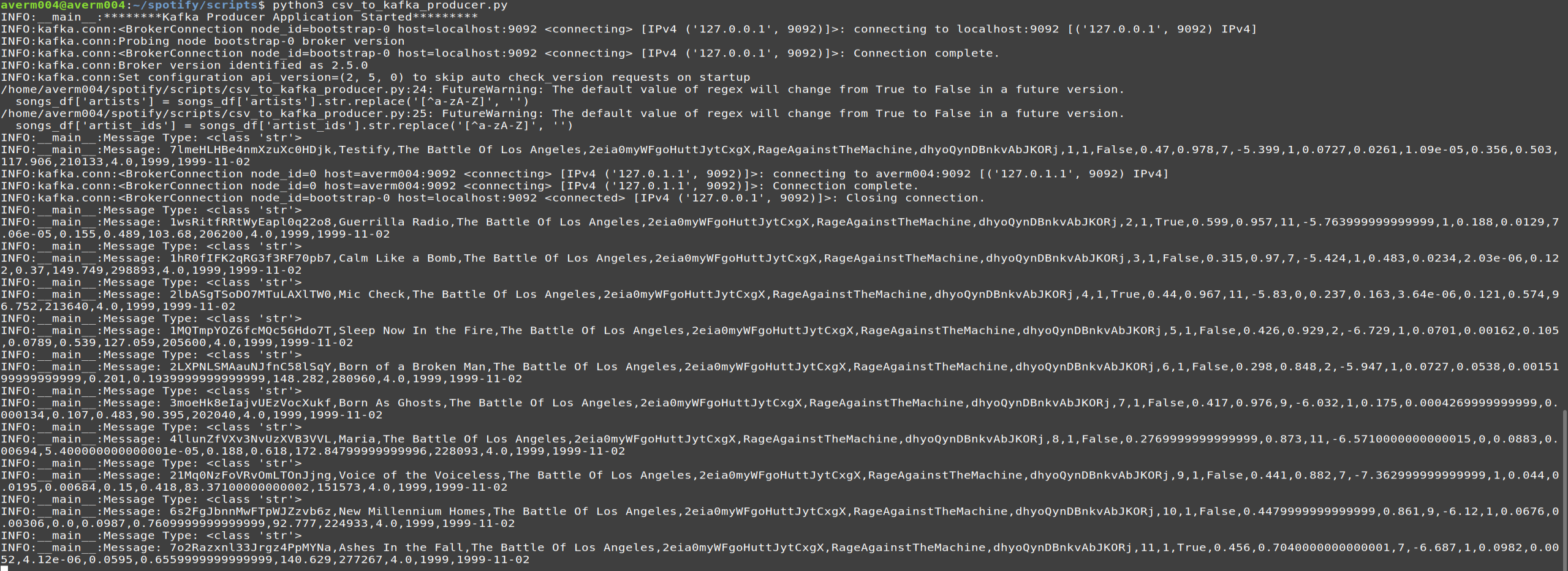
Spotify Song Recommendation System

Reading data from Csv file(1.2 million songs data) into Kafka topic.

Csv\_to\_kafka\_producer.py



Python3 csv\_to\_kafka\_producer.py



<https://www.kaggle.com/code/fusshandschuhe/spotify-eda-with-pyspark-and-clustering>

<https://www.mattmoocar.me/blog/Spark-Music-Rec/>

<https://www.kaggle.com/code/prasadpil/bitcoin-tweet-nltk-sentiment-analysis-pyspark>

<https://www.kaggle.com/code/sandeepanmukherjee/pyspark-bitcoin-trend-analysis#Building-A-Machine-Learning-Model-With-Spark-ML>

<https://www.kaggle.com/datasets/mczielinski/bitcoin-historical-data>

<https://learn.microsoft.com/en-us/fabric/data-science/time-series-forecasting>

https://stackoverflow.com/questions/42918663/is-it-better-to-have-one-large-parquet-file-or-lots-of-smaller-parquet-files

spark.conf.set(“spark.serializer”, “org.apache.spark.serializer.KryoSerializer”)

spark.conf.set(“spark.sql.shuffle.partitions”,10)

spark.dynamicAllocation.enabled,spark.dynamicAllocation.minExecutors,

spark.dynamicAllocation.schedulerBacklogTimeout,

spark.dynamicAllocation.maxExecutors,

spark.dynamicAllocation.executorIdleTimeout

<https://medium.com/@hrushikesh.raghavendra/end-to-end-ml-pipeline-using-pyspark-and-databricks-part-ii-f25f512a8626>

display(dbutils.fs.ls("dbfs:/databricks-datasets/wiki/"))