import sys

from pyspark.context import SparkContext

from awsglue.context import GlueContext

from awsglue.job import Job

from pyspark.sql.functions import col, regexp\_replace, lower

# Initialize Glue context

sc = SparkContext()

glueContext = GlueContext(sc)

spark = glueContext.spark\_session

job = Job(glueContext)

job.init('clean\_reviews\_job', args={})

# Load the CSV from S3

input\_path = "s3://csv-cleaner-data/raw\_reviews.csv"

df = spark.read.option("header", True).csv(input\_path)

# Define list of profane words

profane\_words = ['badword1', 'badword2', 'worst']

# Step 1: Clean irrelevant characters

df\_cleaned = df.withColumn("review\_text", regexp\_replace(col("review\_text"), r"[^a-zA-Z0-9\s]", ""))

# Step 2: Convert to lowercase

df\_cleaned = df\_cleaned.withColumn("review\_text", lower(col("review\_text")))

# Step 3: Remove profanity

for word in profane\_words:

df\_cleaned = df\_cleaned.withColumn("review\_text", regexp\_replace("review\_text", word, "\*\*\*"))

# Step 4: Normalize ratings

df\_cleaned = df\_cleaned.withColumn("rating", col("rating").cast("float"))

# Step 5: Save output to S3

output\_path = "s3://csv-cleaner-data/output/cleaned\_reviews/"

df\_cleaned.coalesce(1).write.mode("overwrite").option("header", True).csv(output\_path)

job.commit()