CS 6350

ASSIGNMENT \_\_\_\_\_\_2\_\_\_\_

Names of students in your group:

Rui Yang

Netid:rxy200003

Number of free late days used: \_\_\_\_\_\_\_\_\_\_\_\_\_1\_\_\_\_\_\_\_   
Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

<https://stackoverflow.com/questions/57703270/writing-databricks-dataframe-to-s3-using-python/59670324#59670324>

<https://docs.databricks.com/data/data-sources/aws/amazon-s3.html#language-python>

<https://sparkbyexamples.com/pyspark/pyspark-filter-rows-with-null-values/>

<https://spark.apache.org/docs/latest/ml-tuning.html>

<https://spark.apache.org/docs/latest/ml-pipeline.html>

<https://spark.apache.org/docs/latest/mllib-evaluation-metrics.html>

<https://stackoverflow.com/questions/29717257/pyspark-groupbykey-returning-pyspark-resultiterable-resultiterable>

<https://stackoverflow.com/questions/36392938/spark-expansion-of-rddkey-list-to-rddkey-value>

Part1: <https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/8750190078783311/3984119822501300/5504697096027154/latest.html>

Part2: <https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/8750190078783311/3444072263407973/5504697096027154/latest.html>

dataset.csv for part1

Tweets.csv for part2

Both of them read input file to get arguments.

part-00000-tid-205.. is the output of part1

Part-00000-tid-858.. is the output of part2