CS 6350

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

Names of students in your group:

Yuan Zhuang (yxz157830)

Jia Zheng Liu (jxl210006)

Number of free late days used: \_\_\_\_\_\_\_\_0\_\_\_\_\_\_\_\_\_\_\_\_   
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.

1. <https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-output.html> Configure input and output location in Amazon AWS S3

2. <https://docs.aws.amazon.com/AmazonS3/latest/userguide/create-bucket-overview.html> How to create AWS S3 bucket

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

<https://stackoverflow.com/questions/57703270/writing-databricks-dataframe-to-s3-using-python> How to write data output to AWS S3 bucket using python.

4. <https://supsystic.com/documentation/id-secret-access-key-amazon-s3/> How to get access key fron AWS S3

5. <https://medium.com/@shamnad.p.s/how-to-create-an-s3-bucket-and-aws-access-key-id-and-secret-access-key-for-accessing-it-5653b6e54337> How to configure access key via awscli in system

6. <https://docs.databricks.com/data/data-sources/aws/amazon-s3.html#mount-aws-s3> Access S3 bucket through DBFS

7. <https://sparkbyexamples.com/pyspark/pyspark-filter-rows-with-null-values/> How to deal with the null value Pyspark

8. <https://stackoverflow.com/questions/29717257/pyspark-groupbykey-returning-pyspark-resultiterable-resultiterable> Pyspark result iteration

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

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

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

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

13. <https://www.youtube.com/watch?v=PUHZbVo2qwA> Read csv file from GitHub

14. <http://lintool.github.io/UMD-courses/bigdata-2015-Spring/slides/session05.pdf> PageRank

ToubleShoot:

1. <https://aws.amazon.com/premiumsupport/knowledge-center/iam-assume-role-cli/> S3 bucket access denied issue

**2.** [**https://grabngoinfo.com/databricks-mount-to-aws-s3-and-import-data/**](https://grabngoinfo.com/databricks-mount-to-aws-s3-and-import-data/) **Complete guild of connection between AWS and databricks (most useful.)**

3. <https://sparkbyexamples.com/spark/spark-write-dataframe-single-csv-file/> Write dataframe parittion into same csv file.

------------------------------------------------------------------------------------------------------

Question 1

PageRank with MapReduce on Databricks:

<https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/519162593337409/2053566011294921/863443026627963/latest.html>

Input file on GitHub:

<https://raw.githubusercontent.com/JiazhengLiu-utd/CS6350/61471b9179a4bbd5f2b2ee9d7d551f58ddc2f9e8/T_T100D_SEGMENT_US_CARRIER_ONLY.csv>

Input file on AWS S3:

<https://utd-cs6350-output.s3.us-west-2.amazonaws.com/T_T100D_SEGMENT_US_CARRIER_ONLY.csv>

Output file on GitHub:

<https://raw.githubusercontent.com/JiazhengLiu-utd/CS6350/main/part-00000-tid-4660999883501531712-2ba2112b-3d9e-494c-94e3-0ebfab81d2d5-9161-1-c000.csv>

Output file on AWS S3:

<https://utd-cs6350-output.s3.us-west-2.amazonaws.com/pagerank/output/part-00000-tid-4660999883501531712-2ba2112b-3d9e-494c-94e3-0ebfab81d2d5-9161-1-c000.csv>

Question 2

For multi-class evaluation the overall data, some of the statistics such as precision and recall caused boardcast error, so that there are more weighted statistics evalued.

<https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/4175355552096756/3879227318673298/7031722269867713/latest.html>

Question 2 output file name : part-00000-tid-5304242498049628626-a71f366e-8a40-40ab-9343-1771e24cf4ac-2569-1-c000.csv