CS 6350
ASSIGNMENT 2

Names of students in your group:

Anirudh Thakur (axt190073)
Rathang Rajpal (rxr210009)

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://python.hotexamples.com/examples/pyspark.ml.feature/StopWordsRemover/-/python-stopwordsremover-class-examples.html

https://george-jen.gitbook.io/data-science-and-apache-spark/tokenizer
https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/3923635548890252/1357850364289680/4930913221861820/latest.html

https://www.analyticsvidhya.com/blog/2019/11/build-machine-learning-pipelines-pyspark/

https://towardsdatascience.com/countvectorizer-hashingtf-e66f169e2d4e

DataBricks link :
Part 1:
https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/1923465492721157/2734248298629331/3673770524474654/latest.html

Part 2:

https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/1923465492721157/2785769998928179/3673770524474654/latest.html

The python notebook, Input data and output data are also available in the zip file

Entire project GitHub link :
https://github.com/Anirudh-thakur/AirportPageRanking

Output file :
Part 1:
https://raw.githubusercontent.com/Anirudh-thakur/AirportPageRanking/main/Output/Q2.1/AirportPageRank.csv

Part 2 :
https://raw.githubusercontent.com/Anirudh-thakur/AirportPageRanking/main/Output/Q2.2/Q2Predictions.csv

https://raw.githubusercontent.com/Anirudh-thakur/AirportPageRanking/main/Output/Q2.2/Q2Metric.csv