[Fall-2021]

HW5

Deadline: December 1st 11:59PM ET (8:59PM PT)

General Description:

In this homework, you are asked to build the inverted index algorithm, discussed in the lecture, using Apache Spark.

- Your application should use the Data.zip posted on Canvas for Course Project Option-2. You may unzip the file if it will make your task easier.
- For each Term, your posting list should contain the file name and the frequency of the term in each file. Your term may have multiple posting lists depending on its repetition across several files.
- Your "stop word list" should contain the following terms:
    - they
    - she
    - he
    - it
    - the
    - as
    - is
    - and
- Use the following docker container for running Spark (You can use PySpark): https://hub.docker.com/_/microsoft-mmlspark
- You will have to make your data available to the Docker container.
- You may use Jupyter notebooks inside that container to run your code.
- You use the following guide to develop Spark in Python
    - https://spark.apache.org/docs/latest/quick-start.html

Submission Guidelines:

- Post URL for your GitHub repository to Canvas. Make sure to keep your GitHub repository public.
- Your GitHub repository should contain the following:
    1. Copy of Your PySpark code for the inverted index algorithm (or equivalent in another programming language) (50% of total grade)
    2. Copy of your output (50% of total grade)

Extra Credit:

- (+10%) Run your Apache Spark Code on Microsoft Azure. (You may create a free trial account to run it). (You will have to run it on Azure in addition to running your code in Docker as well).
    - If you do so, submit a screenshot of the execution on Microsoft Azure.
- (+10%) Deploy your Apache Spark Code to Kubernetes on Microsoft Azure.
    - If you do so, submit a screenshot of the application's execution on Kubernetes hosted on Microsoft Azure.

Common Penalties:

- Your GitHub repository is not public: 100% reduction (won't be graded)
- Late submissions on Canvas or GitHub: 100% reduction (won't be graded)