DS 561 Assignment 2

Vishwas Bhaktavatsala - U74206902
Github: git@github.com:vishwas21/DS561-vishwas-assignemnt2.git

---

File Description

Assignment2.py: Involves the complete code to open the bucket, read contents, parse the file, find out the incoming links, outgoing links, its counts, mean, median, max, min, and page rank.

Steps to Run the code

Clone the code using the command:
git clone https://github.com/vishwas21/DS561-vishwas-assignemnt2.git

Install the requirements using the command: pip3 install -r requirements.txt.
Execute the assignment2.py file to generate the results.

Function Descriptions:

Main(): This is the main function which can be set as the driver for the whole program. This code. This function sets the default values for the global variables which are used in almost all the functions.

checkPageRankUpdate(): This function findout the difference between the previous sum of page ranks and the current sum and decides weather to run one more iteration of the page rank algorithm or not

calculatePageRank(): This function is the driver code which manages the execution and process of finding the pageRank.

calculatePRX(): This function calculates the inner part of the formula which is required to calculate the page rank of a particular node.

operationsForMetrics(): Function used to set the base for all the metrics which is required all the other operations which are needed for this code. Operations such as running the required

functions to parse the files, create the graph and running the output code which shows us the 5 top significant files.

calculateStatistics(): This function as the name suggests calculates the Mean, Median, Min, Max, and Quintiles for the all the incoming links and outgoing links across all the files.

saveToFile() and readFromJson(): This is just a temporary functions which is created to write the graph dictionary to a json file and read it from it.

parseFileHtml(): This function reads each file from the bucket and parses the content so that the program would be able to create the graph and keep track of all the incoming and outgoing links and their counts

readStorageBucket(): This function is the main function which houses the code for reading the bucket and its file.