**DS 561 Assignment 2**

Harshitha Tumkur Kailasa Murthy U00683580

Files and Descriptions:

fileGeneration.py: code used to generate 10K files with 250 links in each file. Random.seed set to 0.

Upload.py: file used to upload all the generated files to the storage bucket by creating a folder.   
  
pageRank.py: This file includes all the functions to run the page rank algorithm.

requirements.txt: The file that contains all the required modules to be installed. Command: pip3 install -r requirements

How to run a program for a new user

Clone the GitHub repository:

To generate the files which would not be present on GitHub please execute python3 fileGeneration.py after cd to fileGeneration folder.

To execute the page rank algorithm, execute the command python3 pageRank.py to generate output.

Function descriptions:

readFromStorageBucket : read from the GCP storage bucket created.

HTMLFileParser : Function written to parse the HTML files generated and list the incoming and outgoing links.

CalculateStatistics: By considering the incoming and ougoing links this function calculates all the required stats like mean max median min and quintiles.

CalculatePRInner: this function is to calculate the inner part of the formula

CalculatePageRank: Calculate page rank completely.

PageRankUpdate: function to check the page rank update is required based on the previous page ranks.