Homework 2 (Report)

Rhythm Somaiya – U84158310

GitHub Link: *https://github.com/IAmRhy31/CDS-DS-561-Homework-2*

- **Project Name: *ds-561-project-1***
- **Bucket Name: *hw2-rhythm***
- **Directory Name inside the bucket: *generated-content/***
- **Service Account Name: *hw2-rhythm-service-account@ds-561-project-1.iam.gserviceaccount.com***

HOW TO RUN AND VERIFY THE CODE:

- Clone the GitHub repository (link mentioned above)

- Run the code in your PowerShell/Terminal using the command:
  python3 pagerank.py

BRIEF DESCRIPTION ON THE CODE AND PARAMETERS IT ACCEPTS:

Firstly, the 'calculate_pagerank' function allows us to compute PageRank scores for a collection of web pages by using an iterative algorithm. It accepts three parameters: graph_outlinks, graph_inlinks, and pagerank_values. graph_outlinks and graph_inlinks describe the web page link structure, where graph_outlinks shows the outgoing links for each page, and graph_inlinks shows the incoming links. pagerank_values is a list initially containing PageRank values for each page. Hence, this function iteratively updates these values until they converge, ensuring that more influential pages receive higher scores. Lastly, it gives the calculated PageRank values.

Further, the 'main' function serves as the entry point for our program and e use it to initialize data structures for web page relationships, including incoming and outgoing links, PageRank values, and page statistics. Then, we process HTML files that represent web pages and populate our data structures accordingly. Once that's done, we call the calculate_pagerank function to compute PageRank scores. Finally, we print the top 5 pages with the highest PageRank scores and calculate statistics for incoming and outgoing links, such as averages, medians, maximums, minimums, and quintiles.

RELEVANT OUTPUTS:

```
Top 5 Pages by PageRank Score:
1. Page 2526: PageRank Score = 2.35081717640534
2. Page 6846: PageRank Score = 2.1160231497129116
3. Page 5971: PageRank Score = 2.0614418606857203
4. Page 5778: PageRank Score = 2.0494611341913664
5. Page 4961: PageRank Score = 2.047160786180241
Incoming Links Statistics:
Average: 123.6451
Median: 124.0
Max: 188
Min: 82
Quintiles: [114. 121. 126. 133.]
Outgoing Links Statistics:
Average: 123.6451
Median: 123.0
Max: 249
Min: 0
Quintiles: [ 49.  98. 149. 198.]
```