CSEN 272 Web Search and Information Retrieval

PageRank Programming Project

Manasa Madiraju (07700009942)

Execution Commands:

- Created a graph with 10,000 web pages. Below are the instructions on the execution of the program in its default state.
- Sample execution command is as below

```
python PageRank.py input.txt 0.75 > output.txt
```

- **(Optional)** True - Passing in another parameter **True** at the end of the input parameters will also show the metrics like 1.)The number of iterations. 2.)The amount of time taken for convergence. 3.)The top 10 web pages with their pagerank values

```
python PageRank.py input.txt 0.75 True > output.txt
```

Observations:

- What convergence criterion did you use?
    - Used the sum of absolute differences between the new and previous PageRank vectors falls below a threshold of 1e-10 and a maximum iteration limit of 1000.
- How long did your PageRank algorithm converge in one run on average ?
    - For this 10,000 web page graph, the time to converge in one run on average was around 0.277909 seconds (0.233644 seconds, 0.258223 seconds, 0.270343 seconds, 0.295331 seconds and 0.332004 seconds for d values 0.75, 0.80, 0.85, 0.90 and 0.95 respectively).
- How different are the top sites for each d ? How different are the PageRanks of the top sites? How does the PageRank vector change as a whole? Write any observations you find.
    - The top two web pages remained the same across different d values. However, as d increased, the ranking order of the remaining web pages shifted, and PageRank values became more concentrated at the top.
    - Additionally, 15 webpages consistently appeared in the top 25 across different d values.
    - PageRank Vector as a Whole : As d increases - The PageRank vector becomes more concentrated on a smaller set of highly linked pages. The distribution of PageRank values becomes more skewed, with a few pages having very high values and most pages having very low values. As d decreases - The PageRank vector becomes more uniform, with less concentration on highly linked pages. The distribution of PageRank values becomes flatter, with less skew.
    - Observations:
        - Higher d values result in larger PageRank scores for the top nodes.
        - The number of iterations required for convergence increases with d as teleportation (random jumping probability) decreases.
        - Showing the top 3 values for each d value and the number of iterations that has taken for convergence

0.75	0.80	0.85	0.90	0.95
P(5566): 4.1562108522e-04 P(5377): 3.7040129708e-04 P(3555): 3.6674313761e-04	P(5566): 4.5417644457e-04 P(5377): 4.1080307425e-04 P(3555): 3.9440513769e-04	P(5566): 4.9539343622e-04 P(5377): 4.5574346550e-04 P(7705): 4.2766492703e-04	P(5566): 5.3925203441e-04 P(5377): 5.0579102318e-04 P(7705): 4.6484380489e-04	P(5566): 5.8567842700e-04 P(5377): 5.6164246788e-04 P(3890): 5.0701064415e-04
25 iterations	27 iterations	29 iterations	32 iterations	34 iterations