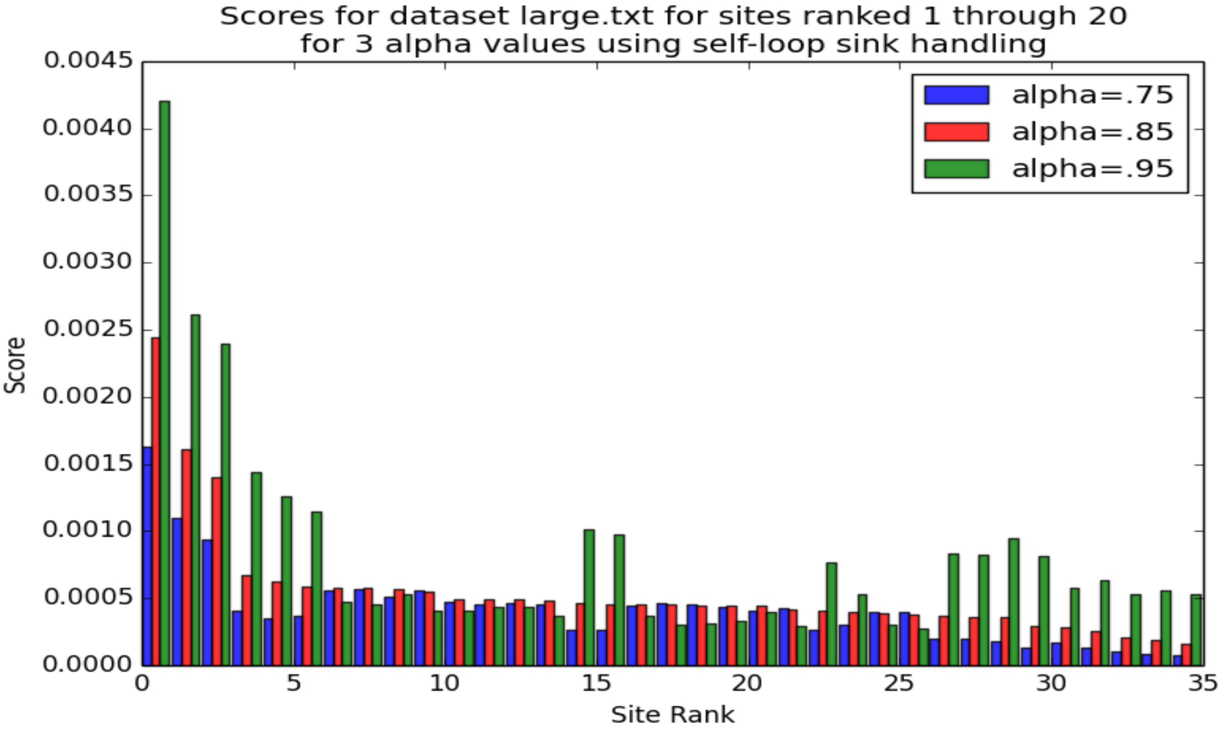
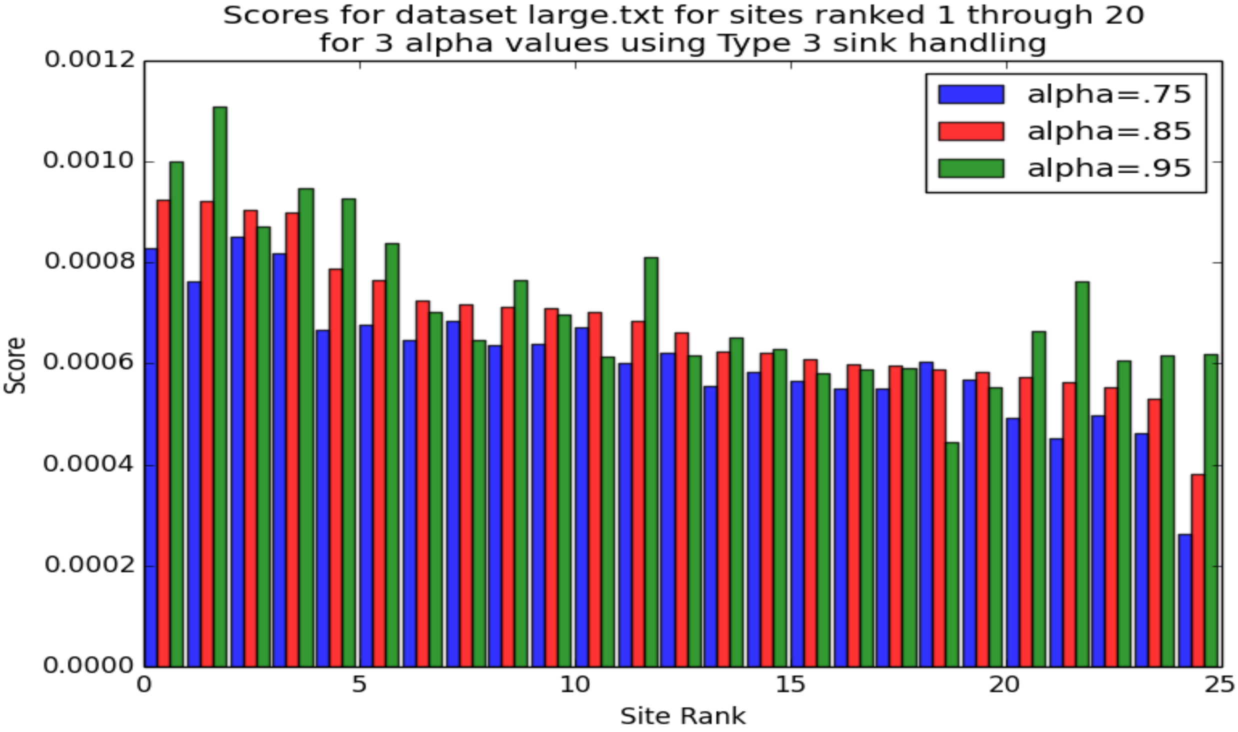
PageRank Project Report

Name: Chao Yang (cyang359)

Graph of results for 3 alphas for self-loop sink handling



Graph of results for 3 alphas for Type 3 sink handling



PageRank Project Report

Name: Chao Yang (cyang359)

Plots Discussion:

What are the results you see in the plots?

The results show the top 20+ site ranking in large.txt. With different Damping Factor(alpha), they have different score. For the type 3, the score is more even than Type 1(0.0004~0.001). On the type 1, the score is very high for first 5 sites. In conclusion, the higher-ranking site has the higher score.

What do the values represent?

In two types results, the higher-ranking site has the higher score. The X-axis shows the top 20+ sites' rank, and Y-axis shows each site's score. For each site, it has three columns (Blue, Red, Green) which match the three alpha values (0.75, 0.85, 0.95). The 0.95 alpha for major sites has the higher score.

What does increasing alpha toward 1.0 mean?

For the PageRank:

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

If the d(alpha) = 1.0, which means $\frac{1-d}{N}$ = 0, there will be no random selection link any more. The PageRank score only depend on the links provides by the site and they will always end up in a sink. This will give an infinite number of iterations to convergence.

Why do the results change with changing alpha?

From the equation:

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)} \qquad \frac{(1-\alpha)}{N} + \frac{\alpha}{|out(y)|}$$

The damping factor(alpha) allows a heuristic to offset the importance of those sinks. That is why the first term of the PageRank equation, (1−d)/N, is included. With the low damping factor will make calculations easier (The front part will take more proportion). Since the flow of PageRank is dampened the iterations will quickly converge. Also, low alpha means that the relative PageRank will be determined by PageRank received from external pages - rather than the internal link structure.

Why does changing alpha affect the runtime?

All clicks are random restarts, which are uniformly distributed (the 1/N coefficient in the first term) by definition. For the runtime, the smaller alpha will have low iteration times which could decrease the runtime. Because we calculate the computation ends when for some small ϵ:

$$|\mathbf{R}(t+1) - \mathbf{R}(t)| < \epsilon$$

When the alpha decrease, the PR will decrease. This makes we could easy to reach the ϵ.

How does the sink handling strategy impact the results and the runtime?

For large file (alpha = 0.85):

Type 1: 355s

Type 3: 310s

As the results show in the picture above: for the type 3, the score is more even than Type 1(0.0004~0.001). On the type 1, the score is higher for first 5 sites especially for alpha = 0.95. The reason for the type 3 is faster, I consider is I separate the calculation for sink node. However, the self-loop I need iterate all the node include the self-loop link which may make the time longer.