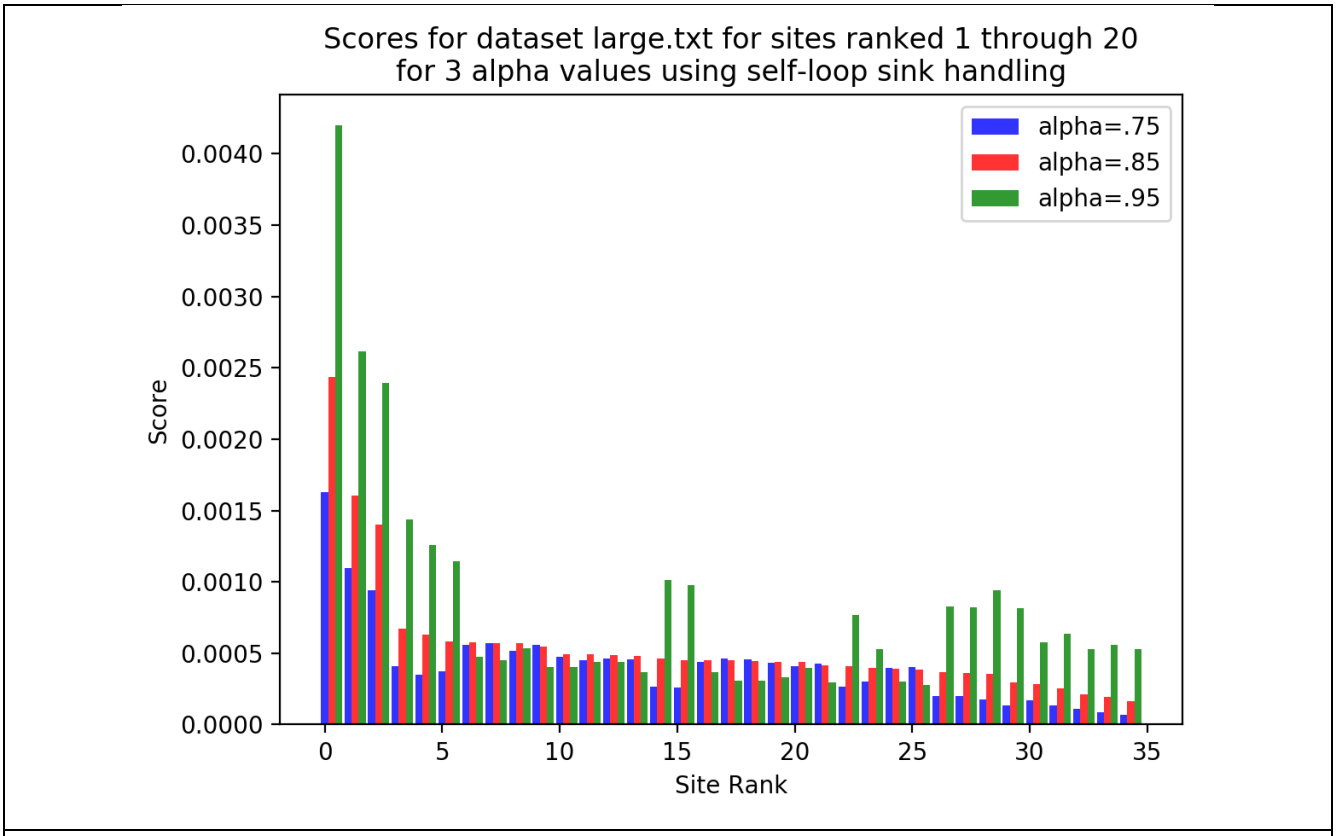
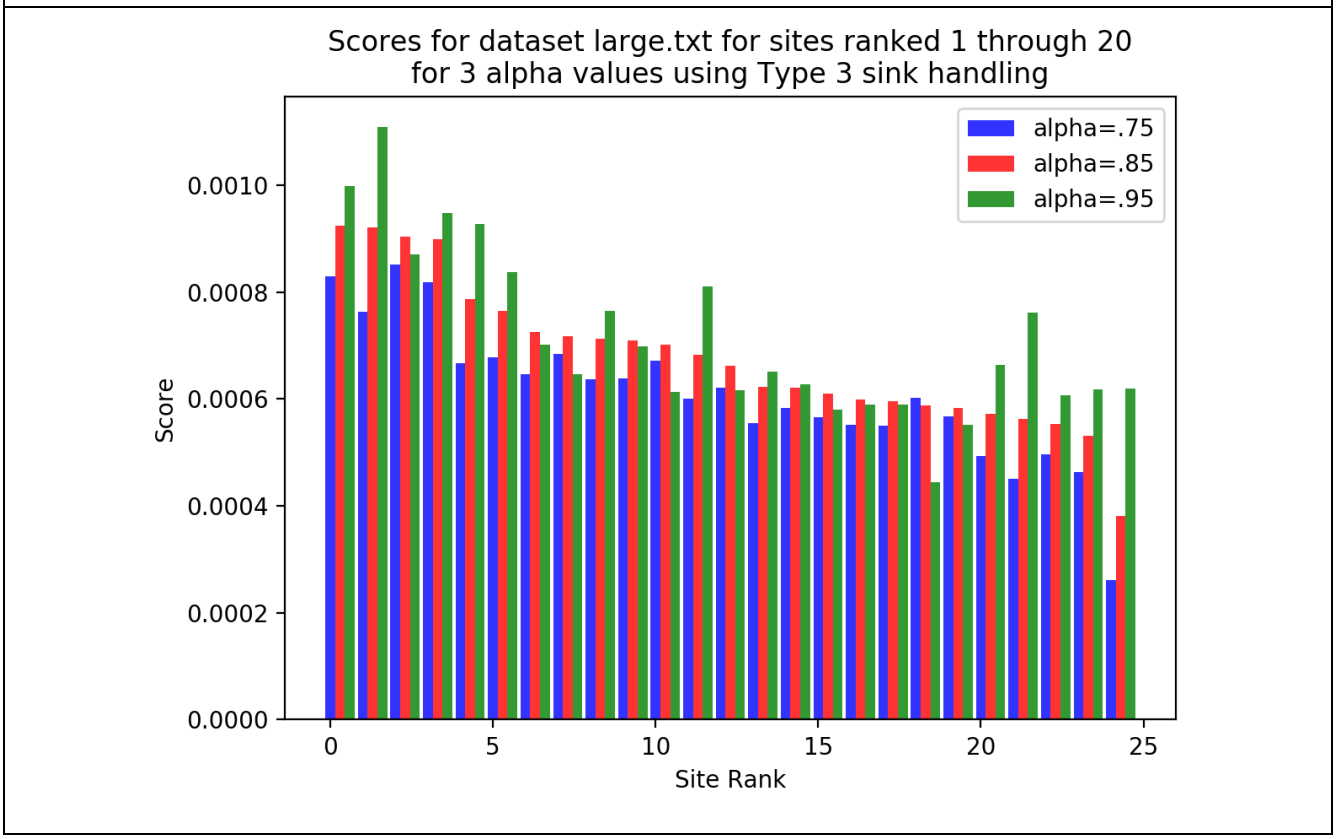
PageRank Project Report

Name : Maksim Kazakov





PageRank Project Report

Name : Maksim Kazakov

Plots Discussion :

Both plots above display PageRank scores for sites with top 20 score results for all three choices of alpha. On the plots themselves sites are placed in rank order based on the order of the scores calculated with alpha 0.85. That's why only scores for alpha 0.85 are displayed in strictly descending order.

The PageRank scores represent stationary distribution of the websites in network graph where websites are the nodes and links between them are directed edges. And as alpha parameter changes behavior of the random walk on which PageRank can be said is based upon it also changes transition probability matrix. As the result, scores or even ranking order for website can change for different choices of alpha.

The closer alpha is to 0 the more random transitions become (links between sites become less relevant) and closer alpha is to 1 the more reliant on the links random walk is. In both graphs we see an expected behavior when scores become closer to uniform distribution with lesser alpha values, which can degrade their predictive power and making them less informative in general. On the other hand, the closer alpha value is to 1 the more time it takes to compute scores as it takes more iterations for PageRank to converge to stable solution. This is because it takes more time for the influence of websites from one strongly connected component to propagate to other websites (or even just between components with a few links connecting them). In this regard alpha parameter in PageRank algorithm is similar to exploration/exploitation ratio in many Reinforcement learning algorithms.

Two plots are different in the way PageRank algorithm handles sink nodes (websites with no outgoing links). In the first case PageRank adds self-loops to nodes. The result of this is that nodes retain part of their score values at future iterations. This makes PageRank to converge faster than the second method where sink nodes are virtually connected to every other node, making their influence on themselves insignificant as it's divided among all the nodes in the graph. But self-loops also cause extreme scores to become even more extreme. As the result it can be seen on the first plot that scores of the top sites are usually visibly greater than scores of the websites with lower ranks.