CSC 391/691 Project 3 Overview
Due 4/17/14

This project will involve implementing a simplified PageRank algorithm.

Download from Sakai → Data two files named WFU_links.txt and WFU_URLs.txt.

WFU_URLs contains approximately 10,000 URLS that were crawled starting at http://www.wfu.edu/. Each URL has a unique integer identifier. You may or may not use this file for the bulk of the project.

WFU_links.txt contains approximately 10,000 "from→to" links stored in a very simple format. Each line contains *fffff:ttttt* where fffff and ttttt are integers corresponding to a URL in the URL file.

[NOTE: The data is also available as a single SQLite file on Sakai.]

1- Implement the simplified PageRank algorithm and identify the ten web pages with the highest ranking (report the URLs) – unless they all turn out to be 0 after 50 iterations. You will probably need to use a compact repre
2- Add a *teleportation* factor to prevent ranks from going to 0.
   a. With Beta set to 0.9, what are the ten web pages with the highest ranking (report the URLs).
   b. With Beta set to 0.8, what are the ten web pages with the highest ranking (report the URLs).
3- Using the top ten list from Part (2), select the highest ranked page to test for spam, as follows. Using the other 9 URLs as trusted pages, compute the *spam mass* of the highest ranked page.

Deliverables

Your answers from parts 1-3 along with a narrative description of how you arrived at those answers and any interesting observations during your exploration of the data.