Short Problem Set (Module 9)

Special note: This problem set is due at the end of Module 10.

1. [30%] Give a short definition or explanation of the following concepts:
   - web spam
   - Broder's taxonomy
   - out-degree
   - robots exclusion protocol
   - priority queue (in the context of web crawling)

2. [20%] Describe in your own words the process described in the course text to efficiently identify near duplicate documents in a large collection.

3. For this problem work with the directed web graph shown below. In the graph there are six nodes: Y, B, F, G, T, R (for the websites Yahoo, Bing, Facebook, Google, Twitter, and Reddit). Use a teleport probability of 0.20. Assume no other pages or links exist beside those shown in the figure.
(a) [15%] Provide (i.e., write) the six recurrence equations that indicate how to iteratively calculate the PageRank score of each page at time t given scores from time t-1.
(b) [25%] Using the brute-force iterative method of calculation shown in the video lecture calculate two iterations of PageRank scores for each page in the graph. Be sure to show scores at times t=0, t=1, and finally at t=2. Report scores using three digits of precision (e.g., 0.247, not 0.2 or 0.24696485932). Show work and do not merely provide a table of values.
(c) [5%] Which page (or pages) has/have the lowest PageRank score after two iterations?
(d) [5%] Which page (or pages) has/have the highest PageRank score after two iterations?