Short Problem Set 9

Max Robinson

1 Problem 1

Question: Describe in your own words the process described in the course text to efficiently identify near duplicate documents in a large collection.

Answer:

2 Problem 2

Question: Give a short definition or explanation of the following concepts:

1. web spam
2. near duplicate
3. in-degree
4. robot exclusion protocol
5. priority queue (in the context of web crawling)

Answer:

3 Problem 3

Question: For this problem work with the directed web graph shown below. In the graph there are six nodes: S, B, F, G, T, and I (for the websites Snapchat, Bing, Facebook, Google, Twitter, and Instagram). Use a teleport probability of 0.25. Assume no other pages or links exist beside those shown in the figure.