LIS 544 / IMT 542 / INSC 544
Assignment: Web Crawlers
Due: November 22

Long Answer

1. Compare the features and crawl architecture of four publicly available web crawlers. Be sure to examine how the crawler handles duplicate detection, URL normalization, archiving (if it's an feature), politeness (robots.txt and throttling), distributed crawling, and DNS cacheing. Also compare each crawler to the idealized crawling process presented in class. You may compare more features if you choose to do so.

   We've discussed a number of crawlers in class. You can also find a nice list of crawlers on Wikipedia. Feel free to examine any crawler with enough documentation. Be sure to document your sources!

2. Discuss the ethics of ignoring a site's robots.txt file. Is it ever appropriate for a crawler to ignore this file? Why or why not? If so, give and example of when this would be ok.

Short Answer

3. What is URL normalization? Explain the process, its purpose, and give an example.

4. How do you tell if your web crawler is stuck in a loop or spam trap (continually loading the same page)? (Extra credit for more than one method)

5. List three constraints on the speed of a crawler. Can you suggest solutions to mitigate these constraints?