Patrick Sheehan
Final Report
TAMU CSCE 463 - Homework 1
github.com/pas0185/csce463-hw1
11 February 2015

1. The overall code architecture for this project is as follows:
    a. Parse command line arguments for N threads and an input file or URLs
    b. Initialize dedicated Parameters class which contains:
        i. A mutex for regulating single thread access to critical sections
        ii. A semaphore used to regulate shared access to urlQueue by...
            1. Producer - FileReader (places URLs in queue as they are parsed)
            2. Consumers - Crawler threads (remove and parse URLs as they become available in the queue
        iii. An event to signal when all URLs have finished being processed
    c. Spawn File Thread
        i. Read input file and safely place URLs into the shared queue one by one
        ii. Must do it within mutex lock because other threads are using this queue
        iii. Release semaphore when done to let crawlers know another URL is available
    d. Spawn Statistics Thread
        i. Every 2 seconds, print status (within mutex b/c of printf)
        ii. When the quit event is triggered, it will print the final report
    e. Spawn N Crawler Threads
        i. Removes URLs from the queue as they become available to parse them
        ii. A semaphore is waited on to know when at least one is available
        iii. Also listens to the event for when the file finishes being read. When no URLs in the queue and the file is done being read, the event will be set indicating that the task is finished
    f. Wait until all threads finish before exiting main thread

    <SHOW 1M TRACE>

2. The average number of links per HTML page is 54.36
   If Google's web-graph has 1 trillion nodes, and each node is adjacent to roughly 54 other nodes. Then the adjacency list, with 55 trillion nodes (each represented by a 64-bit URL) would occupy 3.52 * 10^15 bits (409781 TB)

3. The average HTML page size was 28.2 bytes
   For yahoo to crawl 10 billion web pages per day, the total number of bytes would be 282 billion (268936 GB) per 86400 seconds in a day. In other words, the bandwidth

needed would be 3.112 GBps

4. The probability that a link in the input file contains a unique host was 13.93%
   The probability that a host has a valid DNS record was 13.93%
   The percentage of sites with a 4xx page was 11.17%

5. The number of crawled 2xx pages with a link to tamu.edu domain was _____