Name: Phong Tran

Class: CSCE 463

**HW1 Report**

* 1. Code architecture:
     1. Socket: a wrapper class for SOCKET. It is responsible for connecting to the host, loading response, and verifying header.
     2. FileReader: helper class for reading file. It can extract text from a file line by line.
     3. URLEntry: struct to store basic information of an url
     4. URLParser: parses an url the user provides and create an object of type URLEntry
     5. Crawler: responsible for crawling a page
     6. SinglePageCrawler: part 1 of the homework
     7. SingleThreadedCrawler: part 2 of the homework
     8. MultiThreadedCrawler: part 3 of the homework
     9. Utils: contains helper functions
  2. Lessons learned:
     1. How to use strncpy, strstr, strchr
     2. Pointer manipulation
     3. strncpy might cause buffer overflow
     4. Winsock, thread, semaphore in Windows
     5. Always remember to release mutex, especially in if statements which break out of a loop
     6. Separate host and IP into different unordered\_set
  3. Complete trace: done in x64 Release mode using 5000 threads on URL-input-1M.txt:

Main thread: read file with size 66152007

[ 2] 5000 Q 1 E 63670 H 10019 D 9620 I 7315 R 2923 C 1642 L 46K

\*\*\* crawling 821.0 pps @ 98.6 Mbps

[ 4] 5000 Q 717484 E 100213 H 15045 D 14760 I 11717 R 5384 C 3405 L 99K

\*\*\* crawling 881.5 pps @ 106.0 Mbps

[ 6] 5000 Q 868972 E 131032 H 19566 D 19188 I 15587 R 8142 C 5784 L 197K

\*\*\* crawling 1189.0 pps @ 188.6 Mbps

[ 8] 5000 Q 840345 E 159659 H 23993 D 23535 I 19357 R 10628 C 8300 L 309K

\*\*\* crawling 1257.5 pps @ 220.4 Mbps

[ 10] 5000 Q 805489 E 194515 H 28314 D 27829 I 23048 R 13082 C 10620 L 427K

\*\*\* crawling 1159.5 pps @ 210.1 Mbps

[ 12] 5000 Q 774835 E 225169 H 32514 D 31917 I 26552 R 15451 C 12912 L 535K

\*\*\* crawling 1146.0 pps @ 208.2 Mbps

[ 14] 5000 Q 749221 E 250783 H 36398 D 35806 I 29801 R 17456 C 14997 L 637K

\*\*\* crawling 1042.5 pps @ 194.6 Mbps

[ 16] 5000 Q 728247 E 271757 H 39890 D 39230 I 32706 R 19311 C 16925 L 745K

\*\*\* crawling 964.0 pps @ 210.0 Mbps

[ 18] 5000 Q 707415 E 292589 H 43537 D 42861 I 35749 R 21241 C 18980 L 858K

\*\*\* crawling 1027.5 pps @ 236.8 Mbps

[ 20] 5000 Q 687973 E 312031 H 46869 D 46123 I 38486 R 23116 C 20805 L 951K

\*\*\* crawling 912.0 pps @ 184.2 Mbps

[ 22] 5000 Q 687973 E 312031 H 46869 D 46123 I 38486 R 23116 C 20805 L 951K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[ 24] 5000 Q 687973 E 312031 H 46869 D 46123 I 38486 R 23116 C 20805 L 951K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[ 26] 5000 Q 687973 E 312031 H 46869 D 46123 I 38486 R 23116 C 20805 L 951K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[ 28] 5000 Q 686738 E 313266 H 46912 D 46297 I 38625 R 24294 C 22712 L 1028K

\*\*\* crawling 953.5 pps @ 570.5 Mbps

[ 30] 5000 Q 643486 E 356518 H 54094 D 52661 I 43972 R 26827 C 24679 L 1313K

\*\*\* crawling 983.5 pps @ 150.9 Mbps

[ 32] 5000 Q 602903 E 397101 H 59195 D 58110 I 48467 R 29721 C 27269 L 1404K

\*\*\* crawling 1294.5 pps @ 174.6 Mbps

[ 34] 5000 Q 561716 E 438288 H 64316 D 63160 I 52567 R 32466 C 29928 L 1519K

\*\*\* crawling 1329.0 pps @ 231.5 Mbps

[ 36] 5000 Q 531045 E 468959 H 68490 D 67096 I 55765 R 34659 C 32164 L 1624K

\*\*\* crawling 1118.0 pps @ 208.1 Mbps

[ 38] 5000 Q 490532 E 509472 H 72512 D 70921 I 58866 R 36788 C 34401 L 1747K

\*\*\* crawling 1118.0 pps @ 235.0 Mbps

[ 40] 5000 Q 464458 E 535546 H 75579 D 73864 I 61168 R 38381 C 36016 L 1834K

\*\*\* crawling 807.0 pps @ 165.7 Mbps

[ 42] 5000 Q 464458 E 535546 H 75579 D 73864 I 61168 R 38381 C 36016 L 1834K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[ 44] 5000 Q 463688 E 536316 H 75635 D 74649 I 61793 R 39298 C 37525 L 1973K

\*\*\* crawling 754.5 pps @ 315.0 Mbps

[ 46] 5000 Q 432695 E 567309 H 79902 D 78433 I 64530 R 40964 C 39270 L 2111K

\*\*\* crawling 872.5 pps @ 219.0 Mbps

[ 48] 5000 Q 385673 E 614331 H 86772 D 84332 I 68329 R 43410 C 41263 L 2176K

\*\*\* crawling 996.0 pps @ 128.4 Mbps

[ 50] 5000 Q 342973 E 657031 H 92635 D 89895 I 71300 R 45302 C 43444 L 2292K

\*\*\* crawling 1089.5 pps @ 229.0 Mbps

[ 52] 5000 Q 334605 E 665399 H 93854 D 91182 I 71971 R 45692 C 43818 L 2313K

\*\*\* crawling 186.5 pps @ 37.4 Mbps

[ 54] 5000 Q 299049 E 700955 H 98978 D 96930 I 74996 R 47753 C 46215 L 2466K

\*\*\* crawling 1198.5 pps @ 298.0 Mbps

[ 56] 5000 Q 246671 E 753333 H 106385 D 104568 I 78826 R 50018 C 48399 L 2577K

\*\*\* crawling 1091.5 pps @ 229.6 Mbps

[ 58] 5000 Q 195206 E 804798 H 113918 D 111386 I 82102 R 52243 C 50619 L 2706K

\*\*\* crawling 1110.0 pps @ 254.3 Mbps

[ 60] 5000 Q 162779 E 837225 H 118833 D 115107 I 83913 R 53479 C 51916 L 2779K

\*\*\* crawling 648.5 pps @ 159.3 Mbps

[ 62] 5000 Q 121332 E 878672 H 124087 D 120348 I 86443 R 55187 C 53610 L 2873K

\*\*\* crawling 847.0 pps @ 204.6 Mbps

[ 64] 5000 Q 75248 E 924756 H 131004 D 127447 I 89897 R 57079 C 55515 L 2990K

\*\*\* crawling 952.0 pps @ 215.2 Mbps

[ 66] 5000 Q 27446 E 972558 H 137355 D 134171 I 93182 R 59038 C 57218 L 3057K

\*\*\* crawling 851.5 pps @ 136.5 Mbps

[ 68] 2054 Q 0 E 1000004 H 139301 D 137610 I 94940 R 60280 C 58679 L 3163K

\*\*\* crawling 730.0 pps @ 189.8 Mbps

[ 70] 1402 Q 0 E 1000004 H 139301 D 137655 I 94964 R 60481 C 59080 L 3218K

\*\*\* crawling 200.5 pps @ 89.2 Mbps

[ 72] 1005 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60622 C 59343 L 3245K

\*\*\* crawling 131.5 pps @ 63.1 Mbps

[ 74] 728 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60699 C 59516 L 3264K

\*\*\* crawling 86.5 pps @ 32.1 Mbps

[ 76] 494 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60721 C 59670 L 3279K

\*\*\* crawling 77.0 pps @ 31.2 Mbps

[ 78] 359 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60725 C 59727 L 3286K

\*\*\* crawling 28.5 pps @ 13.1 Mbps

[ 80] 282 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59749 L 3292K

\*\*\* crawling 11.0 pps @ 6.4 Mbps

[ 82] 251 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59756 L 3299K

\*\*\* crawling 3.5 pps @ 4.4 Mbps

[ 84] 130 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59785 L 3303K

\*\*\* crawling 14.5 pps @ 5.3 Mbps

[ 86] 66 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59799 L 3306K

\*\*\* crawling 7.0 pps @ 11.4 Mbps

[ 88] 29 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59805 L 3307K

\*\*\* crawling 3.0 pps @ 14.7 Mbps

[ 90] 20 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59805 L 3307K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[ 92] 19 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59805 L 3307K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[ 94] 17 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59805 L 3307K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[ 96] 8 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59809 L 3315K

\*\*\* crawling 2.0 pps @ 3.9 Mbps

[ 98] 7 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59809 L 3315K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[100] 7 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59809 L 3315K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[102] 5 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59810 L 3315K

\*\*\* crawling 0.5 pps @ 0.4 Mbps

[104] 5 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59810 L 3315K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[106] 5 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59810 L 3315K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[108] 5 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59810 L 3315K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[110] 5 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59810 L 3315K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[112] 5 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59810 L 3315K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[114] 4 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59811 L 3317K

\*\*\* crawling 0.5 pps @ 1.0 Mbps

[116] 4 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59811 L 3317K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[118] 4 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59811 L 3317K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[120] 4 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59811 L 3317K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[122] 4 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59811 L 3317K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[124] 3 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59812 L 3318K

\*\*\* crawling 0.5 pps @ 0.5 Mbps

[126] 3 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59812 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[128] 2 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59813 L 3318K

\*\*\* crawling 0.5 pps @ 0.0 Mbps

[130] 2 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59813 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[132] 2 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59813 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[134] 2 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59813 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[136] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.5 pps @ 3.1 Mbps

[138] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[140] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[142] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[144] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[146] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[148] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[150] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[152] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[154] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[156] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[158] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[160] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[162] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[164] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[166] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[168] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[170] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[172] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[174] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[176] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[178] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[180] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[182] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[184] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[186] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[188] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[190] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[192] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[194] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[196] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[198] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[200] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[202] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[204] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[206] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[208] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[210] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[212] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[214] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[216] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[218] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[220] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[222] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[224] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[226] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[228] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[230] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[232] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[234] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[236] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[238] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[240] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[242] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[244] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[246] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[248] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[250] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[252] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[254] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[256] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

[258] 1 Q 0 E 1000004 H 139301 D 137657 I 94966 R 60726 C 59814 L 3318K

\*\*\* crawling 0.0 pps @ 0.0 Mbps

Extracted 1000004 URLs @ 3846/s

Looked up 137657 DNS names @ 529/s

Downloaded 94966 robots @ 365/s

Crawled 59814 pages @ 230/s (1622.73 MB)

Parsed 3318307 links @ 12763/s

HTTP codes: 2xx = 47010, 3xx = 5909, 4xx = 6683, 5xx = 212, other = 0

* 1. Average number of links per HTML page (2xx) is 3318307 /47010= 71 links
  2. Google webgraph: there are 1 trillion crawled nodes and each one is represented by a 64-bit hash, thus 64 Tb. Every edge is a pointer to another node. I assume it is 16 bit. Assuming that each node has an average of 71 edges. Thus, the total of:
     1. Edges = 1T \* 71 = 71T
     2. Edge Size = 71T \* 16 = 1136 Tb
     3. Bytes = 320Tb + 64Tb = 1200 Tb

1. Average page size = 1622.73MB \* 106/59814 = 27129.6 Bytes = 27.13 KB

10B pages/day = 10B \* 27.13KB \* 1000 \* 8 = 2170400 Gb/day

Bandwidth = 2170400/86400 = 25.12 Gbps

* 1. Number of unique host: 139301
  2. Probability of being an unique host: 139301/1000004 = 0.14 = 14%
  3. Probability of an unique host having a valid DNS: 137657 /139301 = 0.99 = 99%
  4. Percentage of contacted sites having 4xx robots file: 60726/94966 = 0.64 = 64%

1. There are 79 pages that contain the domain tamu.edu. I did it by using the provided HTMLParserBase to get a list of hyperlinks within the crawled page. I then wrote a function that loops the list and checks if there is one link that has the domain tamu.edu. If there is at least one, the function sets the global variable hasTamuDomain to true. Otherwise, it remains false. For every page crawled, it calls that function and increments the global variable tamuCount if it returns true.