Venkata Satya Kavya Sree Bondapalli
CSCE 612: HW1 Report

Note: My program can do HTTP 1.1 downloads.

#1

- **Code Architecture:**
    a) In main(): Argument parsing and calling the respective function that deals with p1 or p2 or p3 is done.
    b) Common features for Part 1, Part 2, Part 3:
        i) Url parsing: achieved by string manipulations
        ii) DNS lookup: done by gethostbyname()
        iii) Connecting to the page: open a socket and send a http request
        iv) Loading the page: listen on the socket to receive the response
        v) Verifying the header: checking if the header starts with "HTTP/" or not
        vi) Parsing the page: to get the links using HTMLParserBase
    c) Add on features for Part 2 & Part 3:
        i) Host uniqueness, IP uniqueness: done by maintaining separate unordered set for each and checking for if the current host is already available in the database or not
        ii) Connect on robots and check if the status code is 4xx, then proceed with connecting to the page
    d) Part 3:
        i) Initialized the parameters to be shared in common to the threads like
            (1) Mutex: to ensure that only one thread at a time accesses a critical section
            (2) Semaphore: to maintain the urls queue access to both the consumer and producer
            (3) A quit event to signal once all urls have been processed
            (4) Other data variables to store the count
        ii) Read line by line from the file and push the url into the common queue.
        iii) Created the crawling threads and stats thread
        iv) Define the callback functions for thread crawler and for running the stats.
        v) The queue is populated and extracted by using Producer-Consumer technique.Once a url is extracted we repeat the steps in p2(parsing,dns,uniqueness,robots check,crawl the page for links)
- **Lessons Learnt**:
    a) Should take care to null terminate the buffer
    b) Null checks are needed
    c) Url, host, length max checks are needed
    d) HTMLParserBase object should be created once at the start of the thread and use it for all further urls taken care by the same thread
    e) Should make sure lock and unlock of mutex is done in a proper manner whenever accessing critical sections

- **Complete Trace of 1M input URLs:**

```
Opened URL-input-1M.txt with size 66152005
[  2] 5000 Q 934490 E   65514 H  10161 D   9071 I  6450 R  1323 C   117 L    1K
        *** crawling 58.5 pps @ 7.1 Mbps
[  4] 5000 Q 862872 E  137132 H  20448 D  18621 I 14198 R  3280 C   645 L    7K
        *** crawling 264.0 pps @ 16.2 Mbps
[  6] 5000 Q 820189 E  179815 H  26488 D  24242 I 18620 R  4815 C  1543 L   26K
        *** crawling 449.0 pps @ 38.9 Mbps
[  8] 5000 Q 773986 E  226018 H  32614 D  30014 I 23140 R  6283 C  2545 L   58K
        *** crawling 501.0 pps @ 59.9 Mbps
[ 10] 5000 Q 737339 E  262665 H  38343 D  35427 I 27319 R  7604 C  3532 L   87K
        *** crawling 493.5 pps @ 56.7 Mbps
[ 12] 5000 Q 705258 E  294746 H  43901 D  40718 I 31355 R  8913 C  4520 L  126K
        *** crawling 494.0 pps @ 69.4 Mbps
[ 14] 5000 Q 670846 E  329158 H  49629 D  46290 I 35546 R 10306 C  5544 L  159K
        *** crawling 512.0 pps @ 70.6 Mbps
[ 16] 5000 Q 631479 E  368525 H  55618 D  52079 I 39905 R 11726 C  6574 L  197K
        *** crawling 515.0 pps @ 67.6 Mbps
[ 18] 5000 Q 577333 E  422671 H  62190 D  58385 I 44516 R 13167 C  7666 L  240K
        *** crawling 546.0 pps @ 91.8 Mbps
[ 20] 5000 Q 536707 E  463297 H  67752 D  63693 I 48428 R 14416 C  8622 L  271K
        *** crawling 478.0 pps @ 72.0 Mbps
[ 22] 5000 Q 531509 E  468495 H  68426 D  64384 I 48926 R 14436 C  8637 L  273K
        *** crawling 7.5 pps @ 82.8 Mbps
[ 24] 5000 Q 524060 E  475944 H  69416 D  65372 I 49634 R 14453 C  8649 L  273K
        *** crawling 6.0 pps @ 17.6 Mbps
[ 26] 5000 Q 517602 E  482402 H  70112 D  66021 I 50101 R 14474 C  8670 L  275K
        *** crawling 10.5 pps @ 12.8 Mbps
[ 28] 5000 Q 512608 E  487396 H  70608 D  66498 I 50445 R 14492 C  8675 L  275K
        *** crawling 2.5 pps @ 0.4 Mbps
[ 30] 5000 Q 507660 E  492344 H  71060 D  66932 I 50757 R 14510 C  8681 L  276K
        *** crawling 3.0 pps @ 1.1 Mbps
[ 32] 5000 Q 503356 E  496648 H  71437 D  67288 I 51028 R 14538 C  8686 L  276K
        *** crawling 2.5 pps @ 1.0 Mbps
[ 34] 5000 Q 499055 E  500949 H  71837 D  67658 I 51303 R 14560 C  8690 L  276K
        *** crawling 2.0 pps @ 0.7 Mbps
[ 36] 5000 Q 493214 E  506790 H  72262 D  68057 I 51582 R 14573 C  8692 L  276K
        *** crawling 1.0 pps @ 0.5 Mbps
[ 38] 5000 Q 489390 E  510614 H  72637 D  68409 I 51828 R 14587 C  8698 L  276K
        *** crawling 3.0 pps @ 0.4 Mbps
[ 40] 5000 Q 485676 E  514328 H  73024 D  68777 I 52085 R 14599 C  8700 L  276K
        *** crawling 1.0 pps @ 0.5 Mbps
[ 42] 5000 Q 483029 E  516975 H  73310 D  69049 I 52276 R 14628 C  8719 L  277K
        *** crawling 9.5 pps @ 0.2 Mbps
[ 44] 5000 Q 480363 E  519641 H  73576 D  69294 I 52432 R 14669 C  8720 L  277K
        *** crawling 0.5 pps @ 0.7 Mbps
```

```
[ 46] 5000 Q 477831 E  522173 H  73916 D  69613 I 52645 R 14736 C  8741 L  279K
        *** crawling 10.5 pps @ 1.5 Mbps
[ 48] 5000 Q 474583 E  525421 H  74304 D  69985 I 52912 R 14815 C  8750 L  280K
        *** crawling 4.5 pps @ 1.6 Mbps
[ 50] 5000 Q 472250 E  527754 H  74603 D  70268 I 53104 R 14883 C  8771 L  282K
        *** crawling 10.5 pps @ 3.0 Mbps
[ 52] 5000 Q 469491 E  530513 H  74977 D  70617 I 53338 R 14952 C  8791 L  285K
        *** crawling 10.0 pps @ 2.6 Mbps
[ 54] 5000 Q 467605 E  532399 H  75204 D  70827 I 53489 R 14995 C  8806 L  289K
        *** crawling 7.5 pps @ 1.8 Mbps
[ 56] 5000 Q 465668 E  534336 H  75431 D  71040 I 53645 R 15056 C  8812 L  289K
        *** crawling 3.0 pps @ 0.9 Mbps
[ 58] 5000 Q 462645 E  537359 H  75799 D  71391 I 53870 R 15128 C  8835 L  291K
        *** crawling 11.5 pps @ 1.9 Mbps
[ 60] 5000 Q 461040 E  538964 H  75968 D  71551 I 53978 R 15163 C  8842 L  291K
        *** crawling 3.5 pps @ 1.5 Mbps
[ 62] 5000 Q 458998 E  541006 H  76235 D  71811 I 54165 R 15219 C  8865 L  294K
        *** crawling 11.5 pps @ 2.7 Mbps
[ 64] 5000 Q 456665 E  543339 H  76505 D  72065 I 54351 R 15283 C  8875 L  295K
        *** crawling 5.0 pps @ 0.9 Mbps
[ 66] 5000 Q 454885 E  545119 H  76735 D  72285 I 54490 R 15325 C  8882 L  295K
        *** crawling 3.5 pps @ 1.1 Mbps
[ 68] 5000 Q 453034 E  546970 H  76957 D  72501 I 54636 R 15372 C  8892 L  296K
        *** crawling 5.0 pps @ 1.2 Mbps
[ 70] 5000 Q 451426 E  548578 H  77176 D  72700 I 54778 R 15425 C  8913 L  298K
        *** crawling 10.5 pps @ 2.1 Mbps
[ 72] 5000 Q 450256 E  549748 H  77321 D  72839 I 54878 R 15475 C  8917 L  298K
        *** crawling 2.0 pps @ 1.7 Mbps
[ 74] 5000 Q 447956 E  552048 H  77650 D  73148 I 55083 R 15543 C  8945 L  301K
        *** crawling 14.0 pps @ 2.7 Mbps
[ 76] 5000 Q 445690 E  554314 H  77996 D  73465 I 55300 R 15614 C  8970 L  303K
        *** crawling 12.5 pps @ 2.0 Mbps
[ 78] 5000 Q 444480 E  555524 H  78188 D  73646 I 55408 R 15646 C  8982 L  304K
        *** crawling 6.0 pps @ 2.2 Mbps
[ 80] 5000 Q 443534 E  556470 H  78327 D  73777 I 55499 R 15679 C  8987 L  304K
        *** crawling 2.5 pps @ 0.4 Mbps
[ 82] 5000 Q 441709 E  558295 H  78594 D  74034 I 55675 R 15732 C  9004 L  306K
        *** crawling 8.5 pps @ 1.2 Mbps
[ 84] 5000 Q 439058 E  560946 H  79029 D  74445 I 55933 R 15795 C  9034 L  307K
        *** crawling 15.0 pps @ 0.9 Mbps
[ 86] 5000 Q 436824 E  563180 H  79383 D  74782 I 56157 R 15876 C  9058 L  311K
        *** crawling 12.0 pps @ 2.4 Mbps
[ 88] 5000 Q 434363 E  565641 H  79772 D  75155 I 56400 R 15953 C  9069 L  312K
        *** crawling 5.5 pps @ 3.0 Mbps
[ 90] 5000 Q 431708 E  568296 H  80252 D  75593 I 56675 R 16039 C  9108 L  315K
        *** crawling 19.5 pps @ 2.0 Mbps
```

```
[ 92] 5000 Q 429540 E  570464 H  80619 D  75937 I 56875 R 16122 C  9125 L  319K
        *** crawling 8.5 pps @ 1.5 Mbps
[ 94] 5000 Q 427272 E  572732 H  80977 D  76278 I 57084 R 16193 C  9138 L  319K
        *** crawling 6.5 pps @ 1.2 Mbps
[ 96] 5000 Q 424803 E  575201 H  81349 D  76620 I 57304 R 16265 C  9163 L  320K
        *** crawling 12.5 pps @ 2.2 Mbps
[ 98] 5000 Q 421421 E  578583 H  81878 D  77103 I 57581 R 16374 C  9202 L  324K
        *** crawling 19.5 pps @ 3.5 Mbps
[100] 5000 Q 419099 E  580905 H  82226 D  77433 I 57777 R 16456 C  9222 L  324K
        *** crawling 10.0 pps @ 3.7 Mbps
[102] 5000 Q 415890 E  584114 H  82637 D  77827 I 58011 R 16535 C  9256 L  327K
        *** crawling 17.0 pps @ 2.7 Mbps
[104] 5000 Q 412284 E  587720 H  83101 D  78243 I 58267 R 16617 C  9282 L  327K
        *** crawling 13.0 pps @ 4.8 Mbps
[106] 5000 Q 409046 E  590958 H  83489 D  78613 I 58480 R 16681 C  9305 L  329K
        *** crawling 11.5 pps @ 1.7 Mbps
[108] 5000 Q 406443 E  593561 H  83852 D  78958 I 58674 R 16743 C  9329 L  330K
        *** crawling 12.0 pps @ 2.7 Mbps
[110] 5000 Q 403086 E  596918 H  84326 D  79387 I 58895 R 16825 C  9352 L  332K
        *** crawling 11.5 pps @ 0.8 Mbps
[112] 5000 Q 399198 E  600806 H  84875 D  79905 I 59176 R 16909 C  9376 L  334K
        *** crawling 12.0 pps @ 3.3 Mbps
[114] 5000 Q 395219 E  604785 H  85431 D  80429 I 59432 R 16997 C  9405 L  335K
        *** crawling 14.5 pps @ 1.5 Mbps
[116] 5000 Q 391963 E  608041 H  85933 D  80899 I 59649 R 17071 C  9426 L  337K
        *** crawling 10.5 pps @ 3.8 Mbps
[118] 5000 Q 388510 E  611494 H  86413 D  81353 I 59845 R 17140 C  9449 L  339K
        *** crawling 11.5 pps @ 1.6 Mbps
[120] 5000 Q 383178 E  616826 H  87110 D  82008 I 60152 R 17242 C  9474 L  341K
        *** crawling 12.5 pps @ 3.7 Mbps
[122] 5000 Q 332075 E  667929 H  94020 D  88073 I 62903 R 19144 C 10265 L  383K
        *** crawling 395.5 pps @ 21.3 Mbps
[124] 5000 Q 255990 E  744014 H 104985 D  98211 I 67507 R 20434 C 10693 L  391K
        *** crawling 214.0 pps @ 16.9 Mbps
[126] 5000 Q 178303 E  821701 H 116440 D 108883 I 72007 R 21830 C 11089 L  401K
        *** crawling 198.0 pps @ 18.8 Mbps
[128] 5000 Q 114656 E  885348 H 125015 D 117109 I 75425 R 22924 C 11442 L  408K
        *** crawling 176.5 pps @ 17.3 Mbps
[130] 5000 Q  64359 E  935645 H 132527 D 124357 I 78484 R 23860 C 11774 L  415K
        *** crawling 166.0 pps @ 22.8 Mbps
[132] 5000 Q   3714 E  996290 H 139300 D 130834 I 81295 R 24771 C 12048 L  420K
        *** crawling 137.0 pps @ 9.9 Mbps
[134] 2362 Q      0 E 1000004 H 139300 D 130965 I 81340 R 24919 C 12198 L  425K
        *** crawling 75.0 pps @ 8.8 Mbps
[136] 2015 Q      0 E 1000004 H 139300 D 131040 I 81366 R 24988 C 12245 L  427K
        *** crawling 23.5 pps @ 4.9 Mbps
```

```
[138] 1835 Q      0 E 1000004 H 139300 D 131049 I 81372 R 24995 C 12300 L  430K
          *** crawling 27.5 pps @ 6.8 Mbps
[140] 1714 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24996 C 12333 L  431K
          *** crawling 16.5 pps @ 3.6 Mbps
[142] 1589 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24996 C 12342 L  432K
          *** crawling 4.5 pps @ 0.9 Mbps
[144] 1274 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12345 L  432K
          *** crawling 1.5 pps @ 0.4 Mbps
[146]  937 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12346 L  432K
          *** crawling 0.5 pps @ 0.0 Mbps
[148]  645 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12348 L  432K
          *** crawling 1.0 pps @ 0.1 Mbps
[150]  397 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12348 L  432K
          *** crawling 0.0 pps @ 0.0 Mbps
[152]  177 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12348 L  432K
          *** crawling 0.0 pps @ 0.0 Mbps
[154]   47 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12349 L  432K
          *** crawling 0.5 pps @ 0.0 Mbps
[156]   44 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12349 L  432K
          *** crawling 0.0 pps @ 0.0 Mbps
[158]   41 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
          *** crawling 0.5 pps @ 0.0 Mbps
[160]   41 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
          *** crawling 0.0 pps @ 0.0 Mbps
[162]   40 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
          *** crawling 0.0 pps @ 0.0 Mbps
[164]   31 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
          *** crawling 0.0 pps @ 0.0 Mbps
[166]   29 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
          *** crawling 0.0 pps @ 0.0 Mbps
[168]   29 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
          *** crawling 0.0 pps @ 0.0 Mbps
[170]   29 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
          *** crawling 0.0 pps @ 0.0 Mbps
[172]   29 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
          *** crawling 0.0 pps @ 0.0 Mbps
[174]   29 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
          *** crawling 0.0 pps @ 0.0 Mbps
[176]   25 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
          *** crawling 0.0 pps @ 0.0 Mbps
[178]   25 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
          *** crawling 0.0 pps @ 0.0 Mbps
[180]   23 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
          *** crawling 0.0 pps @ 0.0 Mbps
[182]   23 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
          *** crawling 0.0 pps @ 0.0 Mbps
```

```
[184]   22 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
         *** crawling 0.0 pps @ 0.0 Mbps
[186]   20 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
         *** crawling 0.0 pps @ 0.0 Mbps
[188]   20 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
         *** crawling 0.0 pps @ 0.0 Mbps
[190]   20 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
         *** crawling 0.0 pps @ 0.0 Mbps
[192]   20 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
         *** crawling 0.0 pps @ 0.0 Mbps
[194]   19 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
         *** crawling 0.0 pps @ 0.0 Mbps
[196]   18 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
         *** crawling 0.0 pps @ 0.0 Mbps
[198]   18 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
         *** crawling 0.0 pps @ 0.0 Mbps
[200]   18 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
         *** crawling 0.0 pps @ 0.0 Mbps
[202]   18 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
         *** crawling 0.0 pps @ 0.0 Mbps
[204]   18 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
         *** crawling 0.0 pps @ 0.0 Mbps
[206]   12 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
         *** crawling 0.0 pps @ 0.0 Mbps
[208]   12 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
         *** crawling 0.0 pps @ 0.0 Mbps
[210]   12 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
         *** crawling 0.0 pps @ 0.0 Mbps
[212]   10 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
         *** crawling 0.0 pps @ 0.0 Mbps
[214]   10 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
         *** crawling 0.0 pps @ 0.0 Mbps
[216]    6 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
         *** crawling 0.0 pps @ 0.0 Mbps
[218]    6 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
         *** crawling 0.0 pps @ 0.0 Mbps
[220]    6 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12350 L  432K
         *** crawling 0.0 pps @ 0.0 Mbps
[222]    5 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12351 L  432K
         *** crawling 0.5 pps @ 0.0 Mbps
[224]    4 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12351 L  432K
         *** crawling 0.0 pps @ 0.0 Mbps
[226]    4 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12351 L  432K
         *** crawling 0.0 pps @ 0.0 Mbps
[228]    4 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12351 L  432K
         *** crawling 0.0 pps @ 0.0 Mbps
```

```
[230]   4 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12351 L  432K
        *** crawling 0.0 pps @ 0.0 Mbps
[232]   4 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12351 L  432K
        *** crawling 0.0 pps @ 0.0 Mbps
[234]   4 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12351 L  432K
        *** crawling 0.0 pps @ 0.0 Mbps
[236]   4 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12351 L  432K
        *** crawling 0.0 pps @ 0.0 Mbps
[238]   4 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12351 L  432K
        *** crawling 0.0 pps @ 0.0 Mbps
[240]   4 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12351 L  432K
        *** crawling 0.0 pps @ 0.0 Mbps
[242]   4 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12351 L  432K
        *** crawling 0.0 pps @ 0.0 Mbps
[244]   4 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12351 L  432K
        *** crawling 0.0 pps @ 0.0 Mbps
[246]   4 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12351 L  432K
        *** crawling 0.0 pps @ 0.0 Mbps
[248]   3 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12351 L  432K
        *** crawling 0.0 pps @ 0.0 Mbps
[250]   2 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12351 L  432K
        *** crawling 0.0 pps @ 0.0 Mbps
[252]   2 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12351 L  432K
        *** crawling 0.0 pps @ 0.0 Mbps
[254]   2 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12351 L  432K
        *** crawling 0.0 pps @ 0.0 Mbps
[256]   2 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12351 L  432K
        *** crawling 0.0 pps @ 0.0 Mbps
[258]   1 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12351 L  432K
        *** crawling 0.0 pps @ 0.0 Mbps
[260]   1 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12351 L  432K
        *** crawling 0.0 pps @ 0.0 Mbps
[262]   1 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12351 L  432K
        *** crawling 0.0 pps @ 0.0 Mbps
[264]   1 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12351 L  432K
        *** crawling 0.0 pps @ 0.0 Mbps
[266]   0 Q      0 E 1000004 H 139300 D 131051 I 81373 R 24997 C 12351 L  432K
        *** crawling 0.0 pps @ 0.0 Mbps
Extracted 1000004 URLs @ 3759/s
Looked up 139300 DNS names @ 523/s
Downloaded 81373 robots @ 305/s
Crawled 12351 pages @ 46/s (220.27 MB)
Parsed 432711 links @ 1626/s
HTTP codes: 2xx = 7953, 3xx = 1686, 4xx = 2601, 5xx = 110, other = 1
```

#2

- Avg no of links / html page with 2xx code = tot no of links obtained / pages with 2xx codes = 54
- Size of Google's webgraph with 1T crawled nodes:

The graph is stored in an adjacency list fashion. And from above result we can assume that each node is adjacent to 54 other nodes. That implies the graph has a total of 54 Trillion edges and size of the adjacency list being (54T * 64bit) = 432TB

#3

- Avg page size in bytes (across all http codes) = The total no of bytes / C = x = 17.8KB
- Bandwidth needed for bing to crawl 10B pages per day (in Gbps) = 10 billion pages * x(in bytes)*8 / (24*3600*10^(9)) = 16Gbps

#4

- Prob that a link in the I/P file contains a unique host = H/E = 0.139
- Prob that a unique host has a valid DNS record = D/H = 0.94
- % of contacted sites that had a 4xx robots file = R/I * 100 = 30.07%

#5

- No of crawled 2xx pages that contain a hyperlink to tamu.edu = 17,
- And 13 of them originated from outside of TAMU

For every successful 2xx page "A", I have taken the links found on this A one by one.
I further parsed this link to get its host, then did a substring search on the host for "tamu.edu" on this link. If "tamu.edu" is found, then the page A contains a hyperlink to our tamu domain.

To decide on how many of the above shortlisted 2xx pages have originated outside, I have further checked if our original page A belongs to our tamu domain or not(by looking for "tamu.edu" in the host of A). If the host of A doesn't contain "tamu.edu" then the hyperlinks have originated from outside of TAMU.

Note:
Where,
E: number of extracted URLs from the queue
H: number of URLs that have passed host uniqueness
D: number of successful DNS lookups
I: number of URLs that have passed IP uniqueness
R: number of URLs that have passed robots checks
C: number of successfully crawled URLs (those with a valid HTTP code)
L: total links found