Team Members:

- Yassine Jaoudi
- Samuel Akpan
- Samantha Clark

Usage:

Use the following command to run the code:

```
python Asg1CrawlThread.py 10 URL-input-100.txt
```

The code accepts two arguments, first one indicates the number of threads to run and the second one the input file

Requirements:

Python 3 with the packages included in the requirements file, use the following command to install:

```
pip install -r requirements.txt
```

ToDo list status:

The code for this assignment needs to meet the following items to meet a perfect score.

- 🚧 Include this emoji if you choose to work on the appropriate task
- ♻️ Include if this task is done, but it needs to be rechecked by another teammate
- ✔️ Include this one if its done

🚨🚨🚨 Choose the interesting tasks you would like to work on.

Function	Points	Break down	Item	Samantha	Samuel	Yassine
Running Output	5	1	Printouts every 2 seconds	🚧		
		1	Correct active threads	🚧		
		1	Correct attempted robots		🚧	
		1	Correct pps			🚧
		1	Correct Mbps			🚧

Function	Points	Break down	Item	Samantha	Samuel	Yassine
Summary	6	1 1 1 1 1 1	Correct URL processing rate Correct DNS rate Correct robots rate Correct crawled rate/totals Correct parser speed Correct HTTP breakdown	🚧 🚧	🚧 🚧	🚧 🚧
Code	6	1 1 2 1 1	>>20Mbps w/ 500 threads >>200MB RAM w/500 threads No deadlocks on exit No issues with the file reader No improper stats thread			
Other	1	1	No Missing files for compilation			
Report	25	5 5 5 5 5	Lessons learned and trace Google graph-size analysis Yahoo band-width analysis Probability analysis Written Report	🚧		

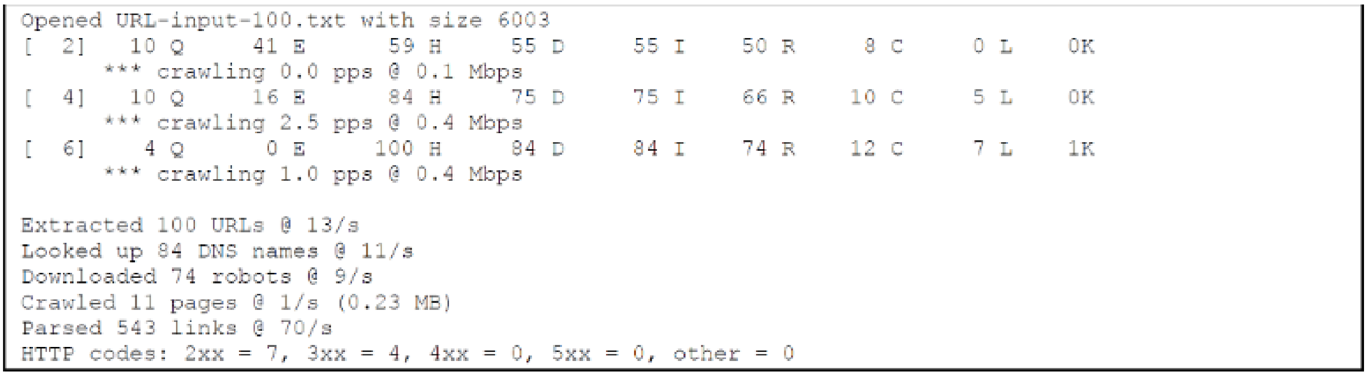
🐛 Bug Fixes 🐛 :

Bug	Status	Fix Implemented	Fixed by

Code Output:

Goal of this part 2 of the assignment:

hw1.exe 10 URL-input-100.txt



```
Opened URL-input-100.txt with size 6003
[  2]    10 Q      41 E       59 H      55 D      55 I      50 R      8 C      0 L      0K
          *** crawling 0.0 pps @ 0.1 Mbps
[  4]    10 Q      16 E       84 H      75 D      75 I      66 R     10 C      5 L      0K
          *** crawling 2.5 pps @ 0.4 Mbps
[  6]     4 Q       0 E      100 H      84 D      84 I      74 R     12 C      7 L      1K
          *** crawling 1.0 pps @ 0.4 Mbps

Extracted 100 URLs @ 13/s
Looked up 84 DNS names @ 11/s
Downloaded 74 robots @ 9/s
Crawled 11 pages @ 1/s (0.23 MB)
Parsed 543 links @ 70/s
HTTP codes: 2xx = 7, 3xx = 4, 4xx = 0, 5xx = 0, other = 0
```

Lessons learned and trace:

- Git version control has been utilized and learned from the team members for better collaboration.
- Team building aspects and workflow has been learned as we have a great team that is motivated to do the work, be there when another member of the team needs help with a certain task, and finish with great results.
- Frequent and great communication throughout the team using zoom meeting or group chat.
- Debbuging techniques.
- Time management.
- 🚨 We need to add the technical lessons learned in this asg 🚨

Google graph-size analysis:

Yahoo bandwidth analysis:

Probability analysis:

Tasks done from previous assignment part:

✨ Future Work ✨

- Find a way to specify the buffer size dynamically.
- Current code runtime is **118.58 ms**, we will be improving this runtime by making the code more effecient in order to decrease the runtime.
- Improve the overall design of the code.