Writeup – Simon Buchheit – 10/6/2019 – Hmwk3

1) Is there certain information about the webserver that you can discern based on what files you can access?

- Based on the directories that I can access, I can safely make the assumption that Hooli, is a tech company. This assumption comes from paths like `/workingtech/` and `/spectra/` (which sounds like a cool technology). I can also guess that Hooli publishes articles from the `/articles/` path. There also appears to be some sort of register located at `/mainregister/`. I am not sure what this contains, but it could be a list of technologies, or employees depending on the company. Finally, I can see that there is going to be busniess information located in the `/business/` path.

2) Are there any ways to improve the speed of your scanner?

- The first way to speed up the scanner would be to use a superior language. Python is inherently slow even though it has great functions for dealing with lists and other data sets. I could speed up my current scanner by using lists more efficiently. There are many places that I have double *for loops* to iterate through lists. These are slow because the lists are likely to have thousands of elements in them. Also, I could have set more aggressive timeouts for my sockets and threads. This would have freed them more quickly if they were hung up on a blocking operation.

3) How can response codes be used in order to more efficiently search your site?

- In my code I check if the response code is a 200. If it is a valid path I append my entire list to my searching list in order to check for deeper paths. EX. `/validPath/x` where x is a path from my path list. If there is a redirect for the response I make sure to follow the redirect location that was provided in the response. Status codes are great because they allow you to check to just the code instead of checking for a body in the response.

4) Are there any common naming patterns that you might expect would yield positive results?

- Based on my findings, `index.html` is definately a common naming pattern for a page. In addition, *dates, images, business, and articles* can all be considered *common* naming patterns. If a website is hosting articles it makes sense for them to be in the *articles* folder. Same for finding the business information in the *business* folder.