Maxwell Fusco
Homework 3 Act 4 Step 2
Writeup

- **Is there certain information about the web server that you can discern based on what files you can access?**

Generally yes, for instance if you can access a directory named admin, you can assume that there is more restricted data within it. Based on the name of directories or pages you can make fairly decent guesses about what kinds of information exist on that page or in that directory. For instance with hooli, we see directories like ‘2004’, ‘business’, and ‘articles.’ We can assume that hooli has been around since 2004 and archives their data based on directories labeled as such. From the directories ‘business’ and ‘articles’ we can assume that there exist things about business and articles in those given directories. Another thing to note is default pages can give a very good hint on what kind of web server is being run; for instance if you have a default nginx or apache server page you can assume that is that type of web server running.

- **Are there any ways to improve the speed of your scanner?**

There are several ways to improve the speed of my scanner. For one, more optimized coding, for a homework assignment like this my focus is on getting a working and robust library for working on the individual assignments and for that optimizations were limited. Another thing that would make my scanner faster is the utilization of faster language, as good as python is to use for this, if this scanner would to be written in a lower level programming language it would be faster. Lastly is to improve hardware, the utilization of better hardware would speed up slower code as well as a better network connection or more network connection could stop rate limiting and improve network speeds.

- **How can response codes be used in order to more efficiently search your site?**

If you are looking for whether a page exists you only need a response code, 200’s will tell you that it exists, 300’s will tell you the page was moved, 400’s will say that the page doesn’t exist, 500’s will say that the page is blocked or another error occurred. This can be used for extremely fast processing when you are looking to footprint a network because you will only ever have to receive the headers of that response and not the page source to tell basic info on the page. This means that the time it takes to download the entire page source is thus saved.

- **Are there any common naming patterns that you might expect would yield positive results?**

There are some naming patterns in terms of how the web server will work. For instance if there is ever a directory, if there is a base webpage for it, it would use index.html to render it. Similarly, organization is extremely important when it comes to making a larger web server. There are some files like sitemap.xml or robots.txt that can give you a better clue how the site is organized as well as some base pages to begin scraping on. Lastly, there are general patterns like /admin being a restricted path, /api is used for api calls, or simple stuff like /images storing images.