Project Overview

I made a program that crawls through Wikipedia to map its internal links. I scraped Wikipedia pages and extracted links from their html. The plan was to create a tree representing the network paths. I was hoping to graph it but after hours of trying I couldn't find a reasonable way within scope.

Implementation

I choose for the primary output of my project to be a list of lists to denote branches on a tree. This seemed like the right combination or obtainable as output and parsable by a graphing function. Many of my internal functions accepted and returned tuples as there were multiple distinct elements I wanted to pass (eg. a string and a starting point for a search).

One interesting data structure choice I made was to store the address of ever page that had been visited in a dictionary. This was on top of storing the order and relationships of the pages I went to. This in some ways is redundant. The list was used because it was the only easy way I could think of to maintain relationships between pages in an interpretable manor. The dictionary was used only to check if I had been to the page before. I did this to stay time efficient when scaling.

Acknowledging I may well have missed something, here was my thought process. For average cases, appending a list is O(1) but searching it is O(n). As my program performs easy operation k*n times (for some constant k) that was going to get very gross very fast. By implementing the dictionary, which searches in O(1) time, I increased the time constant and the space complexity, but kept my time complexity to O(n).

Reflection

When working with younger students on Baja I tell them over and over again that nothing is finished after the first design, especially when learning. I had to eat a lot of my own medicine on this project. There is hardly a line of this code that is original to my first pass. I regret not using git extensively on this project because it would have been a lot of fun to look at. I regretfully only started pushing my code well into the project and I didn't even do it to the right repo. Lesson noted. I'm happy that my code is generalizable (lets you choose the depth and width of your tree) but I'm bummed I couldn't work out a good graphing solution. With mechanical projects you have a trashcan of prototypes to show for your process, here I just have a lot of deleted bits.