HW5: Summary Report

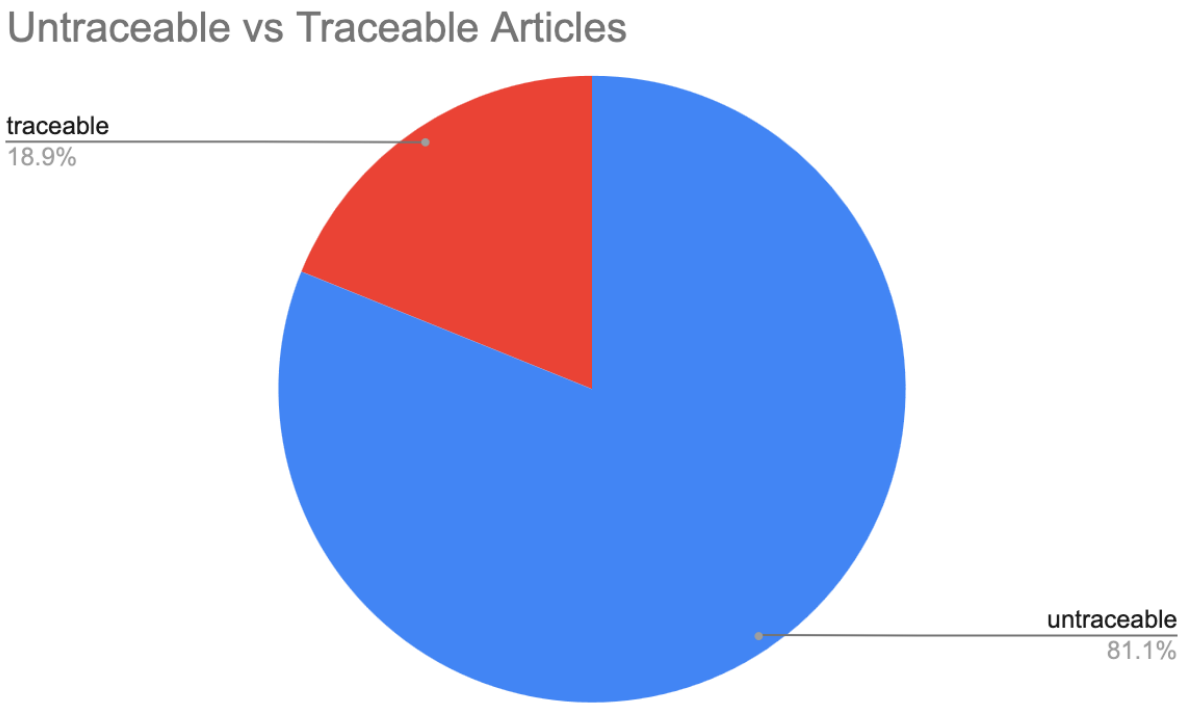
Project Description

For this project, my hypothesis was: no, not every Wikipedia article page could lead back to philosophy (how could a topic such as a water jug lead back to that?).

In order to prove this hypothesis, I downloaded all of Wikipedia's articles from this link[1] in order to not get banned for web scraping. Using this downloaded dataset, I built a graph from exploring every single link on every article. Nodes are the articles, and edges are the links between articles. Then, using BFS, I found the path between every article and the Philosophy page, if it existed, and counted the layers that it took to reach the Philosophy page from the starting article. I put all of the articles that were able to be traced back to the Philosophy page into a tsv file that contains the starting article for BFS and the layers between the starting article and the Philosophy one. Articles that could not be traced back to the Philosophy page were put into another text file. I did my analysis using Rust code, as I was much more comfortable using this language rather than Java for this task.

Analysis Results

From my analysis, it appears that not all article pages can be led back to the Philosophy page. In fact, the untraceable text file contains 1098534 articles that do not lead to the Philosophy page, proving my hypothesis is correct. On the other hand 255728 articles could be traced back to the Philosophy page. Nearly 81.1% of the data was not able to be traced back to the Philosophy page, as shown in the Pie Chart below.



Furthermore, it seems as though an article that could be traced back to Philosophy has a small amount of layers between the starting article and Philosophy–the average amount of layers between articles is 3,

[1] https://meta.wikimedia.org/wiki/Data_dump_torrents#English_Wikipedia

shown in the bar chart below. The highest number of layers between an article and Philosophy is 22. From my analysis, I can conclude that while most Wikipedia articles do not trace back to the Philosophy section, those that do are mostly close in distance to it.