Project Proposal

Darrion Vinson (dmv2124)

Nov. 24th, 2020

1 Introduction

Wikipedia contains over 6 million articles and grows more every year. As one of the largest public information graphs available, there are many interesting graph phenomena that manifest on Wikipedia. There is a popular game played on Wikipedia known as "Six degrees of wikipedia". The idea is that for any two people with Wikipedia pages, there is a maximum of six degrees of seapration between them. Any person's page can be reached from another's with at most 6 pages in between. This is widespread, but not guaranteed due to constant additions and edits of pages. Given Wikipedia's networked link structure, finding a path between two pages can be modeled as a shortest-path graph problem with an undirected unweighted graph.

2 Project Idea

2.1 Objective

The goal of this project is to generate a path between a start and end page for any two pages on Wikipedia. This code will play the Six Degrees of Wikipedia game. I don't plan on restricting the path to 6 steps as of now, since I think this may reduce the number of opportunities to show the benefit of parallelization.

2.2 Algorithms

For the shortest path algorithm, I plan to use one of Djikstra's, Bellman-Ford, or A*. These are all single-pair shortest-path algorithms, as I'm not

looking to solve the all-pairs problem, due to it being too computationally intensive for my laptop.

2.3 Deliverables

For the completion of this project, I will provide serial and parallelized implementations of this path generator. Additionally, I will provide test results showing comparisons of the two versions of the algorithm.

3 Resources

I will make use of the Wikipedia database dump in order to avoid scraping the live Wikipedia website. This database is in XML format, so I will make use of Haskell XML parsing libraries in order to do the graph navigation.