CS685A: Assignment 2
Report: Wikispeedia Paths and Graph

Question 13

Abhishek Bhatia (170022)

Odd Semester, 2020-21

1 Overview

We covered a dataset that contained the human navigation paths on Wikipedia. This was collected through the human-computation game Wikispeedia. In Wikispeedia, users are asked to navigate from a given source to a given target article, by only clicking Wikipedia links. In total, 4604 articles are used covered accross 146 categories, including the ones listed under ”no category”. Various comparison and analysis are done in this assignment, which are briefly explained below.

2 Articles and Its categories

We observed that an article can belong to multiple categories and the data for the same is given in article-categories.csv file. Most of articles belong to one or two categories whereas there are some articles which might belong to three categories.

3 Shortest Distance And Connected Components

We observed that the shortest distance to reach from one article to another is always less than 10. There are some pairs of articles which can't be reached from one another but the average distance between articles is around 4. We have 119772 direct edges, which means these number of pairs have shortest distance between them as one i.e They can be reached from one another with a single link. We can also see the connected components in the graph. We saw that there are around 521 connected components in the graph. One important thing to observe here that the biggest connected sub-graph had 4051 nodes in it with the diameter of 9. This means that 4051 articles fall under one single connected component such that we can reach one article from another without more than 9 clicks since diameter is 9.

4 Human Paths v/s Shortest Paths

We see that there are almost 20% of the cases when the human path is of same length as that of the shortest path. For all the rest paths, the human took a longer path to reach the destination. Considering the cases without clicks, almost 70 percent of the paths had length difference less than or equal to two. Also, we saw that there were only 3 percent of the paths with length difference greater than 8 or more out which only 1.23% were the paths with length difference of more than 10.

5 Finished v/s Unfinished paths

We analyzed the finished v/s the unfinished paths percentage in the data. We saw that for most categories the finished percentage was falling between 60 to 70 percentage. This means that for the given two categories, the finished to unfinished paths ratio is about 3:2 or more. We also analyzed the number of times a category occurs in the paths in both the human as well shortest paths. We saw that a few categories occurred thousands of times combining the paths