CS685: Data Mining

Assignment 2

Shilpa Chatterjee
Roll No. :- 20111057

1 Analysis of Wikispeedia Paths Data

For the analysis I have used data files obtained from wikispeedia paths-and-graph.tar.gz which is also present in `http://snap.stanford.edu/data/wikispeedia.html`.

1.1 Length of Human Paths(finished with no back edge)

The following is the length of human paths with no back edge obtained from finished-paths-no-back.csv.

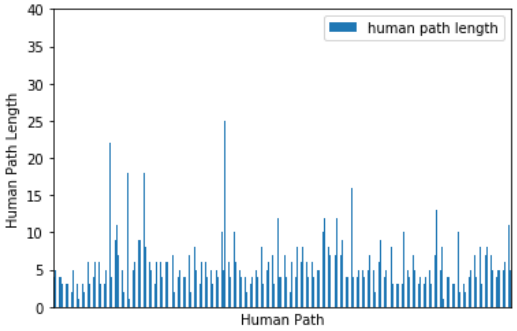


Figure 1: Human Path Lengths

1.2 Length of Shortest Paths corresponding to Human Path(finished with no back edge)

The following is the length of shortest paths corresponding to human paths with no back edge obtained from finished-paths-no-back.csv.

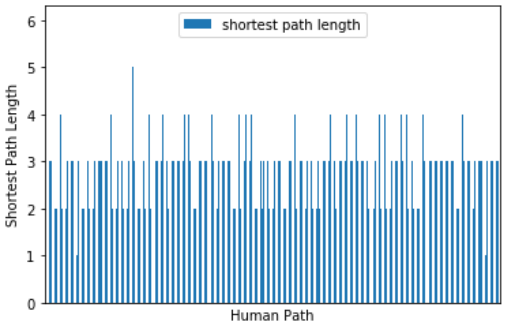


Figure 2: Shortest Path Lengths

1.3 Percentage comparison of human paths with no back edge with shortest paths

The following is the Percentage comparison of human paths with no back edge with shortest paths.

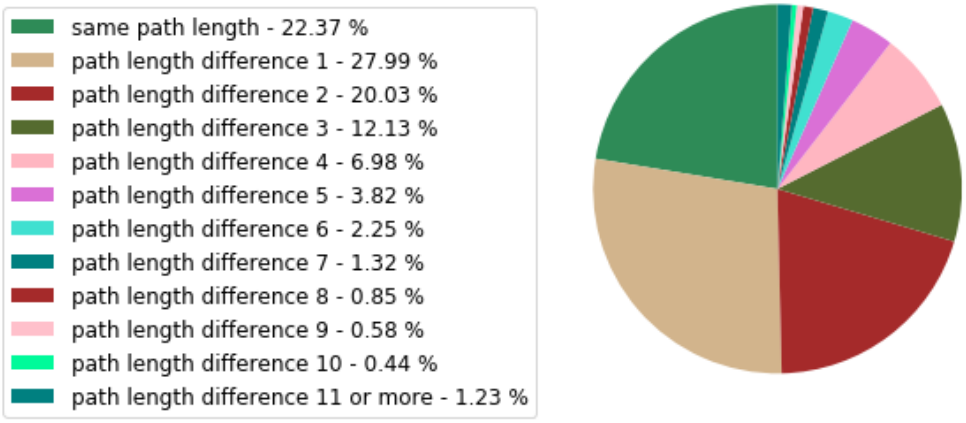


Figure 3: Percentage comparison of human paths with no back edge with shortest paths

The above figure shows that humans were able to take the shortest path only 22.37% which tells us that we humans usually fail to take the shortest path between a source and destination.

1.4 Percentage comparison of human paths with back edge with shortest paths

The following is the Percentage comparison of human paths with back edge with shortest paths.

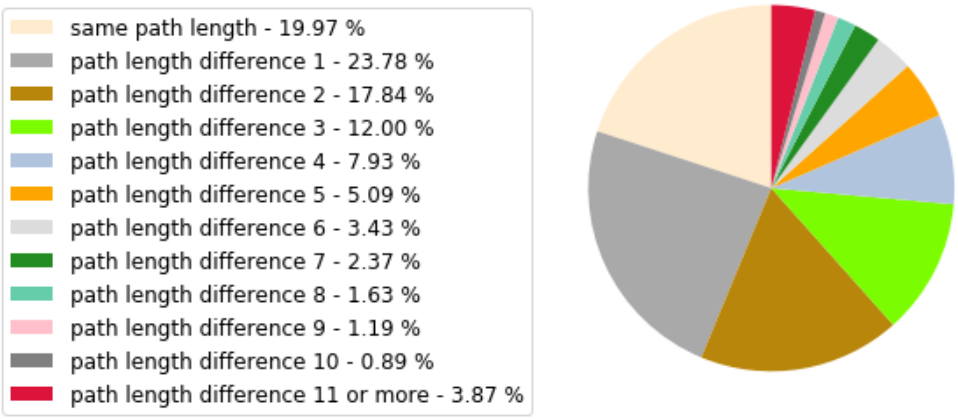


Figure 4: Percentage comparison of human paths with back edge with shortest paths

The above figure shows that humans were able to take the shortest path only 19.97% which tells us that we humans usually fail to take the shortest path between a source and destination.

1.5 Top 5 categories visited in human paths

The top five categories visited in human paths are **C0005 : subject.Countries**,
C0062 : subject.Geography.North_American_Geography ,
C0122 : subject.Geography.European_Geography.European_Countries,
C0056 : subject.Geography.Geography_of_Great_Britain ,
C0132 : subject.Science.Biology.General_Biology.

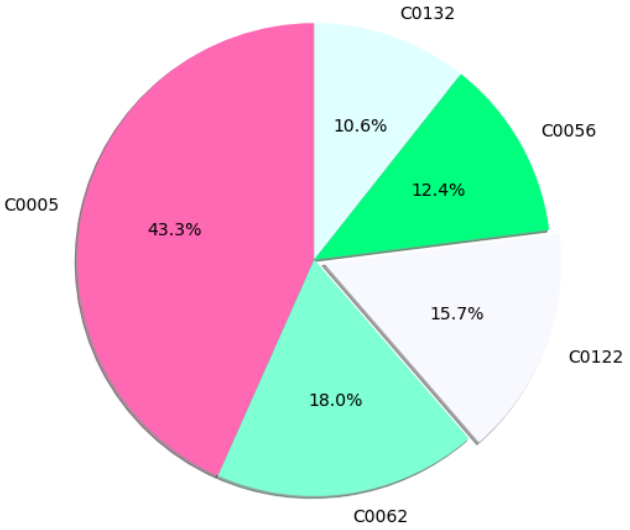


Figure 5: Top 5 categories visited in human paths

1.6 Top 5 categories visited in human paths when parent categories are considered

The top five categories visited in human paths when parent categories are updated when its child category gets visited are **C0001 : subject**, **C0008 : subject.Geography** ,
C0016 : subject.Science, **C0117 : subject.Science.Biology** ,
C0005 : subject.Countries.

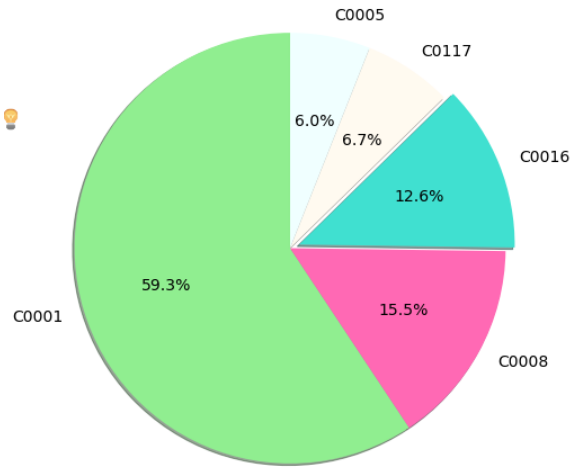


Figure 6: Top 5 categories visited in human paths when parent categories are considered

1.7 Top 5 category pairs visited in finished human paths with no back edge

The Top 5 category pairs visited in finished human paths with no back edge are **(C0002, C0031),(C0002, C0047),(C0002, C0052),(C0002, C0058),(C0002, C0063)**.

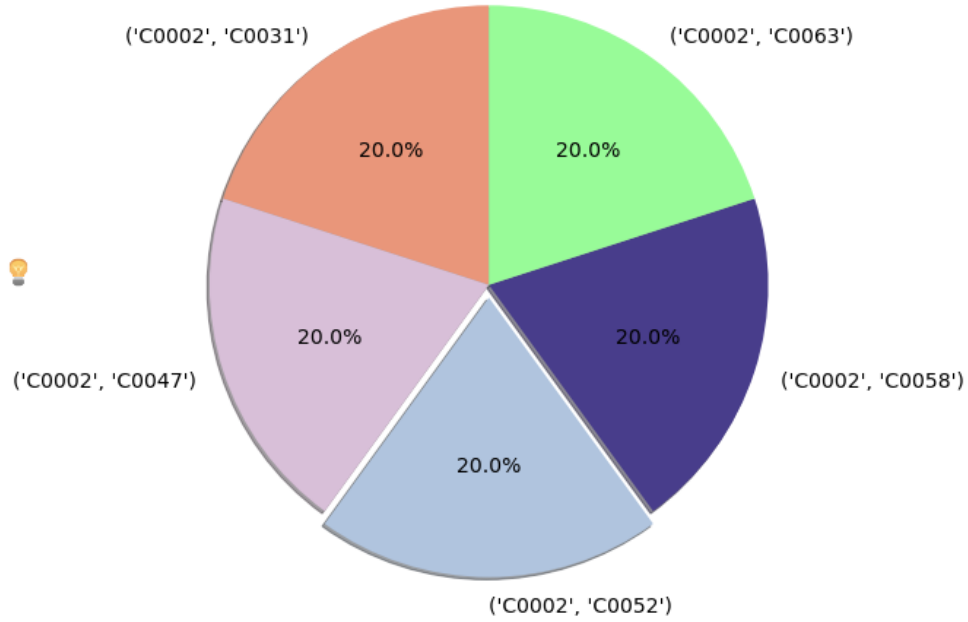


Figure 7: Top 5 category pairs visited in finished human paths with no back edge

1.8 Top 5 category pairs visited in unfinished paths

The Top 5 category pairs visited in unfinished paths are **(C0002, C0012),(C0002, C0022),(C0002, C0026),(C0002, C0032),(C0002, C0033)**.

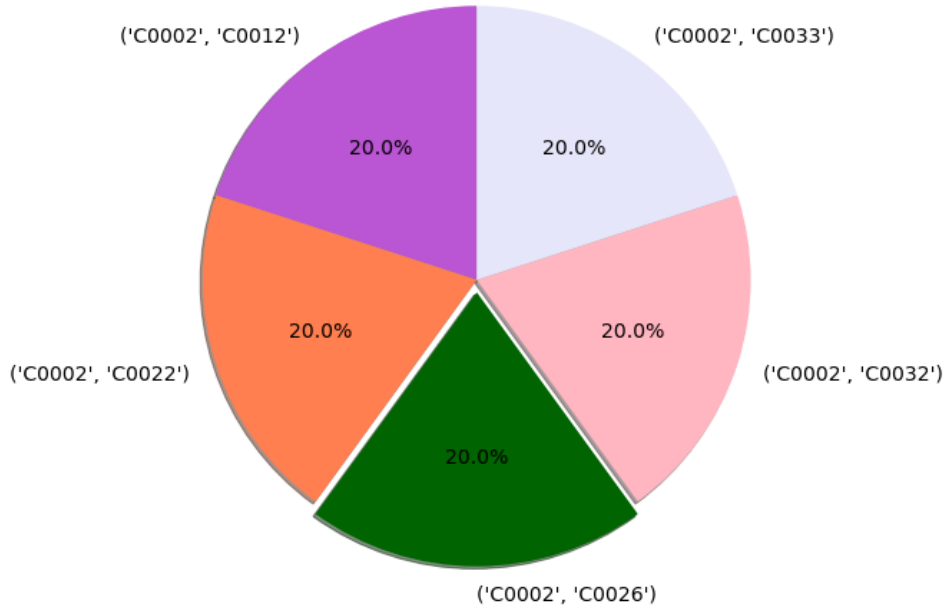


Figure 8: Top 5 category pairs visited in unfinished paths

2 Analysis of Wikispeedia Finished Path Lengths

2.1 Finished Human Path Analysis

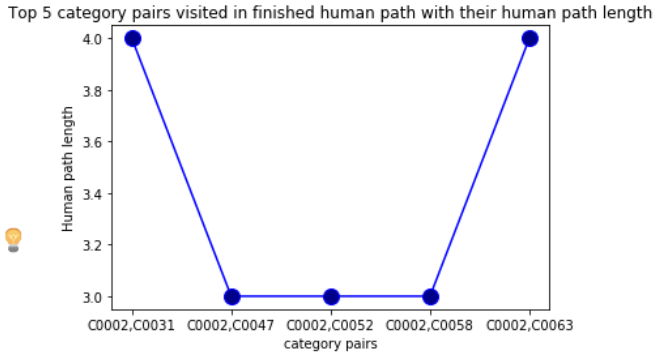
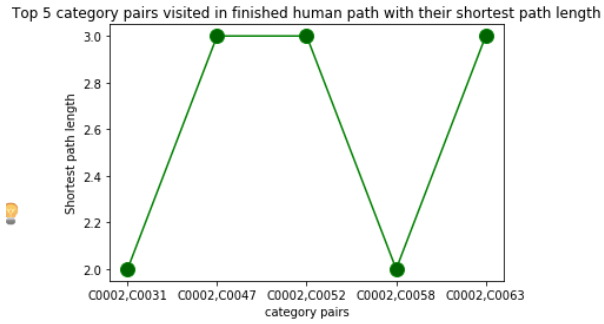
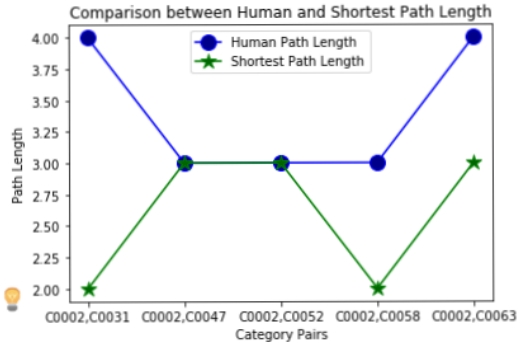


Figure 9: Human Path length corresponding to the top 5 category pairs visited in Finished Paths

2.2 Shortest Path Analysis Corresponding to Finished Human Paths



(a) Shortest Path length corresponding to the top 5 category pairs visited in Finished Paths



(b) Human Path Length vs Shortest Path Length

Figure 10: Comparison between Human and Shortest Path Lengths of Finished Human Paths

From Figure 10(b) it can be easily seen that we humans rarely take the shortest path between a source and destination.