CS685 - Data Mining
2020-2021-I: Assignment 2

Name: Sarthak Singhal Roll No.: 170635

1. Question 13

Assumptions made in the assignment:

- Undirected graph is used for finding connected components in Question 5.
- In questions 6,7,10,11, I have ignored the paths where src $\rightarrow$ dest is not reachable and where src == dest.
- In questions 8,9, I have only ignored the paths where src $\rightarrow$ dest is not reachable.
- For question 10 I have corrected some of the spelling mistakes. The rest article names which are not present in articles.tsv but present in paths_unfinished.tsv and I couldn't find appropriate article names for them for correction are: {The, The_Rock, Netbook, Western_Australia, Test, English, Mustard, Usa, Rat, Macedonia, Fats, Christmas, Black_ops_2, Great, _Zebra, Georgia, Sportacus}. So, I have assigned them the category *subject* as discussed in the class.

Analysis:

- A large number of articles are chained to each other($\sim$4500 out of 4604) i.e. for more than 95% of the articles if we start from an article then we can reach almost all the articles by following links present in them. This means that only less than 5% of the articles are based on some specific topic which doesn't have much relation with other topics(i.e. they are self-complete) or the creator might have forgotten to add hyperlinks.
- Given a source and destination article and given that humans always reach from the given source to the given destination, $\sim$20% of the times humans follow the shortest path (including back clicks). Also, $\sim$60% of the times humans reach destination in no more than 2 steps extra from the shortest path. This means that humans have the tendency to follow the shortest path to complete their tasks.
- The category *subject.Countries* comes in more than 50% of the finished human paths. This means that this is a very famous category and many of the articles have links to the articles which come under this category.
- When including the subtree of a category into itself, the category *subject.Geography* comes in $\sim$80% of the finished human paths. This also means that *Geography* is a very broad category and many articles have links to the articles which comes under this category and humans tend to click on those links maybe because they found them more close to the

destination as compared to other links.

- ∼70% of the source-destination category pairs have more finished paths than unfinished paths. This means that for a larger portion of the pairs people have some domain knowledge about those fields so that they can reach the destination i.e. much portion of the categories are based on general topics on which the common people have knowledge of so that they can find a way to reach the destination.
- For ∼33% source-destination category pairs humans follow path with length less than equal to 1.5 times of the shortest path. And for ∼88% pairs humans follow path with length less than double of the shortest path. They are able to do this without any prior knowledge of the articles' graph. This is done just on basis of intuition and some domain knowledge. This shows that humans want to minimize the time taken to complete their tasks. There is one case when the human confuses very much and takes path with length 17 times more than the shortest path. The source category in that case is *subject.Everyday_life.Films* and the destination is *subject.Music.Performers_and_composers*. The reason might be that these two categories are not much related or the links on those pages are not that intuitive to reach the destination.