CS685: DATA MINING
REPORT ON ASSIGNMENT 2

Abhas Kumar-20111001
M.Tech CSE

20 November 2020

Abstract

This document reports the significant results obtained while working with **wikispeedia-path-grath** dataset , presents analysis of these results and finally draws few conclusions based on the analysis.

RESULTS AND ANALYSIS

1. **About Articles and Categories.**
   Total number of unique articles in this data was **4604**. Total of **146** unique categories were there. Out of 4604 articles **4006** belonged to only **1** category, **590** articles belonged to **2** categories and only **8** articles were having **3** categories.

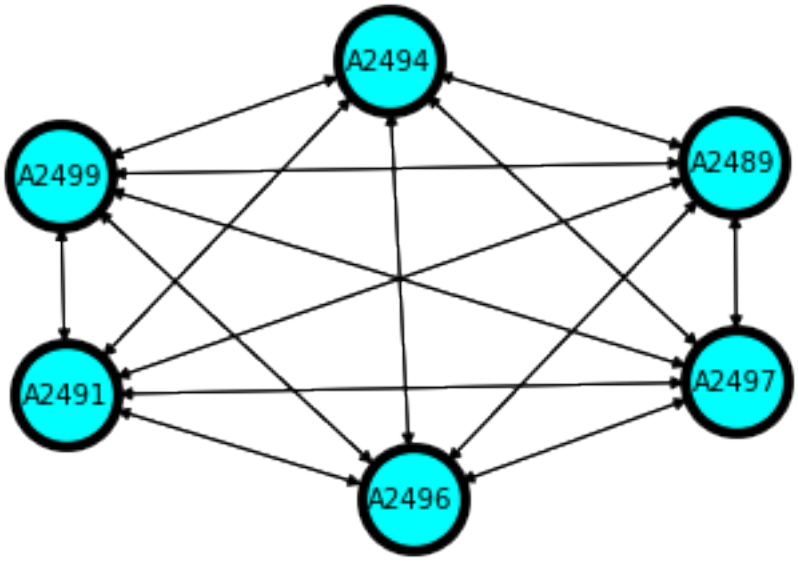

Figure 1: One of the components with 6 nodes (Article IDs as node)

2. **About graph.**
   Directed Graph having articles as nodes can be obtained, where total number of directed edges were **119772**. Entire graph was not **Strongly connected**. Total numbers of **strongly connected components** were **521**. **502** out of these components were having just a single node(thus **0** edge as no self loop was there). **16** out of these components were having **2** nodes each, just **1** component had **3** nodes. **1** component had **6** nodes with total of **30** directed edges among them(figure 1).
   And the **largest** component was with **4051** nodes and **111797** directed edges among them with a diameter of **5**.

3. **About human paths**
   5 leaf Categories that are visited most in path finished by humans are namely **subject.Countries**, **subject.Geography.North-American-Geography**, **subject.Geography.European-Geography.European-Countries**, **subject.Geography.Geography-of-Great-Britain**, **subject.Science.Biology.General-Biology**.

Whereas **5** least visited leaf categories by human namely **subject.Art.Artists**, **subject.History.Historians-chroniclers-and-history-books**, **subject.IT.Cryptography**, **subject.People.Producers-directors-and-media-figure**,**subject.Citizenship.Conflict-and-Peace**.

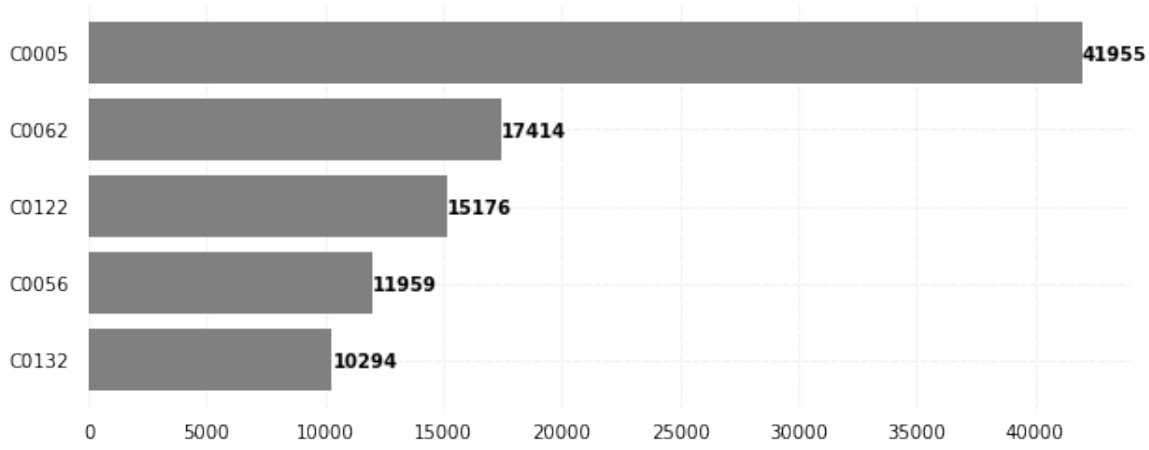

Figure 2: Showing 5 most visited categories(their ids on y-axis) with number of human paths visited them

4. **About source and Destination:**

   Total number of unique source and destination pairs obtained by from finished and unfinished human paths were **17196** out of which **10325** pairs where unique to finished human paths.

   Out of all **10325** source destination pairs obtained from finished human paths **91.6%** were having average ratio of human path length to shortest path length between 1 and 2, **6.6%** were having exactly 1, **1%** between 2 and 3 and only **0.8%** greater than 3.

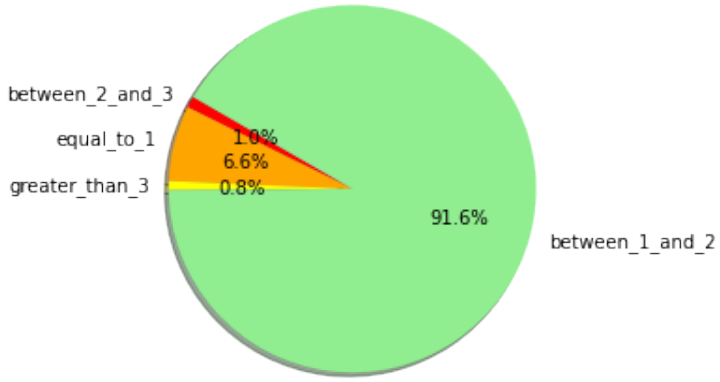

Figure 3: Percentage of pairs having average ratio of human path length to shortest path length equal to 1, between 1 and 2,between 2 and 3 and greater than 3

CONCLUSIONS

1. From the output of Question 7, **70.4%** of the finished human paths were having length of human path greater than the length of shortest path by just 2, and **82.5%** finished human paths were having length of human path greater than the length of shortest path by just 3. Which signifies the fact that human intuitively tends to follow shortest paths without even knowing the actual shortest path among articles.

2. From the pie chart above(Figure 3) , **98.2%** of all source destination pairs in finished human paths were having average ratio of human path length to shortest path length less than equal to **2**. This shows that , when analysis is done on large number of source destination articles most of the time humans do follow the shortest path.