CS685 Data Mining: Assignment 2

Ankita Dey, 20111013

November 20, 2020

1 Introduction

In this report, I have mentioned some of the findings I have come across after analyzing the Wikispeedia navigation paths dataset for CS685 Data Mining Assignment 2.

Wikispeedia is a game in which users are asked to navigate from a given source to a given target article, by only clicking Wikipedia links. The information of finished and unfinished path traversed by humans and the shortest path length between any two articles as well as which article belongs to which category have been given to us.

2 Commonly traversed categories

We can think of categories whose articles are clicked more number of times to be more popular and hence are intuitively thought of by players to lead them to their destination article. More often than not these categories also come more often in shortest path.

2.1 *subject.Geography*

If we check the number of times a category have been traversed, we find *subject.Geography* as one the most common category. Among the subcategories or children of this category *North_American_Geography* and *European_Geography* are the two most common categories. Arguably, these are the most influential continents of the world and their influence leaves their trace even in this data.

The category of *Asian_Countries* comes next among the continents of the world which shows the steady growth and influence of Asian countries in the world.

2.2 *subject.History*

After the more general subcategories of *Ancient_History_Classical_History_and_Mythology* and *General_history*, the popular subcategories of *subject.History* are filled with categories like

British_History.British_History_1500_and_before_including_Roman_Britain,
British_History.British_History_Post_1900,
British_History.British_History_15001750.

This is not a coincidence, History is always written by the powerful side and to quote Wikipedia's 'British Empire' page directly, 'At its height, it was the largest empire in history and, for over a century, was the foremost global power. By 1913, the British Empire held sway over 412 million people, 23% of the world population at the time, and by 1920, it covered 35,500,000 km2 (13,700,000 sq mi), 24% of the Earth's total land area.' Consequently, history is full of them.

Also notably, the *Military_History_and_War* is the fifth most popular subcategory in *subject.History* which shows how gravely War shapes our history.

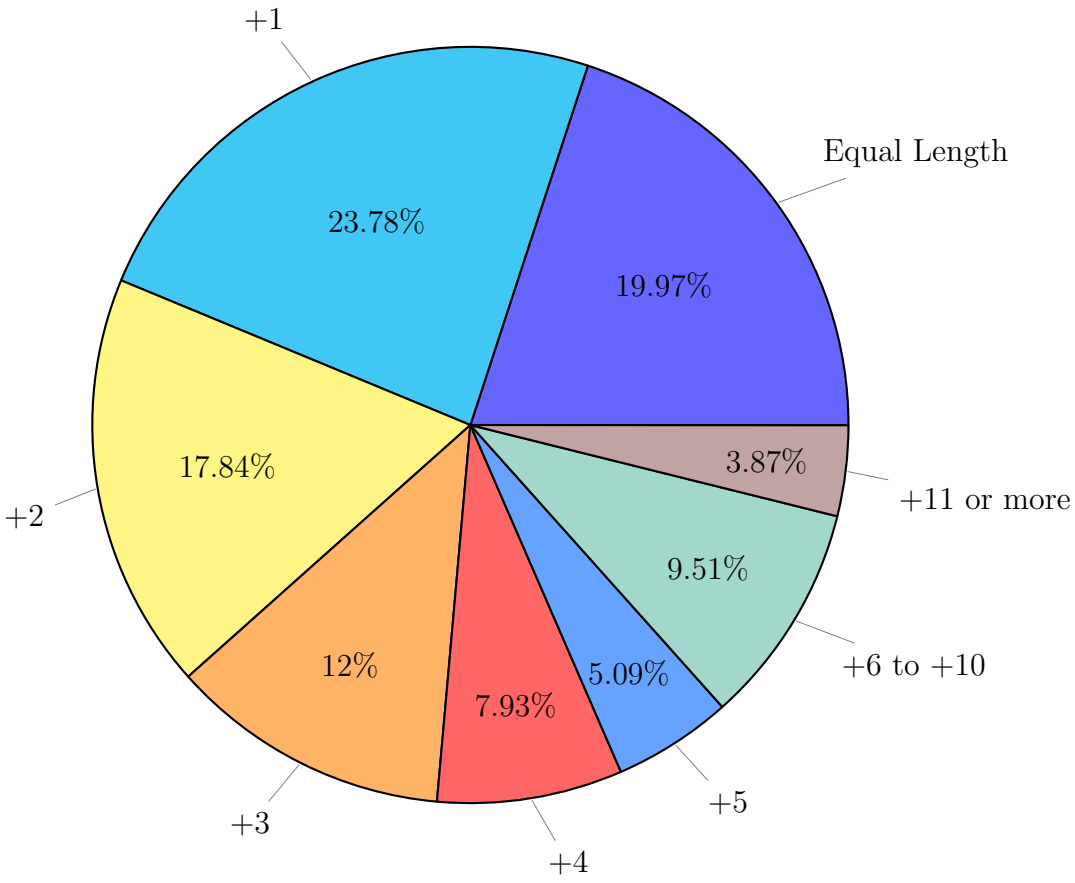
2.3 *subject.Citizenship*

Politics_and_government forms the pillar of citizenship and has been traversed maximum number of times and as expected, also has maximum number of articles among all subcategories of *subject.Citizenship*.

What's striking is that *Media* is the second most popularly traversed subcategory of *subject.Citizenship* in spite of *Environment*, *Law*, *Animal_and_Human_Rights*, *Education* and *Culture_and_Diversity* (other subcategories of *subject.Citizenship*) having more number of articles in them. We all know how media is intricately intertwined with citizenship even without us realising and like always that is reflected in data.

3 How intuitive is the shortest path for human?

The Pie Chart below shows the percentage of human paths that have equal length as the shortest path and where human path length is more than the shortest path. Here +x indicates human path length is x more than shortest path.



I have shown the human path length including the back-links because that is the actual path used by humans to traverse from source to destination article. We can see that the shortest path has been traversed by humans just around 20% of the time. Even if we remove the back links that is ignore the mistakes made in path, still the number of times shortest path has been traversed by humans comes to just around 22%-23%.

Not only that, for around 34% of the paths, humans have taken double or more the path length than shortest path. Thus, all of these clearly shows that shortest path is often not the most intuitive path.

4 Conclusion

We can see how a simple game that involves just clicking article links to reach a particular Wikipedia article can give us so much information on human intuition, influence and power. Thus we see minutely studying any data can give us some information and thus again we see why study of Data Mining is required.

References

[1] http://snap.stanford.edu/data/wikispeedia/wikispeedia_paths-and-graph.tar.gz

[2] https://en.wikipedia.org/wiki/British_Empire