The Wikispeedia dataset is based on a game in which players must find the shortest path from an *initial article A* to a *goal article G* (defined from the very beginning). They must use the hyperlinks contained in each article to move from one article to another (A→ A' → A''...G) to get closer to the goal as quickly as possible.

- Interesting and creative ideas, however, descriptions for some aspects are unclear as explained in the detailed feedback.

Idea 1:

In the article "Wikispeedia: An Online Game for Inferring Semantic Distances between Concepts" (*Robert West and Joelle Pineau and Doina Precup, School of Computer Science, McGill University*), it is written that one of the most common behaviours among players was to first find a **hub** (an article on a very general topic). How do we define these **hubs** (do they contain more hyperlinks than normal, are they linked to more categories than average)? What correlation is there between these **hubs** and the "true" shortest path (heatmap of hubs in the shortest path algorithm matrix)? Once the hubs are defined, it becomes possible to study the paths that don't pass through hubs and to define their characteristics. Are they faster or slower? What proportion of finished and non-finished paths pass through these hubs (is it a condition for finish a path)?

- Idea 1: Hub is standard terminology in the field of network science, and I would implore you to look into https://en.wikipedia.org/wiki/Hub_(network_science) and the citations therein to come up with a definition of Hub in this dataset. Also, some of these questions have already been studied in this paper: https://dl.acm.org/doi/10.1145/2187836.2187920. It would be great if you go through it and try to find novel analysis that you can do.

Idea 2:

The categories of each article along the paths could be used to determine semantic characteristics of the *goal article G*. Are there categories that players choose more frequently than others, at the risk of moving away from the goal? How do the categories of the articles the player clicks on relate to the categories of the *goal article G*? Is it possible to build a larger category matrix, with weights according to the categories encountered in the path (categories close to G in the path and often encountered in different games have more weight). For new parts, would it be possible to "guess" the *goal articles G* by referring to the categories of the articles encountered (using for example a machine learning algorithm)? Such a category matrix for each article would allow to create a semantic network based on human language and culture.

- Idea 2: Interesting idea, however, I don't clearly understand whether you would like to predict the goal article G or its category from the categories of articles encountered in the path: doing the former seems more feasible IMHO? Additionally, once you have such a semantic network, what would you like to do with it? I would encourage you to think of some important applications of such a representation that you aim at building.

Idea 3:

Games with larger than average shortest path length (SPL) present additional difficulties for players (**difficult games**), because the path from *initial article A* to the *goal article G* is naturally longer. Are players able to find efficient paths even in **difficult games**? Do the statistical variables (mean, median, standard deviation) of games played by human differ more from SPL in **difficult games** than in average? What is the link between non-finished games and **difficult games**, do players give up more easily if the game seems long and difficult? It is interesting to look at the specific data related

to these **difficult games** to try to understand what create a distance between the *initial article A* and *goal article G* in difficult games (semantic distance, conceptual distance, "lack of hyperlink" distance).

- Idea 3: Interesting idea, but the focus seems a bit narrow. Additionally, the questions that you pose seem to be quite descriptive. Let's say you have an answer to these questions, what would you eventually present in your data story? Your project mentor throughout the semester will be Akhil Arora: akhil.arora@epfl.ch. For future discussions specific to your P2 and P3 deliverables, you are encouraged to be in touch with your mentor.