ADA Proposal (Wikispeedia)

Below you may find ideas applicable to the Wikispeedia dataset, perceived after reading the corresponding material and exploring the data. Those ideas combine both research and analysis, utilizing one or more of the available data sources, on tasks I find interesting.

[1] Defining a unified similarity distance metric.

Defining a great similarity distance metric is crucial for calculating similarity between articles. For this matter, this idea proposes a unified similarity distance metric, utilizing both Wikispeedia metric and similarity of entities extracted from knowledge graph embeddings. We can integrate knowledge bases (DBpedia/Wikidata) with the Wikipedia graph to assign labels to the network links. Then, we can produce embeddings in the form of multidimensional vectors (using FastText or node2vec) and calculate the semantic distance between articles using cosine similarity. These two metrics (Wikispeedia and Embeddings) will lead to the definition of a unified metric using weights to determine how each one contributes to the result. How do the results (recall, precision) compare to the results of each metric individually? Could this lead to a combination of embeddings and the Wikispeedia algorithm?

- **Decent idea, but some details are missing. (1) What is the Wikispeedia metric (please quote a reference if existent)?, (2) It is unclear why you use the term knowledge graph embeddings, as you don't use the knowledge graph (subject, relationship, object), rather propose to use FastText that captures the text-based embeddings or node2vec that focuses on just the network structure (no labels on the relationship). (3) Again, it is unclear how you would combine the two metrics to unify them? (4) Lastly, I don't understand what you mean by the "precision, recall" of results. What is the underlying task on which you compare these metrics?**

[2] Exploring the structure of Wikipedia graph through the hub nodes.

The initial phase of the game consists of users trying to reach a getaway hub node. Those nodes are crucial, as they enable the player to start the homing phase by narrowing down the search. Thus, it is interesting to analyze these nodes and see how they contribute to the graph structure. What is the distribution of the article categories of these nodes? Do specific categories have more hubs than others? How fast users can reach those hubs? We can do a connectivity or centrality analysis for these nodes to understand the structure of the graph. Does the Wikipedia network form a bow-tie structure like the web?

- **Interesting idea, but 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? 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.**

[3] Link prediction using the user path data.

Automatically predicting connections between semantically similar articles is an important aspect of the Wikipedia semantic network. Using the paths to produce a subgraph that consists of nodes with links defined by the user common sense would result to a smaller semantic network, structured by the game data. Applying link prediction algorithms to this network will automatically detect semantically similar articles, as it will form links in a network built by common knowledge. It is also interesting to examine such graphs to understand if properties such as triadic closure or small world phenomenon apply to such data. Finally, it is a subject of creative graphs, as we could show the automatic link prediction in an interactive visualization of the network.

- **Using human navigation paths to identify new links to be predicted is interesting. It is also interesting to analyze the properties of the network formed by human navigation behavior. That said, have you looked at this paper: https://arxiv.org/abs/1503.04208 that uses Wikispeedia for studying the link prediction task? If not, I would implore you to do so, and also try to differentiate your proposal from what's already done in that paper.**