APMA 4990
Modeling Social Data
Jake Hofman

Human Navigation of Information Systems: Wikispeedia

By Andres Soto (ads2206), Jayati Verma (jv2488), and Tomer Solomon (ts2838)

https://github.com/tomersolomon/msd_finalproject

Introduction

We analyzed a dataset on the game Wikispeedia in order to explore how humans navigate through information systems, such as the Wikipedia article network. Understanding this is critical in designing such systems, because a better understanding of what humans find intuitive and easy to navigate could enable the creation of systems that are more effective and efficient to use. This question is particularly interesting to us because it allows us to use graph theory analysis to dig deeper into the bigger conceptual questions of how we think, connect concepts, and access information.

Dataset

Wikispeedia is a game in which players are given a source article and a target article, and tasked with connecting the two articles using hyperlinks. For example, given a source "Batman" and a target "Vitamin D", an example path between the two is Batman → Superman → Sun → Sunlight → Vitamin D.

The Wikispeedia dataset is from the Stanford Large Dataset Network Collection and was created by pulling information from 51,318 complete games played and 24,875 incomplete games played. These games were played using a subset of the total Wikipedia article network, using 4,604 Wikipedia articles which in total contain 119,882 hyperlinks. The dataset contains each game's path length and duration as well as the optimal (as in shortest) path length between all source and target articles. It also contains information about the Wikipedia articles themselves, including the hyperlinks in each article, the categories (such as Geography or Science) that every article falls under, and the actual text content of each article. The dataset itself is a directed graph in which nodes are articles and directed edges are hyperlinks from a source article to a target article.

Stanford's Robert West and Jure Leskovec, in "Human Wayfinding in Information Networks," use this same dataset and discuss that humans play Wikispeedia relatively efficiently given the large network size. Additionally, humans tend to play Wikispeedia by first going to central "hub"

articles and then to more specific content, giving us an approach for our overview question of how humans navigate through information systems.

For consistency in our analysis, we used the complete games only, rather than the instances where players start the game and then give up mid-play. We used the raw data directly.

Exploratory Data Analysis

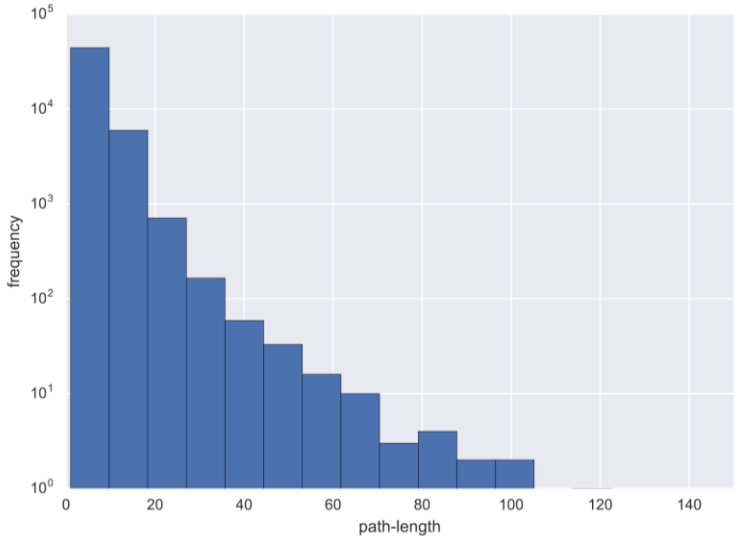

Figure 1: Distribution of path lengths for all games

To gain a baseline understanding of the Wikispeedia network, we computed the mean optimal (as in shortest) path length between all of the articles to be 3.7 clicks, revealing that our Wikispeedia network exhibits properties of small world networks.

However, distribution of path lengths for all of the games played revealed that player performance tended to be worse than the average optimal path length (Figure 1). Though player performance tended to be worse than the optimal, we see a heavy skew towards games completed in less than twenty clicks and there are relatively few games completed in thirty or more clicks.

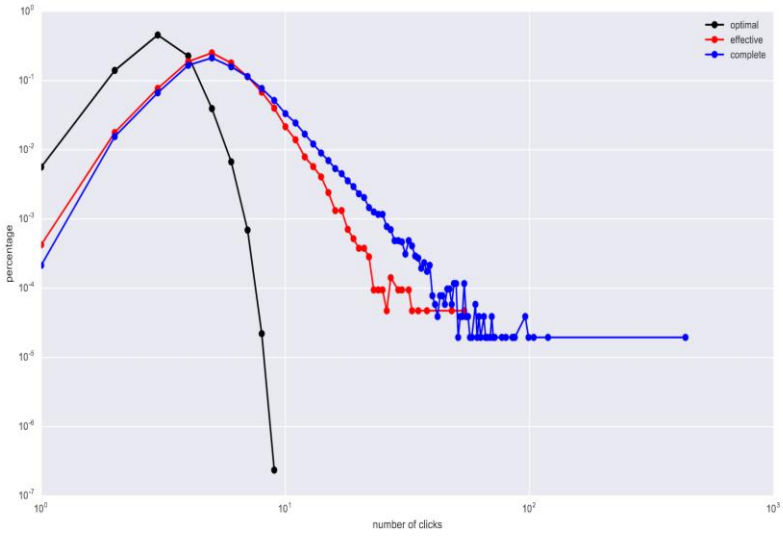

Figure 2: Frequency of games of each path length played (optimal, actual, and effective)

To confirm that player performance tended to be worse than the optimal, we compared the frequency of games played of various path lengths, under an optimal scenario, the observed scenario, and the effective scenario (Figure 2). Our dataset contained the path traversed for each player, including back-clicks; this was the observed scenario. To get an effective scenario, we simply removed all back-clicks and updated the path length for each player accordingly. We saw

that roughly 30% of games had an optimal path-length of 3. Additionally, the key takeaway was that for games of optimal path length greater than 3, player performance was much worse than the optimal.

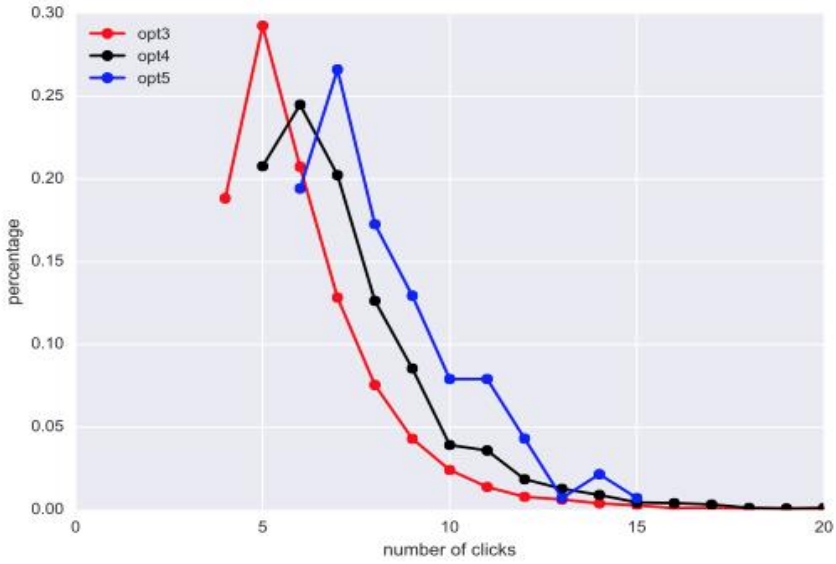

Figure 3: Player performance for games of optimal length 3, 4, and 5

Another important insight was gained when we grouped all games by optimal path length solution and plotted the frequency of player performance for the three shortest optimal path lengths (Figure 3). We clearly see that almost all players are at least 2-3 clicks worse than the optimal solution. Naturally, there is a shift to the right as we increase the size of the optimal solution, and a clear exponential decay in the percentage of players who complete the game in sub-optimal paths.

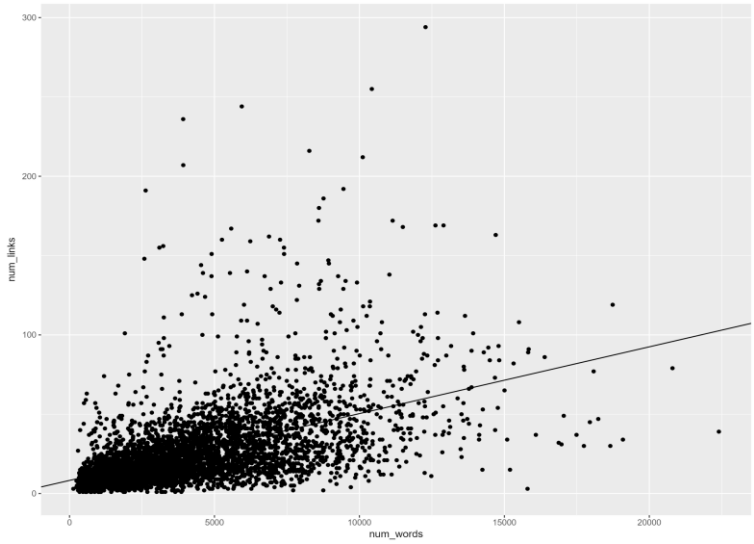

Figure 4: Regression between article size and article connectedness

To further develop our understanding of the dataset, we dove into the actual articles that players were clicking into; we intuited that bigger articles were likely to be more connected to other articles because they would contain more links. We performed a least squares linear regression with article size, measured by number of words, as the explanatory variable and connectedness of article, measured by number of links, as the observed variable (Figure 4). The correlation was extremely weak as R^2 = 27%.

Lastly, we explored the organization of articles into categories. Though the dataset provided article mappings to categories and subcategories, we only used the most general category level

for simplicity in analysis and visualization (there were 15 categories and around 50 subcategories).

Using all of the articles from the paths, we counted how many articles were in each category (Figure 5a). This yielded that certain categories, like Geography and Science, were much more prevalent than others. This matched with the idea from previous research that humans use geographical concepts to navigate through articles. However, our same analysis on just the source and target articles as a control yielded a similar distribution, suggesting that geography may not be significant (Figure 5b). Additionally, the similar distributions could suggest that humans navigate through Wikipedia articles using categorical connections.

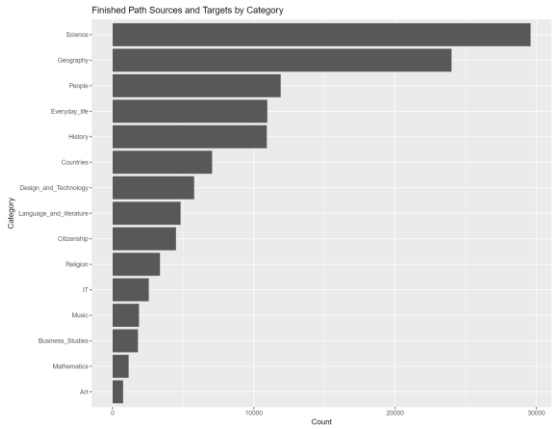

Figure 5a: Categories for articles in all paths

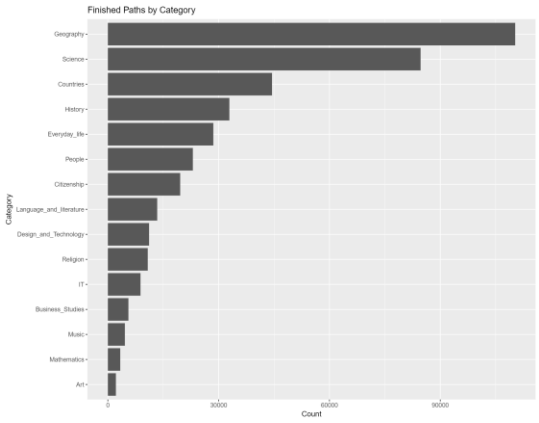

Figure 5b: Categories for source and target articles

Modeling Goal

From West and Leskovec's discussion on article "hubs" and content, as well as our exploratory data analysis on categories, we had the idea that humans navigate through Wikipedia by using conceptual connections between articles. However, we also found that human performance is not optimal, maybe because optimal paths are not necessarily conceptually intuitive.

This disparity motivated us to define a "semantic distance" for the conceptual connection, or content similarity, between two articles as well as a "topological distance" for the graphical connection between two articles. More specifically, we used the dimensionality reduction technique Latent Semantic Analysis (LSA) to calculate semantic distance and optimal path length as the metric for topological distance, both of which are standard.

As such, we distilled our task into attempting to answer two questions: One, can semantic distance be a more effective way to organize Wikipedia categories? Two, can we better predict human Wikispeedia performance by using semantic distance rather than topological distance between source and target? We used the topic model Latent Dirichlet Allocation (LDA) and least squares linear regression to address these two questions.

Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is an unsupervised topic model that helps us discover the main topics and themes within a corpus. Intuitively, each document within a corpus consists of multiple topics, and each word in a document comes from one of these topics. LDA takes the corpus and creates topics defined by significant words. For example, in a corpus about bioinformatics, the topics generated correspond to genetics, evolution, neuroscience, and computers respectively (Figure 6).

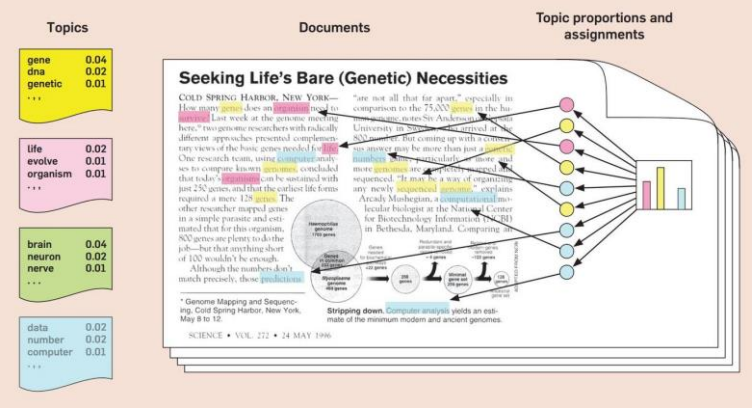

Figure 6: Example LDA

Though a mapping of the articles in our dataset into categories was given, we used LDA to define topics to see whether it would reveal potential improvements in classifying the articles, allowing players to better navigate the network, or insights into how similar/different articles are.

We generated top keywords for 15 topics (since there are 15 categories given by the dataset) using LDA on a Bag of Words representation of the Wikipedia articles; specifically, by using the CountVectorizer() class, a document frequency threshold of 2, and not including English stop words (Figure 7). As we can see, LDA created interesting topics that match conceptually with the categories provided. For example, we see a topic with keywords of “flag”, “government”, “country”, “republic”, which maps to civilization/citizenship. Another topic with “lake”, “river”, “national”, and “state” maps out to geography. We also see a topic with “music”, “new”, “band”, and “album”, which maps out to music. Similarly, “King”, “church”, and “England” can be mapped to history. This mapping confirms that the LDA model make semantic sense; furthermore, it provides insight into the most important words that define a topic. This is important because our understanding of the words that make up different topics might lead to insights on how categories relate to one another.

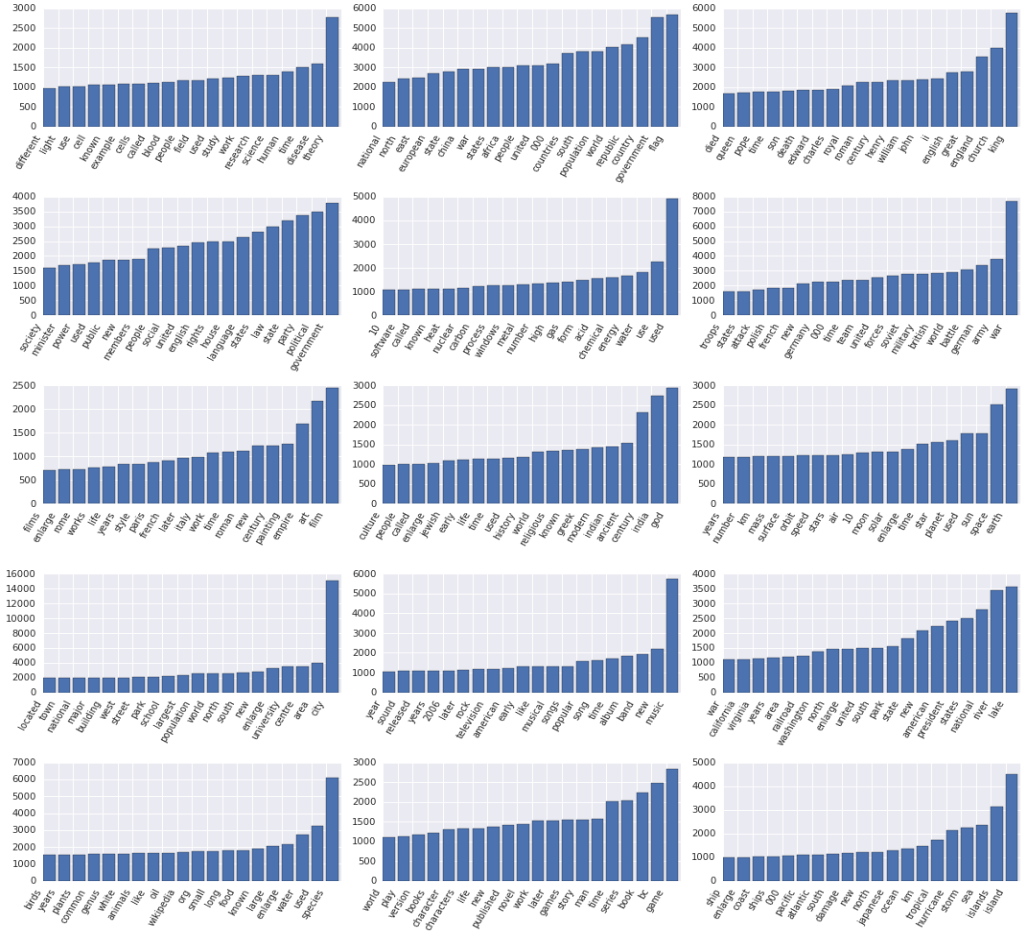

Figure 7: LDA topic divisions and key words

Choosing a Semantic Distance Metric

Though choosing optimal path length to define topological distance was the straightforward choice, we had several options for semantic distance. LDA could capture similarity between two documents by comparing their topic distributions; if two documents are similar, we expected their topic distributions to be similar as well. However, it was difficult to calculate this distribution in sklearn. Additionally, West, Pineau, and Precup used mean observed path length in "*Wikispeedia: An Online Game for Inferring Semantic Distances between Concepts"* so we wanted to use a different metric. Therefore, we decided to use Latent Semantic Analysis (LSA), which is standard in natural language processing, to distill the similarity of source and target articles.

Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) enables us to represent each article as a vector in a concept space. The importance of this analysis is that it reduces each vector to highlight the most important categories. The end goal of this analysis is a square, symmetric matrix that contains a

semantic distance value between 0 and 1 for every combination of source and target articles. The closer the value is to 1, the more similar two articles are; the closer it is to 0, the more different they are.

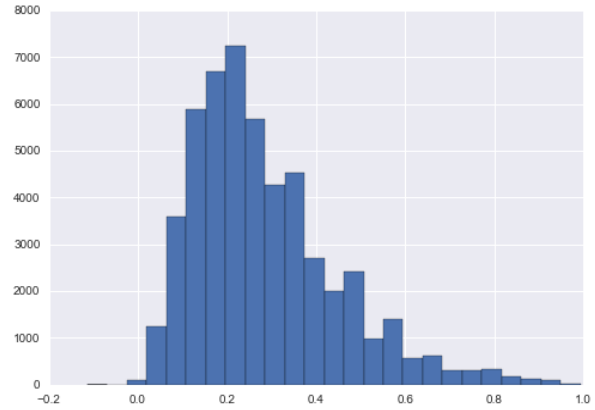

Figure 8: Distribution of semantic distances

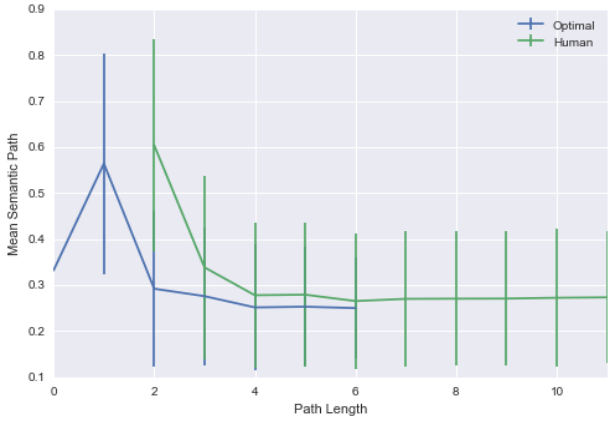

Figure 9: Semantic distance for various optimal and observed path lengths

For our analysis, we first read in each article and vectorized it with CountVectorizer() to create a matrix of article vectors. This matrix could be decomposed via a Singular Value Decomposition (SVD), which helped us describe each article as a weighted combination of different concepts that the SVD determines. We reduced the dimensions of these concepts in order to reduce noise that appeared from the decomposition method. Then, we took the cosine similarity between the reduced vector representations with each other to get our scalar semantic distance. Once we had a semantic distance between two articles, we assigned each game a semantic distance depending on their source and target article.

The distribution of the semantic distances for all of the games looked like a skewed normal distribution centered around a semantic distance value of 0.25 (Figure 8).

Furthermore, we intuited that the more similar the source and target article, the easier the game and the shorted the expected path length. We plotted semantic distance against path length--both human and optimal, and this confirmed our intuition because both path lengths increase as semantic similarity decreases (Figure 9).

Regression

We did linear least squares regressions to address our goal of predicting player performance using semantic distance. Our two metrics for player performance were game duration, in time (seconds), and game path length. We split the data into 80% train and 20% test, fit the regression on the training data, and plotted the predictions for the test data (Figure 10a-b). Neither the regression with semantic distance as the explanatory variable and game duration as observed

variable, nor the regression with semantic distance as the the explanatory variable and game path length as the observed variable, were particularly successful--they had R^2 values of 13% and 18% respectively, indicating weak linear fitting.

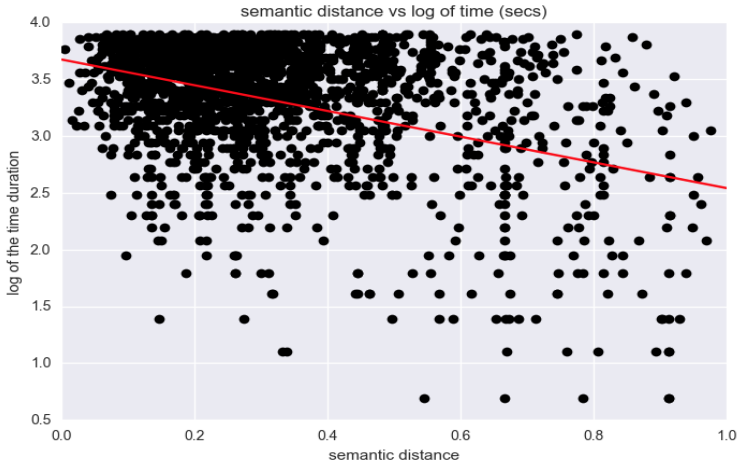

Figure 8: Regression for game duration

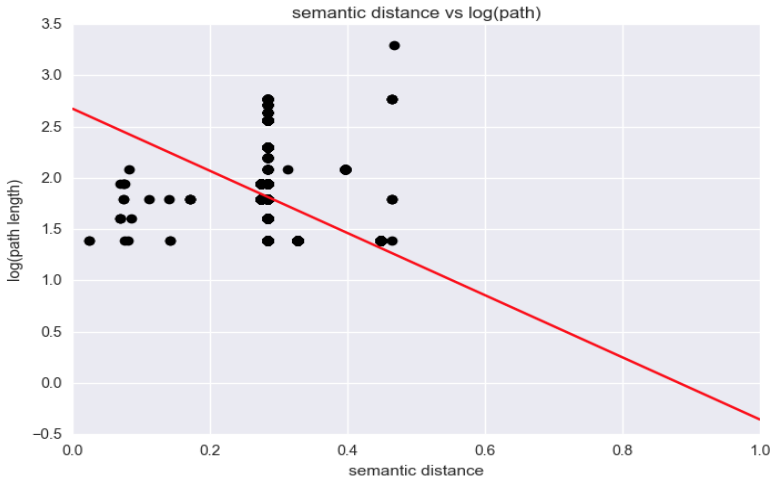

Figure 10: Regression for game path length

Game Dynamics

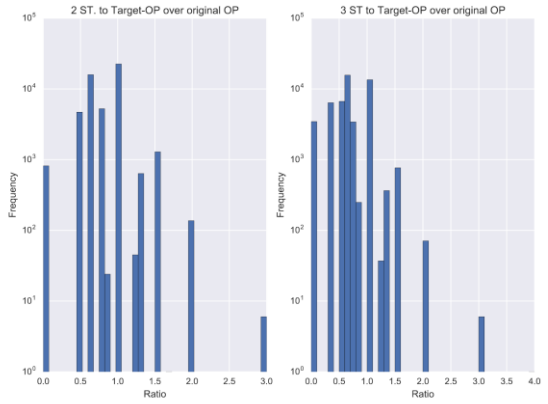

Figure 11: Dynamic analysis

As our regressions did not yield strong correlations, we shifted our focus from analyzing the static distances between source and target articles to analyzing the game dynamically -- as a player progresses through the game, how does distance between their current position and target article change? We chose to use topological distance rather than semantic distance because this information was directly from the dataset and our regressions using semantic distance were weak.

To analyze this, we calculated the ratio of position2-to-target optimal path length and source-to-target optimal path length as well was the ratio of position3-to-target optimal path length and source-to-target optimal path length for all games. If the ratio was over 1, it meant the player was further from the target article; if it was under 1, the player was closer to the target article. We plotted a histogram for this ratio and found that as expected, more players had ratios under 1 at position 3 than position 2 (Figure 11). Additionally, we found two peaks in each histogram, indicating that perhaps there are two groups of players--stronger performers and weaker performers.

Computational Complexity

Regarding the computational complexity of our analysis, many of our algorithms were primarily of the form groupby/apply/count. While we do not know the underlying algorithms that pandas, numpy and dplyr use, we do know that a reasonable upper bound is O(N), where N is the number of rows in a column. There were instances, however, in which we had to apply functions across rows and columns of data and we expect these to be between O(N) and $O(N^2)$ since we were iterating across rows and columns but not the whole matrix. Our LDA and LSA scripts took a while to run since for each article, we had to iterate across each article's raw text. The runtime is on the order of magnitude of O(number of articles * average number of words per article).

Discussion

Our exploratory analysis on the Stanford Wikispeedia dataset clarified that in order to understand how humans navigate through information systems, we should explore differences between semantic distance, or how close in meaning two articles to one another, and topological distance, or the distance between two nodes in the graph.

To determine whether semantic distance could be a more effective way to categorize Wikipedia articles, we ran the topic model Latent Dirichlet Allocation, which generated topics that mapped well to the categories given by the dataset. To determine whether semantic distance could be a good predictor of human game performance, we expressed semantic distance using Latent Semantic Analysis and ran regressions between semantic distance and game path length and semantic distance and game time duration. Both regressions yielded weak correlations on the test data.

Future work could include running additional regressions using topological distance as the explanatory variable to predict player performance; we did not address this because our data on topological distance was heavily skewed towards optimal path lengths of 3, 4, and 5. Additionally, it would be interesting to analyze higher performing players versus lower performing players, which we discovered when analyzing the game dynamically.

We found this dataset extremely interesting to analyze, not only because it is relevant to understand human navigation of information systems with today's increasing information growth rate, but also because the way humans think can be mapped out as a network similar to that of Wikipedia, where topics and ideas are related to one another through intermediate paths. This means that analyzing the human intuition behind navigation of such articles is a step in understanding how humans think and process information.

References

1. http://wikispeedia.net/
2. http://snap.stanford.edu/data/wikispeedia.html
3. Robert West and Jure Leskovec. *Human Wayfinding in Information Networks.* 21st International World Wide Web Conference (WWW), 2012. http://infolab.stanford.edu/~west1/pubs/West-Leskovec_WWW-12.pdf
4. Robert West, Joelle Pineau, and Doina Precup. *Wikispeedia: An Online Game for Inferring Semantic Distances between Concepts.* 21st International Joint Conference on Artificial Intelligence (IJCAI), 2009. http://infolab.stanford.edu/~west1/pubs/West-Pineau-Precup_IJCAI-09.pdf
5. https://cseweb.ucsd.edu/~jmcauley/cse255/reports/wi15/Shelby_Thomas_Moein_Khazraee.pdf
6. http://www.cs.columbia.edu/~blei/papers/Blei2012.pdf