Simon Fraser University

An Analysis of Wikipedia Article Vital Scores

Nicholas Chan (301543674)
nca100@sfu.ca
CMPT 353, Summer 2025
Final Project Report

Overview

On the English Wikipedia website, an interesting and largely unintentional phenomenon can be found: if you click on the first hyperlink in the main text of an article, and repeat this process for subsequent articles, you will almost always end up on the article on Philosophy. This begs the question on whether or not how close a Wikipedia Article is to Philosophy determines how foundational it is.

Wikipedia also maintains a user-reviewed, albeit incomplete list of "Vital Articles", which organize articles based on their perceived importance to the website, where articles are rated on a 1 to 5 scale, based on their overall importance to the site.

In this project, I aim to use this "Path to Philosophy metric" as a basis alongside the vital score to examine whether or not an article being closer to Philosophy is associated with a higher perceived importance, and therefore more foundational.

To answer these questions, I will collect a large sample of articles, recording the number of steps it takes to reach the final attractor, Philosophy, and recording the data along the way. A series of statistical tests to first determine proper values in determining how "foundational" an article is, and using machine learning techniques, attempt to classify the data in order to get a bigger picture of what articles are most important on Wikipedia.

Data Collection

In order to collect the data for this analysis, two sources were needed. Of course, clicking on the first hyperlink within an article's main text isn't something that we can get from an API, and for this reason, a web scraper was first used to automate this process.

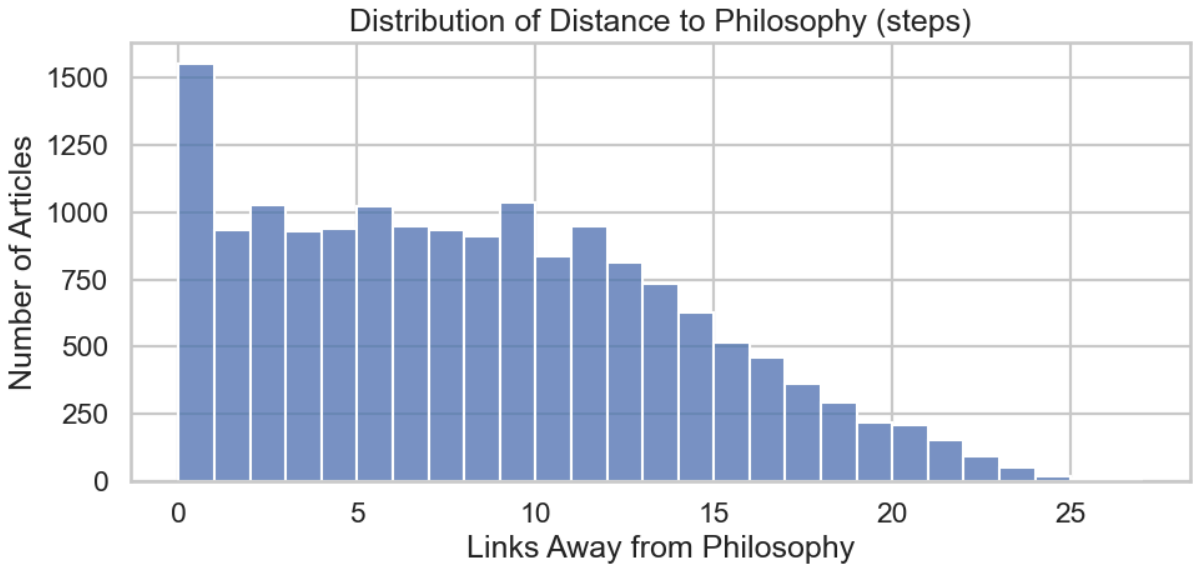
To ensure my data was free of bias, I needed to ensure I had a large amount of Wikipedia pages chosen at random. Fortunately, Wikipedia has an easy way to query a random article through their Special:Random URL (https://en.wikipedia.org/wiki/Special:Random), so this was a trivial task.

Every run of the web scraper would query up a random article, and keep note of its title. By doing this, I was able to get vital scores and additional metadata such as creation dates or article length for each article by querying the title to Wikipedia's MediaWiki API. This data, along with the data gathered from clicking the hyperlinks, was gathered and stored into CSV files. If the Philosophy article was reached, I was also able to calculate how many steps away each article was from Philosophy.

This scraper and API combo was run 970 times, resulting in 15, 271 page visits for us to analyze.

Analysis

Most of the articles in the data seem to be within a few links away from the Philosophy article, with a gradual drop-off in counts as the path length grows.



To check whether or not these articles closer to the Philosophy page tend to have higher vital levels than those further away, we define the following hypotheses:

H_0: Pages further from Philosophy tend to have lower or equal vital level scores compared to those closer to Philosophy.
H_A: Pages further from Philosophy tend to have higher vital level scores compared to those closer to Philosophy.

Note that a score ranges from the highest score of 1 to the lowest score of 5. To evaluate the hypothesis, I first applied a Mann-Whitney U-test to compare the distribution of vital scores between articles that were closer (took lesser steps) to reach the Philosophy page, and those further away. Since the data does not seem to be normal, a non-parametric test may help us understand our results better.

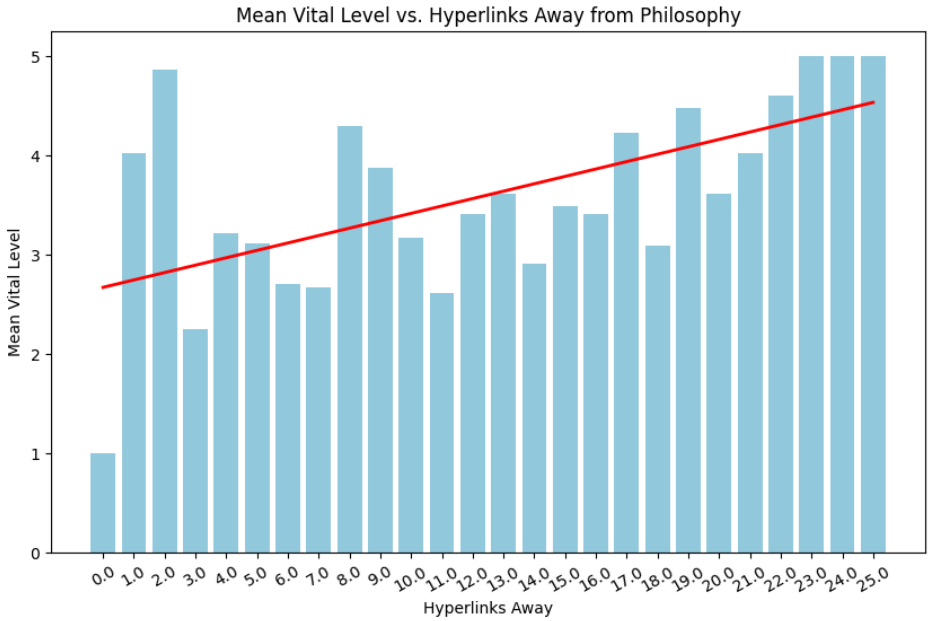
We can further confirm this relationship with an Ordinary Least Squares (OLS) regression to model the relationship between the number of links away from the Philosophy Wikipedia Article and the vital score. By finding the slope of the line, it may help to see what direction the data seems to be going.

Results

The Mann-Whitney U-test produced a p-value of 0.216, meaning there is no statistically significant difference between the two groups, meaning we fail to reject the null hypothesis in this test. However, Using an ordinary least squares regression instead, we can get a result with a p value of 5.43e-8, which is also far beneath our 0.05 significance threshold. Based on this regression model, we reject the null hypothesis and accept the alternate hypothesis, confirming that vital levels are higher the further away from Philosophy they are.

We can further confirm this by plotting the mean vital levels of articles compared to how far they are from the Philosophy page on a histogram. By fitting a line with regression, we can see indeed that a positive slope grows towards larger hyperlinks away. This relationship makes sense, as pages that are closer to Wikipedia's Philosophy page are often more foundational and related to a great number more pages, giving them a higher importance and thus a vital level of 1.

It should be noted that the mean vital level at 0 hyperlinks away is consistent, as 0 hyperlinks away refers to only the Philosophy page itself, which according to Wikipedia's API already has a vital score of 1.



Limitations & Issues

An issue with using a webscraper to obtain data is that the data may often come up in strange formats. Wikipedia often has additional notes or hyperlinks before the main text content of an article which could potentially lead it into an infinite loop. As such, the scraper needed to take this into consideration. Another issue I realized too late into the project is that unfortunately, the "leading to philosophy" phenomenon does not have much relation with other variables, resulting in not a lot of analysis regarding the actual philosophy phenomenon.

Project Experience Summary

- Developed a Python-based webscraper and API integration pipeline to collect and analyze Wikipedia's "Getting to Philosophy" phenomenon.

- Applied Statistical Analysis techniques such as OLS regression and a Mann-Whitney U-test to evaluate the relationship between article distance and importance variables.

- Created data visualizations using Seaborn to illustrate trends in vital scores and hyperlink distance.