Algorithmic Study of the Relation of Peaceful and Violent Subjects to Philosophy Using Wikipedia and a Python Crawler

Antoine Boucher and Dino Ronco

April 8th, 2017

Abstract

By clicking on the first link in the main text of any Wikipedia article over and over again, you will eventually reach the Philosophy article. We have created a computer program to find the relation of peaceful and violent subjects to Philosophy by analyzing their distribution of click distances. After testing this with an assortment words related to both concepts, our results show a strong pattern that violent subjects are closer to Philosophy than their peaceful counterparts.

1 Introduction

The Internet revolutionized our daily life, not only in making it more productive and more entertaining but also in making us better understand our own humanity. This is happening due to "big data", which refers to the nearly infinite amount of information created continuously by people around the world. This data covering nearly all human knowledge is a highly complex network of information. This"rabbit hole" that is the web can be efficiently analyzed by ageless and simple mathematical theorems. These powerful mathematical tools can uncover patterns in human knowledge, difficulty observable before the advent of the world wide web. Even deep abstract concepts can be patterned form mundane data or words, such as Facebook posts, and better understood to bring us closer to the very essence of humanity.

Wikipedia is the Internet's largest and most popular information source.The free on-line encyclopedia can be described as the best-developed attempt thus far to gather all of Man's knowledge in one location. The multitude of user generated articles are riddled with links that bring you to new ones. This ends up creating a web of knowledge that can transport you from one article to another that may seem completely unrelated at surface value. What if I told you that the center of Wikipedia's interconnected network of knowledge is the article on Philosophy [1]. Over 97 percent of articles on English Wikipedia lead to this central article [3]. While this may seem absurd at first, after testing this theory using a Python web crawler, it is an indisputable fact that almost any article you can think of, with few exceptions, follow this rule.

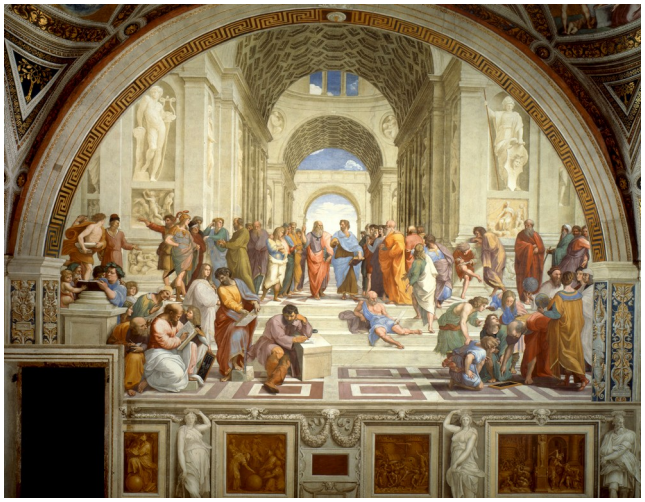


Figure 1: The school of Athens by Raphael, representing the different schools of philosophy and science [2].

2 Prerequisites

In order to understand that Philosophy is the center of all Wikipedia topics and how our research demonstrates this, the following prerequisite knowledge is required:

- Basic Python programming skills and know-how on installing it
- Basic web page structure knowledge
- Basic data format like CSV (comma separated values)
- Basic knowledge of web crawlers
- Knowledge on guidelines on how the lead section of articles should be written
- Convergence series
- Cycle graph

Python is a high-level programming language that is widely used for database searches by companies like Wikipedia and Google. We chose Python over other high level languages because it is relatively simple to use and its uses are well documented on the Internet. Also, it has a great library of software functions for processing web pages, string manipulation and exporting data in table formats readable by Excel.

A Web crawler, or spider, is a program that browses web pages linked to each other and that systematically extracts and catalogs information from each page. This is a first step in organizing information from the Internet. This is in exactly what "Google Search" does on a massive scale, indexing the complete web and storing this information in a database that can be quickly searched by "googling" information. Similarly all the Wikpedia web pages can also be crawled and this is how, in 2008, a wikipedia user discovered the surprising phenomena that all wikipedia web pages "converge" to the Philosophy article!

The Wikipedia convergence phenomena to Philosophy is a form of mathematical series. There multiple ways to explain a series.The characteristic of series is evaluated by examining the finite limit or infinity limit.If a series goes near a certain value call this a converging series.These two type of converging series: conditional and absolute.The type depend on the scale of finite for conditional and infinite for absolute.Conditional convergent series can also be converted to an absolute convergent series.

Directed cycle graph is a part of graph theory.It is a version of a cycle graph with direction that follow the cycle. It really useful for series that loop around themselves and visualizes this phenomenon.The component of cycle graph are the vertices and edges .The vertex are the point that are link by a directed edges. A vertex can be connected by two edges.

Ever since this discovery, mathematicians and programmers have been intrigued by this convergence to the "mother of all science". However, most studies on the subject have been challenged by issues such as loops or cycles due to poorly structured web pages or poorly written articles.

In our program, we implemented the following algorithm:

- Random selection of wikipedia articles on violent or peaceful subjects
- Navigate to subject page and click on the first link in the main text of each article
- Ignore external links, links to the current page, or red links (links to non-existent pages)
- Stop when reaching "Philosophy", or a page with no links, or a page that does not exist, or when a loop occurs
- Count the number of clicks it took to get to the "Philosophy" article

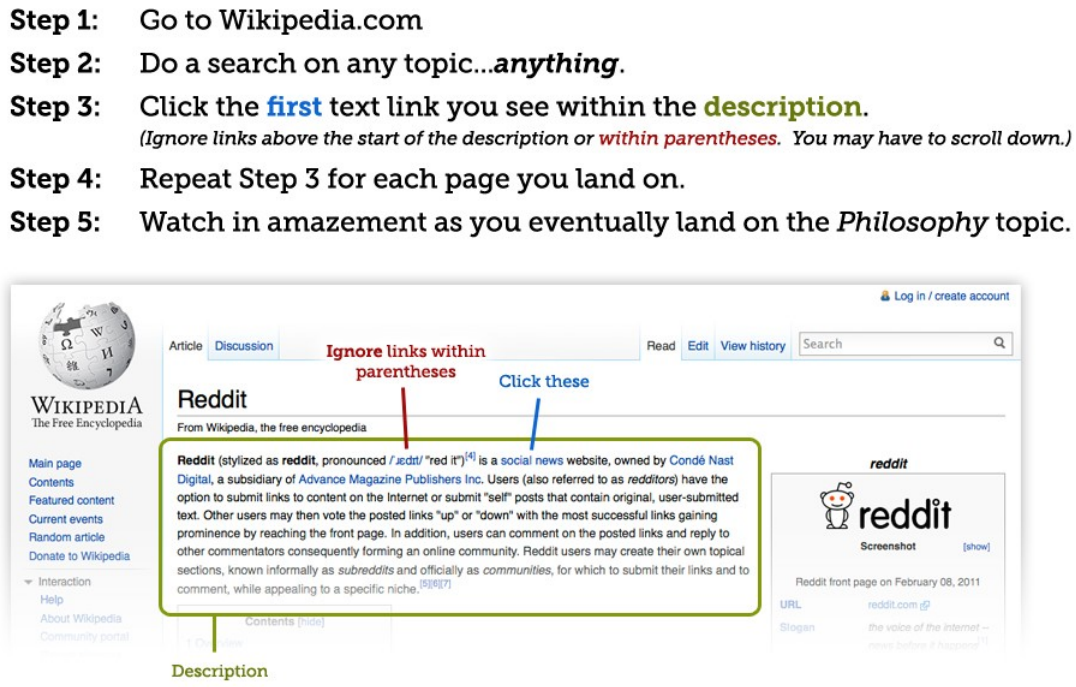


Figure 2: A
meme posted on Reddit referencing this phenomenon [4]

Python Code

item.py

```
import scrapy
from scrapy.item import Item, Field

class WikipediaItem(Item):
    title = Field()
    link = Field()
```

spider.py

```
import scrapy
from scrapy.spider import BaseSpider
from scrapy.selector import HtmlXPathSelector
from bs4 import BeautifulSoup
from wikipedia.items import WikipediaItem

class Spider(scrapy.Spider):
    name = "wikipedia"
    allowed_domains = ["en.wikipedia.org"]
    start_urls = ["https://en.wikipedia.org/wiki/Wikipedia"]

    def parse(self, response):
        yield {
            'title':response.css('div#mw-content-text > p > a')
            .xpath('@title').extract_first(),
            'link':response.css('#mw-content-text > p > a')
            .xpath('@href').extract_first(),
              }
        next_page = response.css('#mw-content-text > p > a')
        .xpath('@href').extract_first()
        if next_page is not None:
            next_page = response.urljoin(next_page)
            yield scrapy.Request(next_page, callback=self.parse)
```

2.1 How It Works

We begin by installing the Python "scrapy" library for web crawling and data extraction.

In our case, we are interested in designing a specialized web crawler, that scrapes only the information we need from articles, specifically the title and the first valid hyper link on the page.

We created a "scrapy" project by setting the "item" name to Wikipedia and creating two storage fields for the article title and the first valid link. This information is then stored in a simple Comma Separated Value (CSV) file format, i.e., the title information is separated from the hyperlink by a comma, and each title-hyperlink pair is separated by a carriage return. CSV files are readable by table manipulation programs like Excel.

For the spider that will crawl Wikipedia, we first created a new folder named spider and created a new python file named spider.py. In this file, we initialize the starting URL and the allowed domain.

(a) Wikipedia folder(main folder) (b) Spider folder

Figure 3: Directory of the scrapy project

An HTML selector and xpath are used to select the first link of each page that we crawl and save. HTML selector and xpath are used in Cascading Style Sheets (CSS) and JavaScript to add functionality or visual style to a specific group in the web page.

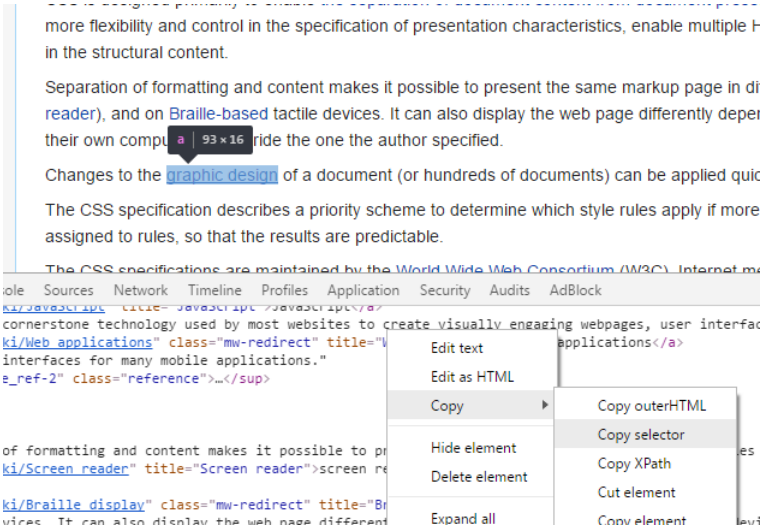


Figure 4: Selector use in the website.

Using a simple command I can start crawling and choose the format in which to store the extracted data.

```
c:\User\Name\>cd Desktop
c:\User\Name\Desktop>cd wikipedia
c:\User\Name\Desktop\wikipedia>scrapy crawl wikipedia -o wikipedia.csv
```

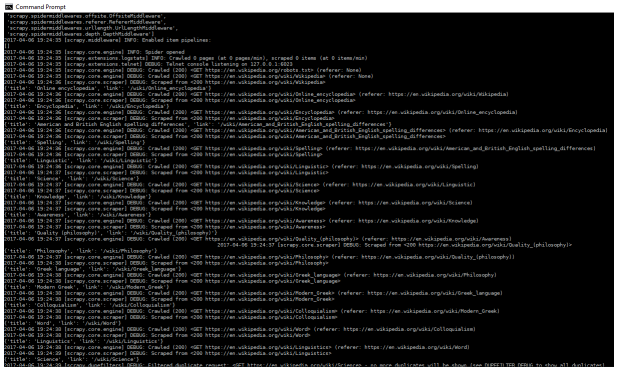


Figure 5: Example of my crawler being used

Other Wikipedia crawlers that do a similar job can be found online. For example: the node.js library also named "Getting-to-philosophy" [10] or

"Xefer" that makes a tree like fractal structure with the names of the articles that were crawled [9] or "wikiloopr" that finds cycles in the First-Link network [7]. All of those crawlers have issues of the same sort to mine.

3 Results

In the analysis of results from our program and through manual testing of 75 random articles in French and English we found the following conjecture:

Conjecture 1. *Violent concepts are closer to Philosophy than peaceful concepts*

This major finding in our research, that violent Wikipedia articles are more closely related to Philosophy than their peaceful counterparts is statistically clear. On average, fewer clicks are required for violent articles (13.7) to reach the Philosophy article than the peaceful ones (15.3). Only once did a peaceful set fare better than a violent set. On average, peaceful historical figures, e.g., Gandhi, required 16.4 clicks compared to violent counterparts which required a slightly higher 16.6 clicks. What was striking, however, was the fact that all the violent sets (except for the set of user provided articles) had a higher standard deviation (2.7) than the peaceful ones (1.4). This stems from violent articles having keywords related to radical ideas and that Wikipedia doesn't enforce a standard procedure to write an article, something that will be further discussed below.

Similar results were also found in French articles but with fewer clicks required for violent articles (10.6) as compared to peaceful articles (10.9). Interestingly we found a reference citing that for French, one reaches Philosophy only 85% of the time, as compared to 97% in English [5]. it is simply because there are significantly less articles in French (1 million as compared to the 5 million in English) which creates gaps in the French network to Philosophy? or/and is it because French speakers are complying less to the standard procedure in writing articles? [5]

We observed another interesting finding in is that the first link out of the Peace treaty set of articles almost invariably is through war articles! With

some reflection this is not too surprising since Peace is an abstract concept that could only have been created after the realization of what Violence was. All in all, however our results are interpreted, seemingly philosophical concepts are the root of all words and concepts.

Table 1: Relation between violent concepts and Philosophy

Concept	Clicks on English Wikipedia	Clicks on French Wikipedia
War	8	9
Genocide	15	6
Violence	13	10
Racism	11	7
Anger	4	12
Average	10.2	8.8
Standard Deviation	4.3	2.4

Table 2: Relation between peaceful concepts and Philosophy

Concept	Clicks on English Wikipedia	Clicks on French Wikipedia
Peace	10	3
Diplomacy	17	19 (through English articles)
Pacifism	9	4
Equality	15	2
Happiness	12	10
Average	12.6	7.6
Standard Deviation	3.3	7.1

Table 3: Relation between violent historical figures and Philosophy (X means that Philosophy was never reached due to a cycle)

Historical figure	Clicks on English Wikipedia	Clicks on French Wikipedia
Adolf Hitler	20	9
Joseph Stalin	20	X
Attila the Hun	17	15
Mao Zedong	8	X
Caligula	18	6
Average	16.6	10
Standard Deviation	5.0	4.6

Table 4: Relation between peaceful historical figures and Philosophy

Historical figure	Clicks on English Wikipedia	Clicks on French Wikipedia
Gandhi	13	10
Mother Teresa	17	11
Buddha	19	9
Jesus Christ	16	11
Nelson Mandela	17	11
Average	16.4	10.4
Standard Deviation	2.2	0.8

Table 5: Relation between deadly wars and Philosophy

War	Clicks on English Wikipedia	Clicks on French Wikipedia
World War I	10	10
Spanish-American War	13	12
Napoleonic Wars	14	12
Thirty Years War	22	13
Seven Years War	9	15
Average	13.6	12.4
Standard Deviation	5.1	1.8

Table 6: Relation between peace treaties and Philosophy

Peace treaty	Clicks on English Wikipedia	Clicks on French Wikipedia
Treaty of Versailles	13	15
Treaty of Paris (1898)	14	15
Congress of Vienna	21	15
Peace of Westphalia	19	14
Treaty of Paris (1763)	9	16
Average	15.2	15
Standard Deviation	4.8	0.7

Table 7: Relation between various violent articles and Philosophy

Article	Clicks on English Wikipedia	Clicks on French Wikipedia
Gun	25	7
Holocaust	10	14
Suicide	11	9
Genocide	12	6
Apartheid	14	13
Average	14.4	9.8
Standard Deviation	6.1	3.6

Table 8: Relation between various peace articles and Philosophy[6]

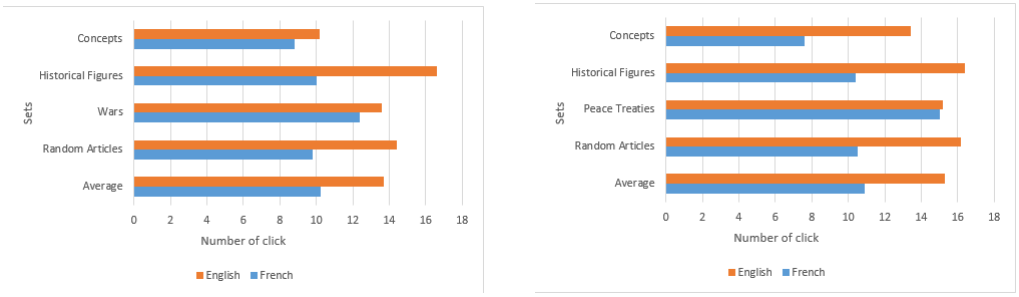
Article	Clicks on English Wikipedia	Clicks on French Wikipedia
Greenpeace	10	18
Nihonga	24	13
John Lennon	22	9
Hippie	5	X
Declaration of Human Rights	20	12
Average	16.2	10.5
Standard Deviation	8.3	3.7

Table 9: Average number of clicks for each of the violent Sets

Set	Clicks on English Wikipedia	Clicks on French Wikipedia
Concepts	10.2	8.8
Historical Figures	16.6	10
Wars	13.6	12.4
Random Articles	14.4	9.8
Average	13.7	10.25
Standard Deviation	2.7	1.5

Table 10: Average number of clicks for each of the Peaceful Sets

Set	Clicks on English Wikipedia	Clicks on French Wikipedia
Concepts	13.4	7.6
Historical Figures	16.4	10.4
Peace Treaties	15.2	15
Random Articles	16.2	10.5
Average	15.3	10.9
Standard Deviation	1.4	3.1



(a) Violent sets (b) Peaceful sets

Figure 6: Average number of clicks for all sets for both concepts.

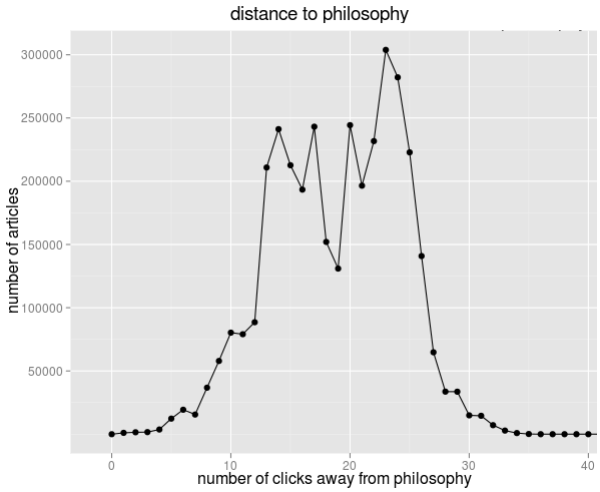


Figure 7: The relation between the number of clicks to reach Philosophy and the number of articles[11]

We found an average of (13.7) clicks to Philosophy for the violence sets and (15.3) clicks for the peaceful set. Comparing this to the data in the above graph, we observe that our measurements are not far off even though our sample set was much smaller.

Conjecture 2. *The First-link network of Wikipedia rooted at Philosophy has a conditional convergent characteristic. When it reaches the convergent point, Philosophy, a cycle is created that loops around itself with vertexes that are the same as the last network vertices to reach this point.*

Wikipedia has a deep interconnected network that goes from precise subjects to subjects with wider and wider scope. This network can be visualized in multiple ways, but most of these visualization will have a fractal structure. As the saying goes all roads lead to Rome. We can also see wikipedia articles as a current in a river. Most rivers will reach bigger rivers to eventually reach the ocean. The ocean has it own current that loops back on itself.

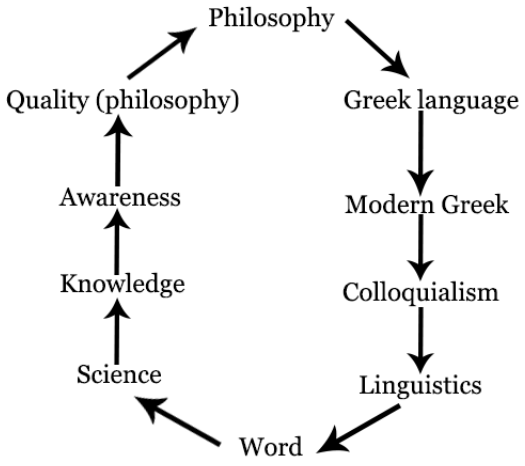


Figure 8: Directed cycle graph of the article on Philosophy

Just like river and ocean metaphor we observed the same thing in our results. The convergence process doesn't stop at Philosophy, rather once we reached it we then go through a sequence of articles that loop back to Philosophy. Thus the overall network of articles going to Philosophy is a converging series but with a particular type of convergence. The convergence point is a finite point or article, but continues on with an infinite cycle on itself. This type of series could be a new discovery in mathematics.

What is also interesting is that the articles in this cycle, which include Greek, Language, Science, Awareness and a number of others, are also most likely to be visited from any article, on the path to Philosophy.[7]

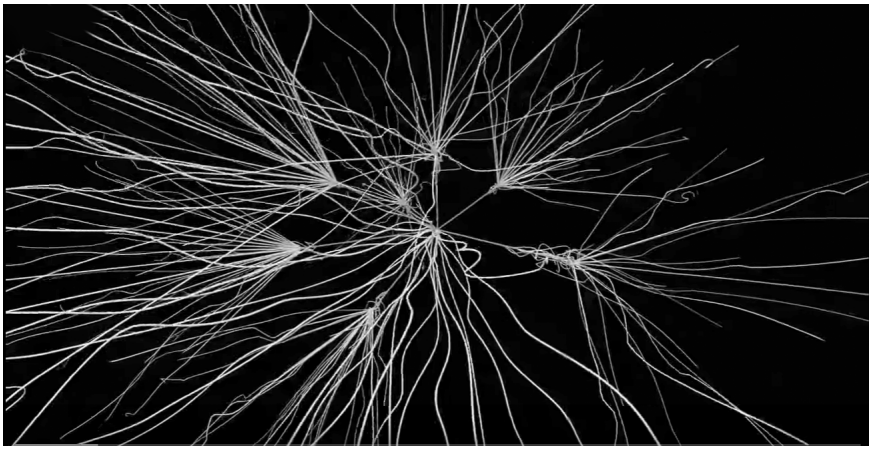


Figure 9: A
3D approximation of the tree-like First-Link network of Wikipedia converging to Philosophy. [8]

4 Conclusion

4.1 Analysis

We expected to find that peaceful subjects would be closer to Philosophy, probably because we inherently believe that peace, or good, should be the foundation of philosophy, however the contrary turned out to be true. We don't know the exact reason why violent subjects are closer to Philosophy, but we theorize that it might be that the peaceful subjects cannot be defined without the concept of "conflict". In other words, peace wouldn't exist without the concept of violence. There was a second important observation from our results in that the network of first-links most always goes to Philosophy, but after this convergence, there is a cycle out of Philosophy and back on itself.

4.2 Problems Encountered

The problems encountered are mostly due to challenges in defining the rules of the algorithm because of the slackness of article authors to adhere to Wikipedia guidelines. Basically each Wikipedia article is constructed by different authors with different views on how to structure articles. The first link to click to choose is not always clear, for neither a human or computer, since the conditions that makes it the right link to click is subjective. The hand-full of algorithms found on-line and the Python crawler we created, all demonstrate that Philosophy is the network root, however, the number of clicks to reach Philosophy varied from one approach to another. This issue is caused by the lack of a general cohesive structure in Wikipedia articles. While some articles can start with the etymological root language of the word (e.g. Greek or Latin), others have no clear pattern on which word is designated as the first link. This problem becomes much more apparent when trying to run the algorithm on French Wikipedia articles.

The second problem encountered stems from the fact that there still exists 3 percent of English articles that never lead to "Philosophy". A notable one we encountered was the article on Genghis Khan. This infamous loop of "Mongolian" to "Mongolian Script" to "Mongolian Cyrillic" was caused by the fact there is no article or link that describes what a "Script' is. In brief, the cause of these errors stems from the fact that the editors of Wikipedia articles don't follow a standard writing guideline and the rare occurrence when an article doesn't exist. For French, we observed that only 85 percent of articles converged to Philosophy. This might be because they contain less reference links, in turn maybe because there 5 times less articles in French than in English. These 15 percent of articles that

didn't converge ended up in a never-ending loop, that didn't lead or include the "Philosophy" article.

4.3 Potential Paths

Wikipedia is the largest and most popular general reference work on the Internet. Therefore, making a clear and standard format to create articles is of the utmost importance. The quality of any article is based solely on the number of the editors and their knowledge on the subject. To make a good Wikipedia article, the page should follow a standard guideline and have a lot of internal references to lead to further articles to enrich research.The way articles are structured is really important for people to find information more efficiently. We observed the French collection of articles to be somewhat more disorganized than for English. Our work shows that making Wikipedia more structured will make it easier for bots or AI to do data mining and learn pattens from this rich knowledge bank. Also, we should explore other on-line encyclopedias and dictionaries to see if we observe the same or similar phenomenon as on Wikipedia. By exploring multiple sources as a whole, we could possibly further clarify how information is interconnected, by the increased variety of ways describing it. We could also analyze many other languages to better understand the differences of one language to another. The only limit to this quest in understanding is ourselves and how we want to see the world.

Acknowledgements

We wish to thank Patrick St-Amant for guidance while we developed our paper. We also wish to thank the philosophy faculty at CiSA for giving us the inspiration to write a mathematical paper on philosophy. Finally, we thank our classmates who encouraged us along the way.

References

[1] Wikipedia, Wikipedia:Getting to Philosophy, Retrieved on April 5, 2017, https://en.wikipedia.org/wiki/Wikipedia:Getting to Philosophy.

[2] Wikipedia, Philosophie https://fr.wikipedia.org/wiki/Philosophie.

[3] D. Brezhnev, S. Trusheim, and V. Yendluri. All paths lead to philosophy. part of the Stanford Network Analysis Project, 2013.

[4] Pixelcrak, Try this Wikipedia MindFuck, Reddit, April 13 2011, https://www.reddit.com/r/pics/comments/gpdhb/try hiswikipedia mindfk

[5] D. Lamprecht, D. Dimitrov, D.Helic, and M. Strohmaier. "Evaluating and Improving Navigability of Wikipedia." Proceedings of the 12th International Symposium on Open Collaboration, 2016. http://www.daniellamprecht.com/wp-content/uploads/2016/08/Evaluating-and-Improving-Nav

[6] A. Boucher, D.Ronco, Getting to philosophy, Google form, 30 March 2017. https://docs.google.com/forms/d/e/1FAIpQLSfbO8AfXw29iJM4KgUKCYviBRZPGPEwFwj5uRmcE703C0

[7] S.Gransee, WikiLoopr, 2012, http://www.wikiloopr.com/.

[8] Luciano Floridi, The joy of Data, BBC Documentary, 18 December 2016, https://youtu.be/l6oKriR-RjM.

[9] Xefer, wikipedia tree, https://xefer.com/wikipedia.

[10] getting-to-philosophy v2.1.1, Node.js, https://runkit.com/npm/getting-to-philosophy.

[11] M. Kelcey, do all first links on wikipedia lead to philosophy?, August 13, 2011, http://matpalm.com/blog/2011/08/13/wikipedia-philosophy/.