

# Determining Correlation Between Wikipedia Categories Based on Shortest Path of Hyperlinks

Adam Lang

[adam.lang@temple.edu](mailto:adam.lang@temple.edu)

## Abstract:

Wikipedia is a web service that provides a free encyclopedia that is available for all to use. Hyperlinks are used to connect two articles that have similar or related content. The Stanford Network Analysis Project created a dataset to model Wikipedia hyperlink structure and hyperlinked articles, further grouped into categories of related articles. This project seeks to understand underlying connections between these categories based on the average shortest path length of hyperlinks between articles of different categories, specifically the most popular and a random set. Results exhibit evidence to support related categories will have similar rank structure as well as average path lengths when being compared against a set of categories indicating there may be a deeper link between related categories and average shortest path length.

*Supplementary Materials:* [https://github.com/adamlang739/CIS5524\\_FinalProject](https://github.com/adamlang739/CIS5524_FinalProject)

## Introduction:

Wikipedia is a multilingual free online encyclopedia which has become one of the most popular websites on the internet since its inception. Wikipedia is the home of many informational articles that have hyperlinks to other website articles. ‘The Wiki Game’ is an online game where the player is given two random Wikipedia articles. The goal of the game is to get from the starting article to the ending article using only hyperlinks between them in the shortest possible time (ex. ‘Futurama’ to ‘Network Topology’). The game inspires interesting questions about how those two topics may be related, and how one might get from one topic to the other. There may be many underlying connections between the two topics that are not blatantly obvious, which has become the main inspiration behind this project.

The connections between Wikipedia articles through hyperlinks can be represented as a directed network. Stanford Network Analysis Project (SNAP) created a dataset [1] which models the Wikipedia network. The dataset collected is a web graph of Wikipedia hyperlinks collected from the website in its state as of September 2011. The network was formed by taking the largest strongly connected component. Once found, the developers of the dataset restricted the network to pages that fell within the top set of categories. Categories are defined as groups of information that have at least 100 pages that relate to the topic. Once the network was restricted, the developers took the largest strongly connected component of that network, meaning that it is possible to reach every node from any other node.

The network contains approximately 1.79 million nodes, which represent Wikipedia articles/pages, and 28.5 million edges, which represent directed connections between pages via a hyperlink. The dataset also includes a set of 17,364 categories and a list of pages associated with each category. Articles may be a part of a single category or multiple categories. However, during

analysis, if a category is present in multiple categories, it was kept as the samples were large enough that outliers would have a minimal effect on the overall outcome. Categories span a variety of topics, such as 'Eye Anatomy' or '2002 American Television Series Endings', with varying lengths of how many articles each category has ranging from only a few to thousands.

The overall goal of the project is to find correlations between different & top/most popular Wikipedia categories based on the shortest path between them to understand underlying similarities, differences, and connections that may be unknown. Correlation in this project is defined as the average shortest path length between all the nodes of two distinct categories. Categories are groups of articles that are similar, so it is a reasonable conclusion that average path length between categories will be a good metric for understanding how topics are connected. Given that the network is one strongly connected component, there will always be a route or path connecting topics and categories.

While average shortest path is the main metric for this study, it is important to consider more deeply what is being measured: number of hyperlinks connecting topics. A hyperlink is a digital reference to data. Most hyperlinks on Wikipedia exist as links to other pages on the website that have been referenced in an article, for example, 'Insertion Sort' is hyperlinked in the page 'Sorting Algorithms'. When there is a direct hyperlink, the pages will be closely related or have a connection. The closer related two topics are, the less hyperlinks there will be between them. For example, 'Insertion Sort' -> 'Sorting Algorithms' -> 'Computer Science', having a path length of 3, one can draw a conclusion that 'Insertion Sort' and 'Computer Science' closely related, which is true. Average path length of hyperlinks between topics becomes a good indicator of relatedness of those topics.

The following experiments and work are done to obtain a better understanding of relatedness between Wikipedia categories. An interesting sub-topic that is studied in relatedness research is semantic relatedness. Semantics is defined as the meaning of language and how an individual understands a combination of lexical characters. Semantic relatedness (SR) aims to measure some form of lexical or functional association between two words or concepts based on contextual or semantic similarity regardless of syntactical differences. SR is an important topic that allows for further study into how people understand language as well as enhancing how computers or deep learning models may better grasp linguistic relationships on a quantitative level. Some motivations behind this research are to begin understanding if there is a relationship between the hyperlink structure of Wikipedia and semantics, if average shortest path between two Wikipedia categories has a connection to the semantic properties of the categories, and how Wikipedia hyperlinks exhibit relatedness between topics. Conducting this research could aid NLP research, ideation, explorative research, finding related sources for more refined citations, etc.

### Related Work:

The main goal of this study is to look more closely at relatedness between Wikipedia categories. Relatedness is the state of being connected and is a common metric that is widely used across multiple areas of study including computer science, linguistics, biology, etc. In the

literature, Wikipedia has become a popular medium for studying relatedness, such as semantic or entity relatedness. The following works go further into different utilizations of Wikipedia for studying different relationships of relatedness as well as the uses of Wikipedia categories.

Singer et al. [2] propose a different approach to understanding semantic relatedness through human navigational paths. They utilize data taken from ‘The Wiki Game’, mentioned previously as the game which inspired this project, in the form of game data, showing how players navigated through the website. Their approach to identifying semantic associations differs from previous work into textual or structural information by seeing how users navigate using their intuition and knowledge of real-world concepts/relationships. The basis of study for this paper is the ability for humans to find intuitive paths instead of short paths, exhibiting how human navigation might produce more semantic richness than algorithmic shortest path. Their work concluded that human navigation paths were a viable option for computing semantic relatedness and can often be more precise than going off link structure alone. However, not all human paths were equally useful in indicating SR. While this work focuses primarily on individual paths between nodes instead of categories, it is useful in highlighting alternative methods for finding relatedness and potential explorations in human navigation paths in comparing categories of Wikipedia articles.

Milne [3] proposes a new model to understanding SR, being the Wikipedia Link Vector Model. While this model also uses Wikipedia and desires to find SR, it does so by solely examining the link structure of articles. Milne’s work addresses finding relatedness by finding the angle between the vector of links found within them. Vectors are built using link counts weighted by the probability of each link occurring. The probability is defined by the total number of links to the target article over the total number of articles. So, links are considered less valuable in calculating similarity if many other articles also link to the same target. The novel approach of this paper is evaluated and compared to other standard models, concluding with areas for improvement and future next steps. The work provides an interesting method for finding relatedness but differs from this project because of the use of a link vector instead of the shortest path as the metric.

Agirre, Barrena, and Soroa [4] explore Wikipedia hyperlinks to study relatedness when looking at the entire graph with respect to direct links only as opposed to other filtering methods. The work utilizes Milne’s previously mentioned method of only working with the skeletal/link structure of Wikipedia articles as a means of understanding relatedness. Their work uses a random-walk algorithm as a method for measuring differences between other filtering methods and their proposed approach. The results of this research show comparable outcomes to other highly regarded combined systems giving weight to the potential of using a graph of hyperlinks. The research gives validity to the use of the SNAP dataset in approaching the problem of relatedness between different categories.

Chernov et al. [5] explore semantic relationships by analyzing links between Wikipedia categories for the purpose of improved search capabilities and more meaningful suggestions for those editing Wikipedia pages. They hoped to see how semantics could allow more complex searches as opposed to traditional full-text searches. Complex searches could enhance article generation related to certain topics where full knowledge of the subject is not known by the

individual, for example from the paper: ‘Find Countries which had Democratic Non-Violent Revolutions’. The researchers measured their experiments using the number of links between the categories, studying separately in-links and out-links, as well as a proposed measure called the Connectivity Ratio. They found that in-links showed greater performance in exhibiting semantic similarity as opposed to out-links, and the Connectivity Ratio was a better measure for extracting semantic relationships between categories. Their work shows applications for better understanding Wikipedia categories and a different method for finding semantic similarity.

Yeh et al. [6] turn Wikipedia into a graph to perform random walks to compute semantic relatedness with the goal of combining previous approaches together to improve the efficacy of the process. Previous work has shown the effectiveness of Wikipedia in providing successful measurements of SR, but where this work differs is that they explore all link types when constructing the graph, such as info-box, categorical, and content. The reasoning behind this approach is that some kind of combination of links will produce a more conclusive measure of relatedness. The researchers found, although small, enhanced results when using this method, showing validity to their approach. While this project focused on links between nodes via hyperlinks alone, there may be future work to be done to support smaller shortest paths when using all types of link structures in Wikipedia articles.

Zesch and Gurevych [7] look deeper at the Wikipedia Category Graph (WCG) and how that might be used for a variety of NLP applications and problems. They investigate the different connections to semantic relatedness when diving into the category graph, which exists in a taxonomy-like structure, and the article graph, which exists as a directed graph. The researchers perform various graph-theoretic analyses on the networks to discover that the WCG is a scale-free network, and that it exhibits the small-world property, meaning high clustering and low average shortest path length. They concluded that Wikipedia can be used for NLP tasks like finding semantic relatedness and similarity. Their results and conclusions follow observations made about the SNAP dataset, where some categories act as hubs with many pages in the category while most categories have a smaller number of pages. Zesch’s research helps show that the SNAP dataset can produce interesting results for different NLP tasks.

These works provide a good basis of understanding relatedness and specifically how Wikipedia is a valuable resource for grasping this problem. There are many applications of these works and to the field of Natural Language Processing which allow for advancement in this subfield of artificial intelligence and machine learning.

### Methodology:

The first step in conducting analysis is network creation. The dataset includes an adjacency list detailing all the links between nodes in the network. NetworkX, a Python package made for the creation, manipulation, and study of networks, is the software utilized during this project, and Python is used for programming all necessary functions.

Once the network is created, analysis can begin to find correlation between two categories. First, the shortest path is found from C1-A1 (Category 1 – Article 1) to C2-A1 (Category 2 –

Article 1). The resulting value is stored in an array. Next, the shortest path is found from C1-A1 to every other article in C2. All those values are stored in the previous array, and the average of all those results is taken. The resulting average shortest path is specifically for C1-A1 to every article in C2 and is stored in an array. The same process continues to repeat for every article in C1, continuing to compare against every article in C2, averaged, and stored in the array. Once all average shortest path lengths are found from every article in C1 to every article in C2, the array where all those values are stored is averaged. The resulting number is the correlation or average shortest path length between C1 and C2.

Now knowing how to find correlation, the next step in the process is to find the most popular categories in the dataset. The most popular categories are the ones that have the greatest number of nodes per category. While finding correlations between random categories leads to potentially interesting results, there is greater interest to be found in how the most popular categories are related to each other. One can hypothesize that because of having a large number of nodes, the top categories have many connections between them and are thus more correlated. To further test this hypothesis, it is important to test how the top categories are related to a set of randomly selected categories.

However, a challenge arises when trying to conduct analysis on the top categories, being the number of articles in each category. Each of the most popular categories has thousands of nodes contained in them making the computational cost for finding correlation incredibly expensive. The proper computational resources required to complete this analysis in a timely fashion were not available. The best way to combat the lack of resources was to take a random sample of the articles from each of the top categories to conduct experiments on. A sample of one hundred randomly selected articles were selected from the top five most popular categories. Samples did not include any repeated articles, so each article in the sample represented a unique article in the category. To keep the experiment consistent, the randomly selected categories were required to have at least one hundred articles in them, and if they had more than one hundred categories, then the same sampling process occurred for them as well. Now all categories had one hundred articles to be used for finding correlation. The sampling process allowed for results to be produced within a reasonable amount of time, which can be replicated with ease.

Next, correlations were found using a variety of experiments: top categories compared against only top categories, top categories compared against the set of top categories and random categories, top categories compared against only the randomly selected categories, and finally the random categories compared against the set of top categories and random categories. Correlation was found for every possible combination of categories, excluding a category being compared against itself. These results were then ranked on the basis of correlation. As correlation is defined as average shortest path length, the smaller the path length, the higher the ranking. The top ranked result is designated as the maximum correlation, having the smallest average shortest path between the two categories. The last ranked result is designated as the minimum correlation, having the largest average shortest path length between the two categories. As mentioned previously, related/correlated categories will have a smaller number of hyperlinks between them.

## Results:

Table 1. Top/Most Popular Categories	
<i>Category (in descending order of popularity)</i>	<i>Number Of Nodes/Articles</i>
Living People	418,223
Year of Birth Missing (Living People)	34,721
English Language Films	22,699
American Films	15,302
American Film Actors	13,938

Table 2. Random Categories
Investment Banks
Actors from Paris
American Heavy Metal Singers (AHMS)
Business Law
16 <sup>th</sup> Century Italian People

\*\* = *maximum correlation, or smallest average shortest path between categories*

\* = *minimum correlation, or largest average shortest path between categories*

Table 3a. Top Categories -> Top Categories + Random Categories		
<i>Category</i>	<i>Ranking</i>	<i>Graph</i>
Living People	1. Business Law - 5.343 ** 2. Actors from Paris - 5.557 3. American Film Actors - 5.569 4. English Language Films - 5.600 5. American Films - 5.667 6. Investment Banks - 5.690 7. AHMS - 5.713 8. 16 <sup>th</sup> Century Italian People - 5.806 9. YOB Missing - 6.135 *	
Year of Birth Missing (Living People)	1. Business Law - 5.368 ** 2. American Film Actors - 5.591 3. Actors from Paris - 5.612 4. English Language Films - 5.622 5. American Films - 5.692 6. Investment Banks - 5.709 7. AHMS - 5.738 8. 16 <sup>th</sup> Century Italian People - 5.936 9. Living People - 6.202 *	

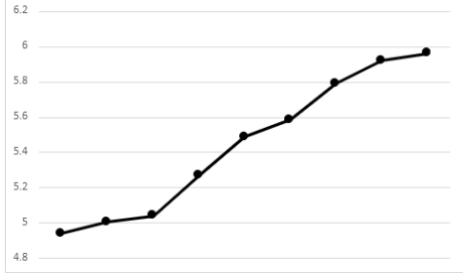
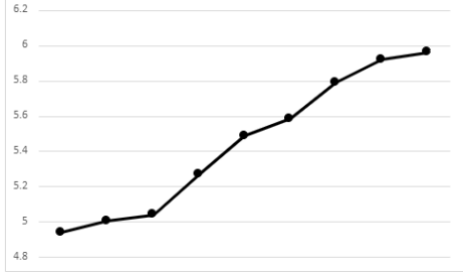
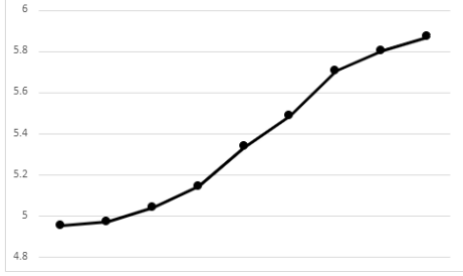
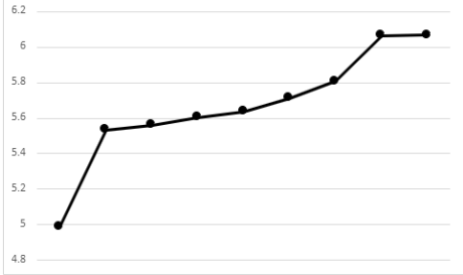
English Language Films	<ol style="list-style-type: none"> <li>1. American Film Actors - 4.942 **</li> <li>2. American Films - 5.005</li> <li>3. Actors from Paris - 5.041</li> <li>4. Business Law - 5.268</li> <li>5. AHMS - 5.489</li> <li>6. Investment Banks - 5.586</li> <li>7. 16<sup>th</sup> Century Italian People - 5.790</li> <li>8. YOB Missing - 5.921</li> <li>9. Living People - 5.962 *</li> </ol>	 <table border="1"> <caption>Data for English Language Films Graph</caption> <thead> <tr> <th>Rank</th> <th>Category</th> <th>Value</th> </tr> </thead> <tbody> <tr><td>1</td><td>American Film Actors</td><td>4.942</td></tr> <tr><td>2</td><td>American Films</td><td>5.005</td></tr> <tr><td>3</td><td>Actors from Paris</td><td>5.041</td></tr> <tr><td>4</td><td>Business Law</td><td>5.268</td></tr> <tr><td>5</td><td>AHMS</td><td>5.489</td></tr> <tr><td>6</td><td>Investment Banks</td><td>5.586</td></tr> <tr><td>7</td><td>16<sup>th</sup> Century Italian People</td><td>5.790</td></tr> <tr><td>8</td><td>YOB Missing</td><td>5.921</td></tr> <tr><td>9</td><td>Living People</td><td>5.962</td></tr> </tbody> </table>	Rank	Category	Value	1	American Film Actors	4.942	2	American Films	5.005	3	Actors from Paris	5.041	4	Business Law	5.268	5	AHMS	5.489	6	Investment Banks	5.586	7	16 <sup>th</sup> Century Italian People	5.790	8	YOB Missing	5.921	9	Living People	5.962
Rank	Category	Value																														
1	American Film Actors	4.942																														
2	American Films	5.005																														
3	Actors from Paris	5.041																														
4	Business Law	5.268																														
5	AHMS	5.489																														
6	Investment Banks	5.586																														
7	16 <sup>th</sup> Century Italian People	5.790																														
8	YOB Missing	5.921																														
9	Living People	5.962																														
American Films	<ol style="list-style-type: none"> <li>1. American Film Actors - 4.895 **</li> <li>2. English Language Films - 4.962</li> <li>3. Actors from Paris - 5.026</li> <li>4. Business Law - 5.267</li> <li>5. AHMS - 5.462</li> <li>6. Investment Banks - 5.628</li> <li>7. 16<sup>th</sup> Century Italian People - 5.801</li> <li>8. YOB Missing - 5.919</li> <li>9. Living People - 5.962 *</li> </ol>	 <table border="1"> <caption>Data for American Films Graph</caption> <thead> <tr> <th>Rank</th> <th>Category</th> <th>Value</th> </tr> </thead> <tbody> <tr><td>1</td><td>American Film Actors</td><td>4.895</td></tr> <tr><td>2</td><td>English Language Films</td><td>4.962</td></tr> <tr><td>3</td><td>Actors from Paris</td><td>5.026</td></tr> <tr><td>4</td><td>Business Law</td><td>5.267</td></tr> <tr><td>5</td><td>AHMS</td><td>5.462</td></tr> <tr><td>6</td><td>Investment Banks</td><td>5.628</td></tr> <tr><td>7</td><td>16<sup>th</sup> Century Italian People</td><td>5.801</td></tr> <tr><td>8</td><td>YOB Missing</td><td>5.919</td></tr> <tr><td>9</td><td>Living People</td><td>5.962</td></tr> </tbody> </table>	Rank	Category	Value	1	American Film Actors	4.895	2	English Language Films	4.962	3	Actors from Paris	5.026	4	Business Law	5.267	5	AHMS	5.462	6	Investment Banks	5.628	7	16 <sup>th</sup> Century Italian People	5.801	8	YOB Missing	5.919	9	Living People	5.962
Rank	Category	Value																														
1	American Film Actors	4.895																														
2	English Language Films	4.962																														
3	Actors from Paris	5.026																														
4	Business Law	5.267																														
5	AHMS	5.462																														
6	Investment Banks	5.628																														
7	16 <sup>th</sup> Century Italian People	5.801																														
8	YOB Missing	5.919																														
9	Living People	5.962																														
American Film Actors	<ol style="list-style-type: none"> <li>1. English Language Films - 4.953 **</li> <li>2. American Films - 4.971</li> <li>3. Actors from Paris - 5.042</li> <li>4. Business Law - 5.144</li> <li>5. AHMS - 5.336</li> <li>6. Investment Banks - 5.486</li> <li>7. 16<sup>th</sup> Century Italian People - 5.706</li> <li>8. YOB Missing - 5.803</li> <li>9. Living People - 5.869 *</li> </ol>	 <table border="1"> <caption>Data for American Film Actors Graph</caption> <thead> <tr> <th>Rank</th> <th>Category</th> <th>Value</th> </tr> </thead> <tbody> <tr><td>1</td><td>English Language Films</td><td>4.953</td></tr> <tr><td>2</td><td>American Films</td><td>4.971</td></tr> <tr><td>3</td><td>Actors from Paris</td><td>5.042</td></tr> <tr><td>4</td><td>Business Law</td><td>5.144</td></tr> <tr><td>5</td><td>AHMS</td><td>5.336</td></tr> <tr><td>6</td><td>Investment Banks</td><td>5.486</td></tr> <tr><td>7</td><td>16<sup>th</sup> Century Italian People</td><td>5.706</td></tr> <tr><td>8</td><td>YOB Missing</td><td>5.803</td></tr> <tr><td>9</td><td>Living People</td><td>5.869</td></tr> </tbody> </table>	Rank	Category	Value	1	English Language Films	4.953	2	American Films	4.971	3	Actors from Paris	5.042	4	Business Law	5.144	5	AHMS	5.336	6	Investment Banks	5.486	7	16 <sup>th</sup> Century Italian People	5.706	8	YOB Missing	5.803	9	Living People	5.869
Rank	Category	Value																														
1	English Language Films	4.953																														
2	American Films	4.971																														
3	Actors from Paris	5.042																														
4	Business Law	5.144																														
5	AHMS	5.336																														
6	Investment Banks	5.486																														
7	16 <sup>th</sup> Century Italian People	5.706																														
8	YOB Missing	5.803																														
9	Living People	5.869																														

Table 4. Random Categories -> Top Categories + Random Categories																																
Category	Ranking	Graph																														
Investment Banks	<ol style="list-style-type: none"> <li>1. Business Law - 4.990 **</li> <li>2. Actors from Paris - 5.535</li> <li>3. American Film Actors - 5.560</li> <li>4. English Language Films - 5.604</li> <li>5. American Films - 5.636</li> <li>6. AHMS - 5.713</li> <li>7. 16<sup>th</sup> Century Italian People - 5.806</li> <li>8. YOB Missing - 6.063</li> <li>9. Living People - 6.067 *</li> </ol>	 <table border="1"> <caption>Data for Investment Banks Graph</caption> <thead> <tr> <th>Rank</th> <th>Category</th> <th>Value</th> </tr> </thead> <tbody> <tr><td>1</td><td>Business Law</td><td>4.990</td></tr> <tr><td>2</td><td>Actors from Paris</td><td>5.535</td></tr> <tr><td>3</td><td>American Film Actors</td><td>5.560</td></tr> <tr><td>4</td><td>English Language Films</td><td>5.604</td></tr> <tr><td>5</td><td>American Films</td><td>5.636</td></tr> <tr><td>6</td><td>AHMS</td><td>5.713</td></tr> <tr><td>7</td><td>16<sup>th</sup> Century Italian People</td><td>5.806</td></tr> <tr><td>8</td><td>YOB Missing</td><td>6.063</td></tr> <tr><td>9</td><td>Living People</td><td>6.067</td></tr> </tbody> </table>	Rank	Category	Value	1	Business Law	4.990	2	Actors from Paris	5.535	3	American Film Actors	5.560	4	English Language Films	5.604	5	American Films	5.636	6	AHMS	5.713	7	16 <sup>th</sup> Century Italian People	5.806	8	YOB Missing	6.063	9	Living People	6.067
Rank	Category	Value																														
1	Business Law	4.990																														
2	Actors from Paris	5.535																														
3	American Film Actors	5.560																														
4	English Language Films	5.604																														
5	American Films	5.636																														
6	AHMS	5.713																														
7	16 <sup>th</sup> Century Italian People	5.806																														
8	YOB Missing	6.063																														
9	Living People	6.067																														

Actors from Paris	<div><div>1. American Film Actors - 5.052 **</div><div>2. English Language Films - 5.074</div><div>3. American Films - 5.132</div><div>4. Business Law - 5.161</div><div>5. Investment Banks - 5.340</div><div>6. AHMS - 5.524</div><div>7. 16<sup>th</sup> Century Italian People - 5.561</div><div>8. YOB Missing - 5.955</div><div>9. Living People - 5.958 *</div></div>	<table><caption>Actors from Paris Correlation Data</caption><thead><tr><th>Category</th><th>Correlation</th></tr></thead><tbody><tr><td>1. American Film Actors</td><td>5.052</td></tr><tr><td>2. English Language Films</td><td>5.074</td></tr><tr><td>3. American Films</td><td>5.132</td></tr><tr><td>4. Business Law</td><td>5.161</td></tr><tr><td>5. Investment Banks</td><td>5.340</td></tr><tr><td>6. AHMS</td><td>5.524</td></tr><tr><td>7. 16<sup>th</sup> Century Italian People</td><td>5.561</td></tr><tr><td>8. YOB Missing</td><td>5.955</td></tr><tr><td>9. Living People</td><td>5.958</td></tr></tbody></table>	Category	Correlation	1. American Film Actors	5.052	2. English Language Films	5.074	3. American Films	5.132	4. Business Law	5.161	5. Investment Banks	5.340	6. AHMS	5.524	7. 16 <sup>th</sup> Century Italian People	5.561	8. YOB Missing	5.955	9. Living People	5.958
Category	Correlation																					
1. American Film Actors	5.052																					
2. English Language Films	5.074																					
3. American Films	5.132																					
4. Business Law	5.161																					
5. Investment Banks	5.340																					
6. AHMS	5.524																					
7. 16 <sup>th</sup> Century Italian People	5.561																					
8. YOB Missing	5.955																					
9. Living People	5.958																					
American Heavy Metal Singers	<div><div>1. Business Law - 5.213 **</div><div>2. American Film Actors - 5.257</div><div>3. English Language Films - 5.323</div><div>4. American Films - 5.375</div><div>5. Actors from Paris - 5.389</div><div>6. Investment Banks - 5.549</div><div>7. 16<sup>th</sup> Century Italian People - 5.797</div><div>8. YOB Missing - 5.897</div><div>9. Living People - 5.950 *</div></div>	<table><caption>American Heavy Metal Singers Correlation Data</caption><thead><tr><th>Category</th><th>Correlation</th></tr></thead><tbody><tr><td>1. Business Law</td><td>5.213</td></tr><tr><td>2. American Film Actors</td><td>5.257</td></tr><tr><td>3. English Language Films</td><td>5.323</td></tr><tr><td>4. American Films</td><td>5.375</td></tr><tr><td>5. Actors from Paris</td><td>5.389</td></tr><tr><td>6. Investment Banks</td><td>5.549</td></tr><tr><td>7. 16<sup>th</sup> Century Italian People</td><td>5.797</td></tr><tr><td>8. YOB Missing</td><td>5.897</td></tr><tr><td>9. Living People</td><td>5.950</td></tr></tbody></table>	Category	Correlation	1. Business Law	5.213	2. American Film Actors	5.257	3. English Language Films	5.323	4. American Films	5.375	5. Actors from Paris	5.389	6. Investment Banks	5.549	7. 16 <sup>th</sup> Century Italian People	5.797	8. YOB Missing	5.897	9. Living People	5.950
Category	Correlation																					
1. Business Law	5.213																					
2. American Film Actors	5.257																					
3. English Language Films	5.323																					
4. American Films	5.375																					
5. Actors from Paris	5.389																					
6. Investment Banks	5.549																					
7. 16 <sup>th</sup> Century Italian People	5.797																					
8. YOB Missing	5.897																					
9. Living People	5.950																					
Business Law	<div><div>1. Investment Banks - 5.590 **</div><div>2. Actors from Paris - 5.725</div><div>3. American Film Actors - 5.771</div><div>4. English Language Films - 5.772</div><div>5. American Films - 5.829</div><div>6. 16<sup>th</sup> Century Italian People - 5.857</div><div>7. AHMS - 5.931</div><div>8. Living People - 6.228</div><div>9. YOB Missing - 6.231 *</div></div>	<table><caption>Business Law Correlation Data</caption><thead><tr><th>Category</th><th>Correlation</th></tr></thead><tbody><tr><td>1. Investment Banks</td><td>5.590</td></tr><tr><td>2. Actors from Paris</td><td>5.725</td></tr><tr><td>3. American Film Actors</td><td>5.771</td></tr><tr><td>4. English Language Films</td><td>5.772</td></tr><tr><td>5. American Films</td><td>5.829</td></tr><tr><td>6. 16<sup>th</sup> Century Italian People</td><td>5.857</td></tr><tr><td>7. AHMS</td><td>5.931</td></tr><tr><td>8. Living People</td><td>6.228</td></tr><tr><td>9. YOB Missing</td><td>6.231</td></tr></tbody></table>	Category	Correlation	1. Investment Banks	5.590	2. Actors from Paris	5.725	3. American Film Actors	5.771	4. English Language Films	5.772	5. American Films	5.829	6. 16 <sup>th</sup> Century Italian People	5.857	7. AHMS	5.931	8. Living People	6.228	9. YOB Missing	6.231
Category	Correlation																					
1. Investment Banks	5.590																					
2. Actors from Paris	5.725																					
3. American Film Actors	5.771																					
4. English Language Films	5.772																					
5. American Films	5.829																					
6. 16 <sup>th</sup> Century Italian People	5.857																					
7. AHMS	5.931																					
8. Living People	6.228																					
9. YOB Missing	6.231																					
16 <sup>th</sup> Century Italian People	<div><div>1. Business Law - 5.349 **</div><div>2. Actors from Paris - 5.408</div><div>3. English Language Films - 5.661</div><div>4. American Film Actors - 5.668</div><div>5. American Films - 5.735</div><div>6. Investment Banks - 5.744</div><div>7. AHMS - 5.879</div><div>8. Living People - 6.150</div><div>9. YOB Missing - 6.242 *</div></div>	<table><caption>16<sup>th</sup> Century Italian People Correlation Data</caption><thead><tr><th>Category</th><th>Correlation</th></tr></thead><tbody><tr><td>1. Business Law</td><td>5.349</td></tr><tr><td>2. Actors from Paris</td><td>5.408</td></tr><tr><td>3. English Language Films</td><td>5.661</td></tr><tr><td>4. American Film Actors</td><td>5.668</td></tr><tr><td>5. American Films</td><td>5.735</td></tr><tr><td>6. Investment Banks</td><td>5.744</td></tr><tr><td>7. AHMS</td><td>5.879</td></tr><tr><td>8. Living People</td><td>6.150</td></tr><tr><td>9. YOB Missing</td><td>6.242</td></tr></tbody></table>	Category	Correlation	1. Business Law	5.349	2. Actors from Paris	5.408	3. English Language Films	5.661	4. American Film Actors	5.668	5. American Films	5.735	6. Investment Banks	5.744	7. AHMS	5.879	8. Living People	6.150	9. YOB Missing	6.242
Category	Correlation																					
1. Business Law	5.349																					
2. Actors from Paris	5.408																					
3. English Language Films	5.661																					
4. American Film Actors	5.668																					
5. American Films	5.735																					
6. Investment Banks	5.744																					
7. AHMS	5.879																					
8. Living People	6.150																					
9. YOB Missing	6.242																					

The results of these experiments yielded a few trends and patterns to be discussed and investigated further. To begin, it should be noted that not all results of each experiment are shown due to size constraints of this paper, but the remaining result tables will be included within the supplementary materials for view and each result table will be named and discussed.

The first experiment tested was comparing the Top Categories against the other Top Categories (Table 3b). Within the top categories, three of the five were categories that were related to the film industry. The results and correlation rankings of those three categories were nearly identical (best visualized by the graph structure), namely the top two most correlated categories



were the other film-related categories. After the top two results, which are very close in distance, there is a very large jump in distance, nearly an entire extra step in the average path length, to the bottom two results, which are also very close in distance to each other. These results exhibit that the film-related categories are much more related to each other than the other top categories. Although these results would make sense at first glance, it is still surprising that the average shortest path length of film-related categories is approximately 5, meaning it takes 5 steps to reach one article to another. This result does not make as much sense given that there is a seemingly close semantic relationship between the film-related categories. The original hypothesis for this experiment predicted a much closer correlation between these categories than what the results show. The results of the other top categories are also worthy of note given that their results were also nearly identical too.

The next experiment to discuss is the results of finding correlation between the Top Categories and the set of only the Random Categories (Table 3c). Similarly, to results of Table 3b, the three film-related categories produced near identical results, with the exact same rank structure and even very little difference in the correlations to each of the random categories. While the other top categories also exhibited nearly identical results to each other. From the perspective of individual who knows the English language, the film-related categories have high semantic relatedness. Someone would look at those categories and conclude that they are similar to each other. Given how the results of the average shortest paths to the randomly selected categories materialized, on a network-level, the film-related categories also exhibit a high level of relatedness. Although the underlying intricacies of this conclusion are not completely clear, these categories show similar paths in their hyperlink structure, possibly showing there were similar, if not the same, paths taken by articles within these categories. These conclusions also apply to the top categories related to living people on Wikipedia and the previous experiment discussed with Table 3b. Based on these results, one can conclude that if there exists a highly similar or identical rank structure after performing this methodology, then those categories have underlying, high relatedness. Although discussed individually, the results from the previous experiments (Tables 3b and 3c) can be seen culminated in Table 3a, which previews the same results.

The final experimental results were to examine are the correlations between the Random Categories and the set of the Top Categories & the other Random Categories (Table 4). Within the set of randomly selected categories, there is another category that has to do with the film industry. The category ('Actors from Paris') shows similar results to the previously mentioned film-related categories further providing evidence to the conclusion that semantically related categories will exhibit comparable results in average path length and rank structure. Although there are some interesting outlying correlations to consider, there do not seem to be any obvious patterns that would indicate underlying trends or other conclusions to be made. However, working on a more diverse and larger subset of categories could yield results that have not been displayed with this particular random selection.

## Discussion:

The main goal of this project and experiments was to look more closely at relatedness in Wikipedia categories using the average shortest path of hyperlinks between those categories as a metric to quantify potential relationships that may exist. To the best of my available knowledge, this work exhibits a novel approach to finding relatedness, namely semantic relatedness, using Wikipedia and its category and article link network as the medium for study. The results of this project provide evidence that related categories will exhibit similar ranking structure and average shortest path length when being compared against the same set of other categories. These results also indicate validity in the proposed methodology and the potential to expand the work for further testing and evidence collection. Currently, this conclusion is only in the beginning stages of being formed and tested, leaving many more experiments to take place in the future.

The next step in this research is to test the proposed conclusion more thoroughly by choosing multiple sets of semantically related categories and compare them against a set of semantically dissimilar and/or random categories to see if results will produce alike ranking structure. Further experiments would test the opposite case of semantically unrelated categories against one another testing the presence of dissimilar rank structures. These procedures will be conducted against the same dataset used in this study, but a way to further reinforce this idea would be to try and compare against other popular datasets that are used for the problem of finding semantic relatedness. Hopefully these experiments can also be conducted with a larger subset of the data than previously worked in this project to better substantiate whether the proposed pattern exists in related categories.

Although there is some promise in this work, there are also some limitations that should be discussed too. The biggest limitation in conducting this research is time. Given the size of the network and the number of calculations that need to occur to obtain results, it takes long periods of time to complete this task on relatively high-performance hardware. However, the sampling process described in the methodology provides an opportunity to reduce processing time while outputting conclusive results. Another limitation is that this proposed approach may be only applicable to the SNAP dataset used in this study. Although it yields promise when looking at Wikipedia's link structure, it may not have the same applications to other common datasets used for computing semantic relatedness and other relatedness tasks. The scope of this work may prove limited, but there are potential benefits to further studying underlying trends and patterns in relatedness using hyperlink structure of Wikipedia and average shortest path length.

## Acknowledgements:

I greatly appreciate the help of Dr. Zoran Obradovic and Dr. Jumanah Alshehri for giving guidance and direction to my project. Their assistance allowed me to improve my work and methods, producing more interesting results, which led to better insights and discussion into Wikipedia's network structure.

## References:

- [1] Leskovec, J. 2011. Wikipedia network of top categories. Stanford Network Analysis Project (SNAP). <http://snap.stanford.edu/data/wiki-topcats.html>
- [2] Singer, P., Niebler, T., Strohmaier, M., & Hotho, A. 2013. Computing Semantic Relatedness from Human Navigational Paths: A Case Study on Wikipedia. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 9(4), 41-70. <http://doi.org/10.4018/ijswis.2013100103>
- [3] Milne, D. 2007. Computing semantic relatedness using wikipedia link structure. *Proceedings of the new zealand computer science research student conference*. Vol. 7. No. 8.
- [4] Agirre, E., Barrena, A., Soroa, A. Studying the Wikipedia Hyperlink Graph for Relatedness and Disambiguation. 2015. <https://doi.org/10.48550/arXiv.1503.01655>
- [5] Chernov, S., Iofciu, T., Nejdl, W., Zhou, X. 2006. Extracting Semantics between Wikipedia Categories. *CEUR Workshop Proceedings*. 206.
- [6] Eric Yeh, Daniel Ramage, Christopher D. Manning, Eneko Agirre, and Aitor Soroa. 2009. [WikiWalk: Random walks on Wikipedia for Semantic Relatedness](#). In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4)*, pages 41–49, Suntec, Singapore. Association for Computational Linguistics.
- [7] Zesch, T., & Gurevych, I. 2007. Analysis of the Wikipedia category graph for NLP applications. In *Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing* (pp. 1-8).