**CSOC20010**

*Introduction to Computational Social Science II*

UCD School of Sociology

Spring, 2020-2021

Module Coordinator: Taha Yasseri

***Final Essay***

**Social Contagion: Mapping the Spread of Covid-19 through Wikipedia**

**Samuel Dornan & Ines Fernandez**

**1.** **Introduction and background**

The aim of this project was to use Wikipedia to measure how Covid-19 has affected different facets of the world. In order to do this, we would take the list of Wikipedia articles that link to the Covid-19 Wikipedia page and see which of the 13 main topic classifications they belong to, with these main topic classifications being reference, culture, geography, health, history, human activities, mathematics, nature, people, philosophy, religion, society, and technology. By counting how many articles in each of these categories link to Covid-19, we would be able to estimate how much each of them were impacted by the pandemic. There is no preexisting or official algorithm for categorising an article into one of these top 13 categories, so the project would involve coming up with our own methodology in order to do this. We decided to use an algorithm from the NetworkX Python library that finds the shortest path between two Wikipedia pages, and classify articles based on which category they had the shortest path to. Ultimately, the research question was: How much were each of Wikipedia’s 13 main topic classifications impacted by the Covid-19 pandemic? This project builds on work such as that by Turki et al. (2021).

**2.** **Data and Methods**

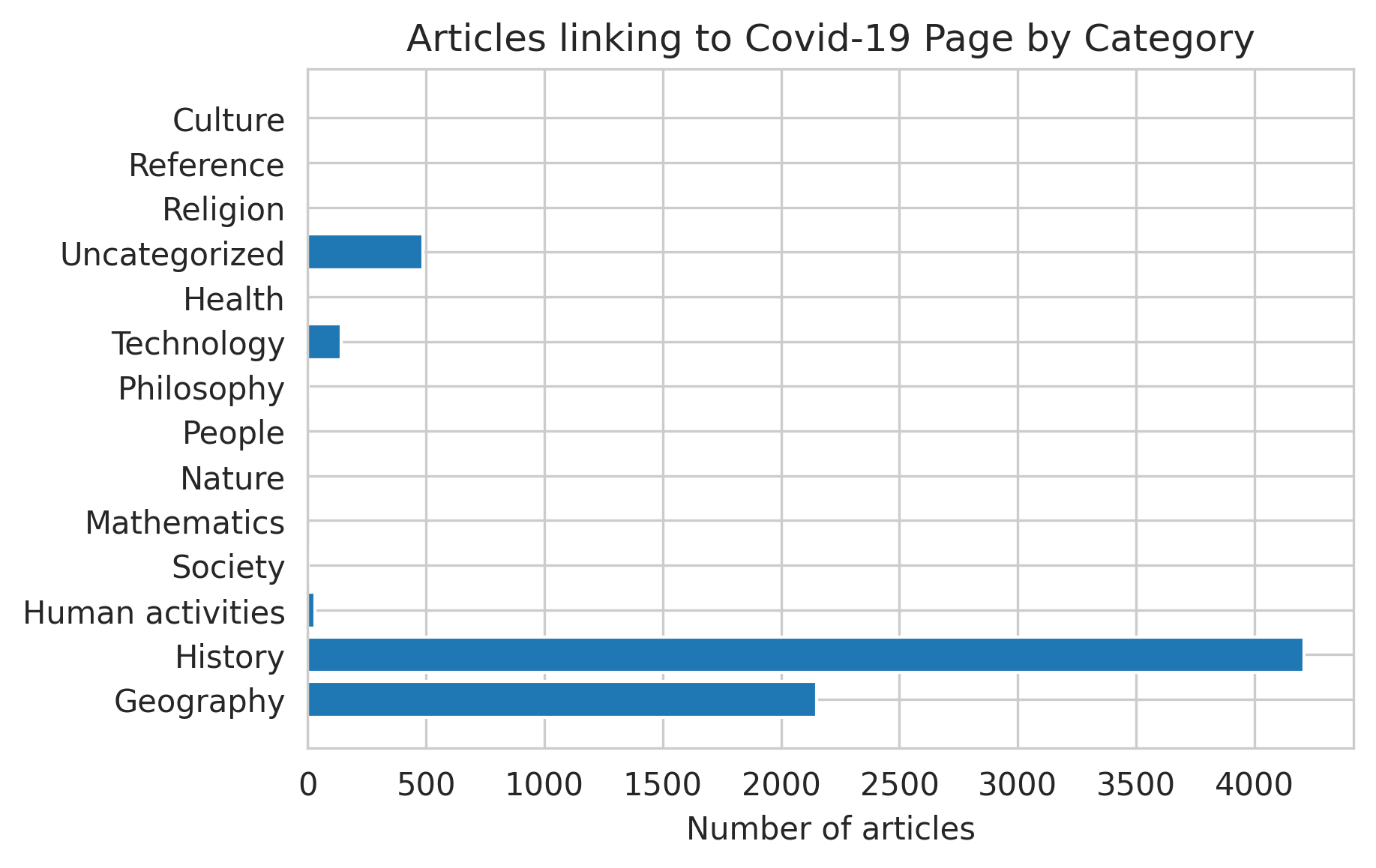
To determine which of the 13 main topic classifications were most heavily affected by Covid-19, we began by collecting the first and second neighbours of the article “Covid-19” on Wikipedia using the SQL API service known as Quarry. However, our attempts to replicate this procedure to collect the relevant topic classifications using Quarry failed, which led us to use the Wikipedia database dumps of all category links across the English language Wikipedia from the beginning of April 2022.[[1]](#footnote-1) After downloading and extracting the category links database table from the database dump, we filtered the table to only show pages linking to subcategory pages, and performed an inner join on the page IDs of the neighbouring articles to merge the neighbouring articles with their corresponding category links. These linked articles were then post-processed to remove as many maintenance categories as possible, before being exported to NetworkX, where the shortest path from the neighbour articles to each of the main topic classifications was found. The main topic classification with the shortest path between itself and the relevant neighbouring article was determined to be the most relevant classification (Haddadi et al., 2008).

**3.** **Results**

The results of this process differed from what was expected. Of the articles in our dataset, 2151 were classified as geography, 4208 as history, 31 as human activities, 2 as society, 0 as mathematics, 0 as nature, 0 as people, 2 as philosophy, 141 as technology, and 1 as health. This is remarkably different from what one would predict—one would, for instance, expect there to be more than one Wikipedia article primarily about health that links to Covid-19. The Wikipedia pages of famous individuals who died due to Covid-19 would also be expected to both link to Covid-19 and to be classified as “people”, but none of them did. The pandemic also had well-documented effects on religion and culture that we would expect to see here (Adegboyega et. al, 2021; Kantor and Kubiczek, 2021). This suggests that our methodology for classifying articles may have suffered from some important shortcomings.

A further 486 articles were left uncategorized, as they were not found in our second dataset. The pages for religion, reference, and culture were not found in this dataset either, and thus all had 0 articles categorised as them. This means that there were problems in our data that affected the quality of our results.

These results can be visualised in the following graph:



**4.** **Discussion**

There are many reasons why our results may have been different from what was expected, pertaining both to the methodology and to the data. First, the classification algorithm based on the shortest path has some important flaws. It may be the case that most articles include a section about the subject matter’s history or its geographical origins, causing them to be incorrectly classified as history or geography articles when that is not their actual primary focus. Moreover, there is an overwhelming amount of maintenance categories that connect articles that are not related in a meaningful way, such as “Articles with a short description” or “Articles using American English from February 2015”. We often found that these pages were being used as loopholes by the shortest path algorithm in order to find the shortest path between two pages, and led to the incorrect classification of many articles. We eliminated thousands of these maintenance categories during the data cleaning process, but given the sheer number of them, it was impossible to eliminate all of them. Thus, it is possible that the algorithm continued exploiting remaining maintenance categories and this led to misclassification of articles.

Moreover, the algorithm may have been unable to categorise some articles due to our use of both the live database and Wikipedia database dumps. More specifically, Wikipedia database dumps appear to lag slightly behind the main Quarry API service which queries against the live database, which may have led to some articles appearing in our dataset of neighbouring articles, but not appearing in our dataset of category links. Alternatively, our use of an inner join to merge the category links with their corresponding page titles may have led to some articles being dropped from the dataset of category links but not the dataset of neighbouring articles, leaving them unable to be categorised.

Our use of the shortest path algorithm required us to manually select the nodes which we thought best represented each of the 13 main topic categorisations. While every effort was made to select the highest-level and most relevant node for each main topic classification, our data was such that not all of these nodes were of the highest level. It is therefore possible that there was significant variance in the degrees of each of the main topic nodes in our dataset, leading to some articles being miscategorized.

Lastly, it is possible that despite all these problems, most of the articles linking to Covid-19 really were about history and geography, and the pandemic had greater implications for history and geography than it did for the other main topic classifications. While we cannot rule out that this may be true, we also cannot confirm that it is without further study addressing the aforementioned shortcomings and seeing whether the results still hold.

**5.** **Potential Extensions**

As outlined above, there are a variety of ways this study could be improved to improve the quality of results. Another important extension would be to look at the distribution of Wikipedia articles in general across the 13 main topic classifications, and then compare these to the distribution of just the Wikipedia articles that link to Covid-19 across the 13 main topic classifications. This would address the underlying assumption that articles are distributed approximately uniformly across the top 13 categories, which could easily not be true. By seeing what proportion of articles belonging to the top 13 categories link to Covid-19, instead of just the absolute number, we would get a better idea of how much each of these was affected by Covid-19.

In addition to improving the quality of results, we could expand on this research to learn even more information: we could classify the second neighbours linking to Covid-19 and use a clustering algorithm to find additional insights in the data. Moreover, if improved, the same methodologies outlined in this paper could be applied to learn how other important historical events, individuals, or cultural phenomena impacted each of the top 13 Wikipedia categories, or the 42 subcategories that are one level down.

**6.** **Conclusion and Reflection**

The results of this study suggest that the Covid-19 pandemic primarily impacted aspects of history and geography, affected technology and human activities significantly less, and had a very small impact on culture, health, mathematics, nature, people, philosophy, religion, society, and reference. However, these results are very counterintuitive and do not align with common perceptions about the effects of the pandemic. There were several shortcomings in the data and methodologies used, and these may explain the results found. Further study is needed to address this study’s shortcomings before any conclusions can be reached regarding their factualness.

**References**

Adegboyega, A., Boddie, S., Dorvie, H., Bolaji, B., Adedoyin, C., & Moore, S. E. (2021). Social distance impact on church gatherings: Socio-behavioral implications. Journal of Human Behavior in the Social Environment, 31(1-4), 221-234.

Haddadi, H., Rio, M., Iannaccone, G., Moore, A., & Mortier, R. (2008). Network topologies: inference, modeling, and generation. *IEEE Communications Surveys & Tutorials*, *10*(2), 48-69.

Kantor, A., & Kubiczek, J. (2021). Polish culture in the face of the COVID-19 pandemic crisis. Journal of Risk and Financial Management, 14(4), 181.

Turki, H., Hadj Taieb, M. A., Shafee, T., Lubiana, T., Jemielniak, D., Aouicha, M. B., ... & Mietchen, D. (2021). Representing COVID-19 information in collaborative knowledge graphs: the case of Wikidata. Semantic Web, (Preprint), 1-32.

1. Database dumps available at https://dumps.wikimedia.org/enwiki/latest/ [↑](#footnote-ref-1)