Gabrielle Avila

Yiheng Ye

Michael Lam

Wikipedia's Response to the COVID-19 Pandemic

Abstract

Through collaborative efforts online, Wikipedia has always been at the forefront of providing information to the public on almost any topic, including a pandemic. Covid-19 has been one of the most relevant topics of 2020 and still remains so as of right now, therefore gathering as much information as possible is essential for the world to combat such a virus. Many official health sources online provide such knowledge with the resources that they have, but false or outdated information can spread quickly. In this article, we perform EDA and LDA on different Wikipedia articles related to coronavirus and compare the results to the word clouds of traditional sources to explore how Wikipedia can provide reliable and updated details and data about Covid-19.

Introduction

Given the current pandemic, up to date information is essential to keeping people safe and informed. Traditional online sources such as the CDC, World Health Organization and John Hopkins, provide up to date and reliable information on COVID numbers and information. However online platforms such as Wikipedia, also provide a comprehensive and real time approach to analyzing a pandemic. There can be complications with online platforms providing false information and reporting conspiracy theories, however we believe these discrepancies are fixed by credible editors preserving Wikipedia's facts. By June 2020, covid articles on Wikipedia had over 400 million pageviews.[1] We are going to investigate how these articles are created and edited, and how an online community can be used to monitor a pandemic. Online communities present real time, unique,

[1] https://wikimediafoundation.org/covid19/data

comprehensive information and a new understanding of the covid pandemic by studying page views, edits and comparing information to reliable sources.

We are going to justify that an online community can be used to provide reliable information on the safety and health of everyone. The problem is that inaccurate information being spread about a global pandemic can be costly and detrimental. We plan on using several methods to quantify the reliability of wikipedia information.

The data is bound to the past year, since the covid pandemic first began back in November of 2019. Our scope is limited to articles about covid from 2019 to present. The bar graph to the left shows the counts of edits made by year on the article titled "COVID-19 Pandemic", it appears all of the edits occurred in 2020. We will also use other aggregated data sources found in the COVID-19 Data Repository[2] by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University to make comparisons between Wikipedia data and traditional data sources.

Since the pandemic is global we are going to analyze the resources presented in other countries, and compare those results/information to more reliable sources. The map to the left is provided by Wikipedia, with the confirmed cases displayed throughout the entire continent. This type of information is available for most locations in the world.

We are going to look at the editors who are editing covid information the most and see the disparities across different regions on this information. The pie chart to the right shows the top 10 editors for the article titled "COVID-19 Pandemic". We are going to explore a deeper analysis of this article on wikipedia.

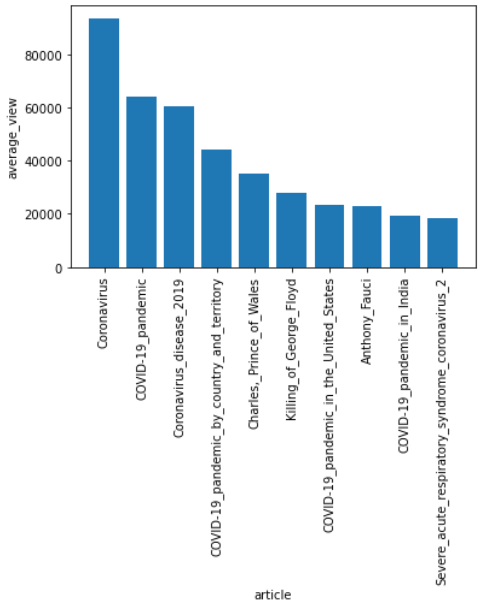
EDA

We explored a single page on Wikipedia to gather general information from the COVID pandemic of deaths, recovery rate, by country. We explored that Mexico had a much larger death rate with Covid than any other country recorded. This could be due to the health care system in Mexico. This result was expected.

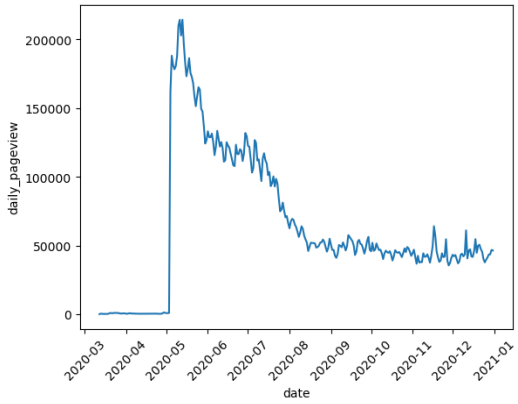
[2] https://github.com/CSSEGISandData/COVID-19

We also gathered the text and citation data from thirteen articles that contained “Covid” in the title on wikipedia. The articles contained similar citations/references to popular trusted articles outside the online community.

We also gathered 1,000 articles on wikipedia and their pageview counts. We analyzed the daily pageviews for popular articles and compared them to significant dates within the past year to see if there were any trends.



[*Figure 1*]: Top 10 COVID-19 related articles in Wikipedia with the most average pageview

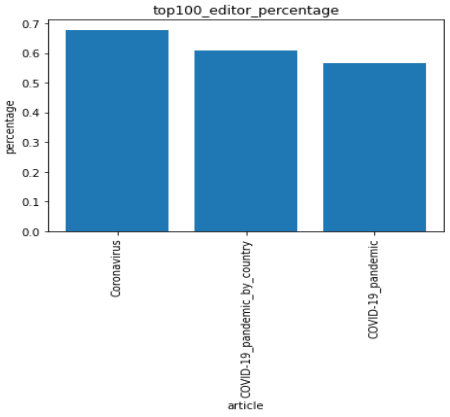


[*Figure 2*]: Daily pageview for article COVID-19 pandemic in 2020

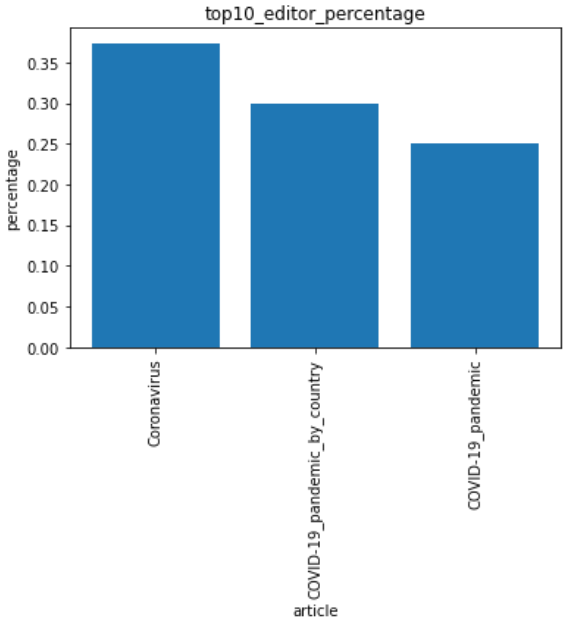
We can see from our data, [Figure 1], that the most popular articles related are comprehensive covid-19 pages as well as some pages for important figures through the whole period. Looking at those tops and the plot of COVID-19 pandemic daily pageview [Figure 2], we can see that the pageview popularity is related with the COVID-19 pandemic in the United States, which makes sense as most of English readers are from America. However, India users seem to also contribute to the pageview here as the pandemic issue in India is also concerned according to our data. This guide us about our future investigation on the information Wikipedia is providing.

Besides the separate Wikipedia page view data, we also count the total pageview for those 1000 popular articles in the year 2020. 387,176,891 pageviews combined for those articles, which is smaller than our expectation. Since John Hopkins University's website exceeds 1 billion visits in January 2021 to their Coronavirus page. That is to say, the whole Wikipedia COVID-19 project is not as popular as the John Hopkins University's website.

Furthermore, we also explored editing history data on important COVID-19 related pages in 2020, and according to our first investigation, it looks like a few editors contribute to the majority of the work in making those articles.



[*Figure 3*]: Contribution for top100 editors in some important COVID-19 articles



[*Figure 4*]: Contribution for top10 editors in some important COVID-19 articles

We can observe from [Figure 4] that even if we only choose top10 editors, they still make a lot of contribution to the formations of those articles.

Therefore, we want to make a further look into those editing data and making investigation on who are responsible for those editing contents

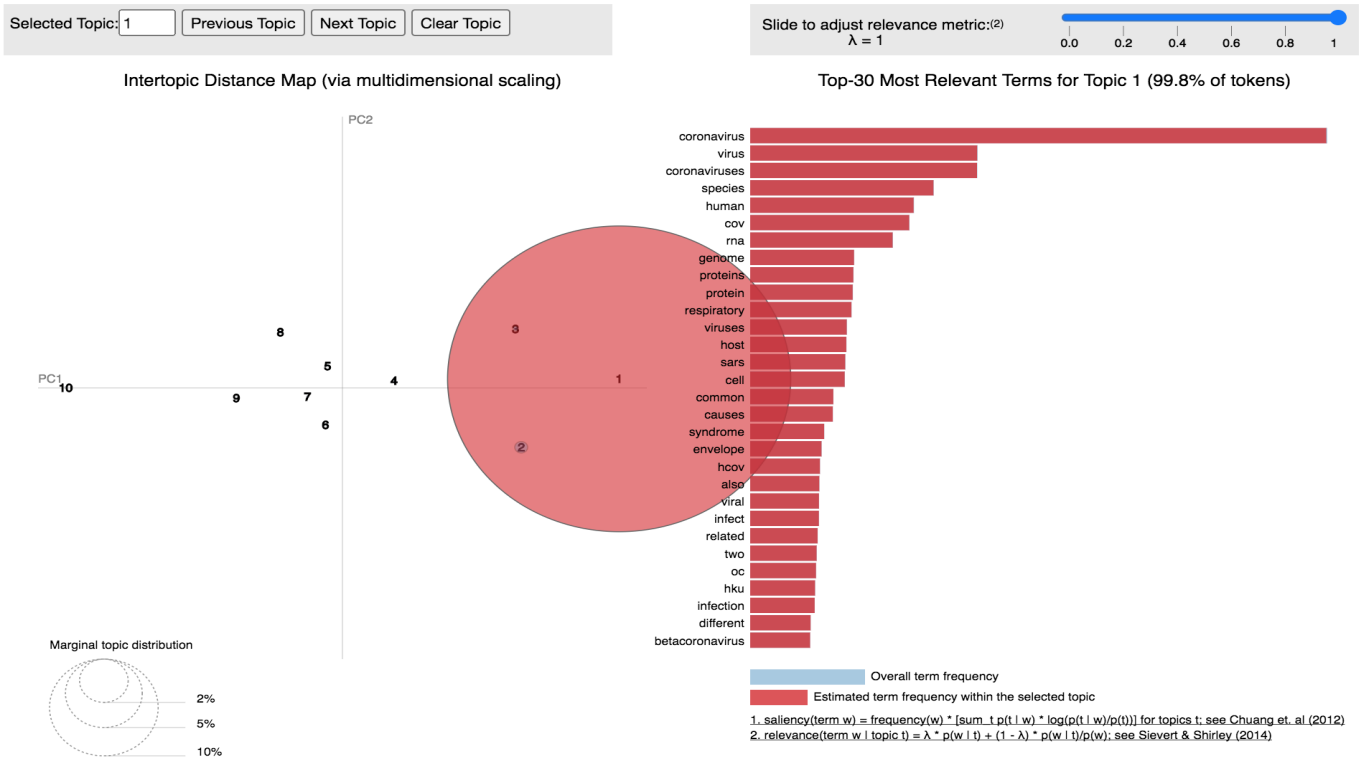
Methods

We are choosing methods that are going to allow us to see if Wikipedia data is similar to other reliable sources. We are going to analyze

their text and editing data, as well as do an analysis into specific popular covid articles, comparing them to more commonly known and trusted researchers.

We are going to incorporate deep data analysis with topic models to see what Wikipedia is focusing on and how they differ from some other websites like JHU (John Hopkins University) or WHO. That is to say, we are going to analyze the editing history of Wikipedia and the contents those websites have. We are using word clouds to show the main contents for three different websites (Wikipedia Coronavirus page, JHU, and WHO), and use LDA (latent dirichlet allocation) model to analyze the topic talked by Wikipedia COVID-19 pandemic page. However, to provide more insight on the Wikipedia COVID-19 page, and find out how they provide intended contents, we need to make more investigation on the editing history. We will study the composition of both editors and revisions and use LDA models on the revision comment to see how those editors collaboratively make contributions to the COVID-19 articles.

Results/Analysis

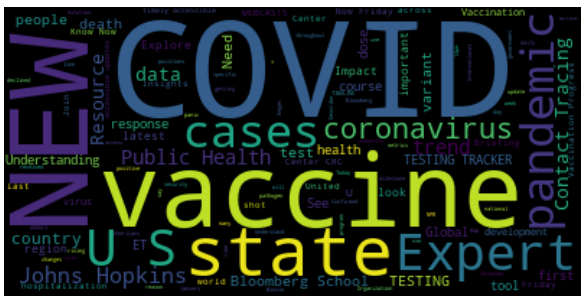


Initially, we decided to analyze Wikipedia's "Coronavirus" article page by utilizing topic modeling. Specifically, we implemented the Latent Dirichlet Allocation method, or LDA, of topic modeling to view the most salient terms for given topics within the Coronavirus article (The above one, treated as Figure 5).

However, after clustering the frequency of terms under certain topics, we noticed that topic 1 contained 99.8% of tokens, which means that 99.8% of all terms within the Coronavirus article were found in topic 1. We find that this

article is constantly talking about coronavirus and the properties of this virus.

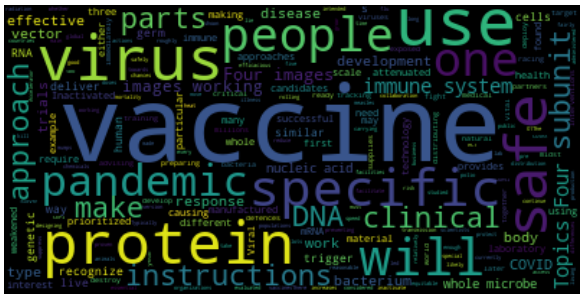
In order to compare the content provided by different websites, we choose to use a simple way to visualize the content of different websites: word cloud. The results are shown below.



[*Figure 6*]: Word Cloud for the John Hopkins University website.

Taken from the main page on the coronavirus center. This figure shows how prevalent they are spreading the vaccine. You can also see the word John Hopkins used frequently.

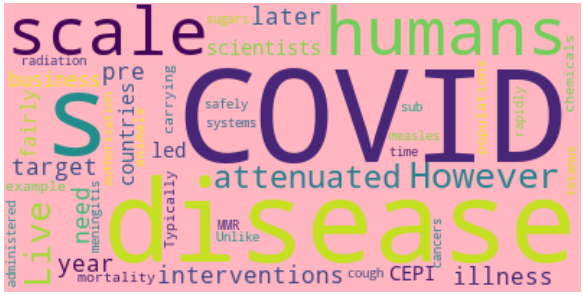
[*Figure 7*]: Word Cloud for the John Hopkins University website using the set of its own words as stopwords.



[*Figure 8*]: Word Cloud for the WHO website, the most popular word is vaccine.

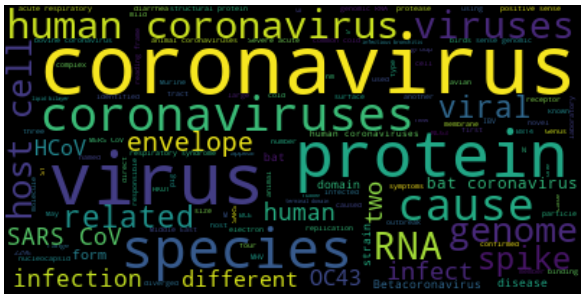
Considering most posts on social media site vaccine information with the WHO, this is expected. We don't see the word covid at all or

coronavirus, we do see the word COVID appear as a stop word in this case.



[*Figure 9*]: Word Cloud for the WHO website using the set of its own words as stopwords.

According to [*Figure 9*], Covid and disease appear in the stopwords, whereas they didn't in the other two.

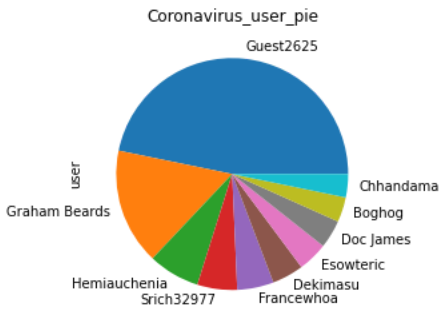


[*Figure 10*]: Word Cloud for the Wikipedia Coronavirus article

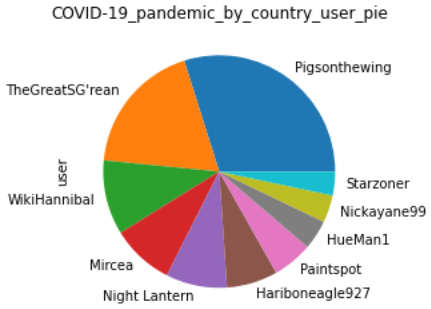
[*Figure 11*]: Word Cloud for the Wikipedia Coronavirus article using the set of its own words as stopwords.

It appears there is more information on [*Figure 11*] where the virus actually is, using a lot of terminology, rather than discussing the vaccine. Wikipedia has a separate page for numbers that are updated on deaths/recovery rates across the globe.

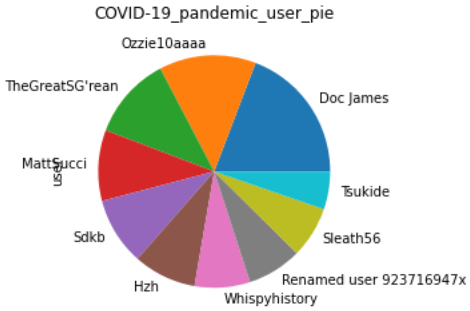
We also investigated the editing history and editor composition in 2020 for three important articles we choose, which are: Coronavirus, COVID-19 pandemic and COVID-19 pandemic by country and territory. The reasons why we choose those three articles are that they are both representative while providing enough information. For example, the COVID-19 pandemic page has 23500 revision history alone in the past 2020. Additionally, we are using comments instead of contents when conducting the LDA model as contents are usually redundant/providing meaningless information while the comments can accurately tell us what those editors are doing.



[*Figure 12*]: Contribution composition in the edits made by them for top10 editors in the article "Coronavirus"



[*Figure 13*]: Contribution composition in the edits made by them for top10 editors in article "COVID-19 pandemic by country and territory"



[*Figure 14*]: Contribution composition in the edits made by them for top10 editors in article "COVID-19 pandemic"

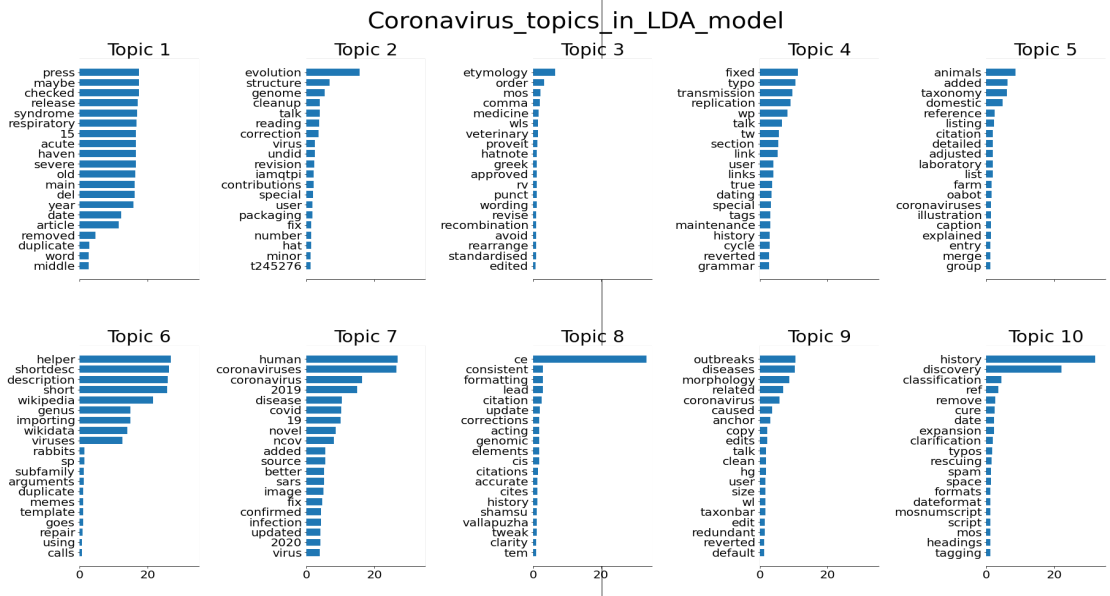
As we have discussed in the EDA part, the top 10/100 editors contribute a lot in making those COVID-19 related articles. However, if we look closer, we can find that not only few editors contribute a lot in editing them, but some of the articles have a "main contributor" who makes significant contributions to the editing work there. "Guest2625" edits nearly half of the editor work in the editing work made by top10 editors in the article "Coronavirus" , while the editor "Pigsonthewing" works over a quarter of the edit works in the top10 for the article "COVID-19 pandemic by country and territory". But the COVID-19 pandemic article does not have a "main contributor" and it has much more revision history in 2020. The "Coronavirus" page has only 1500 revisions, and the "COVID-19 pandemic by country and territory" page has 5000 revisions. But, the "COVID-19 pandemic" page has 23500 revisions. It looks like the more frequently edited an article is, the more equal contributions every editor has, which is good in editing work as they are applying

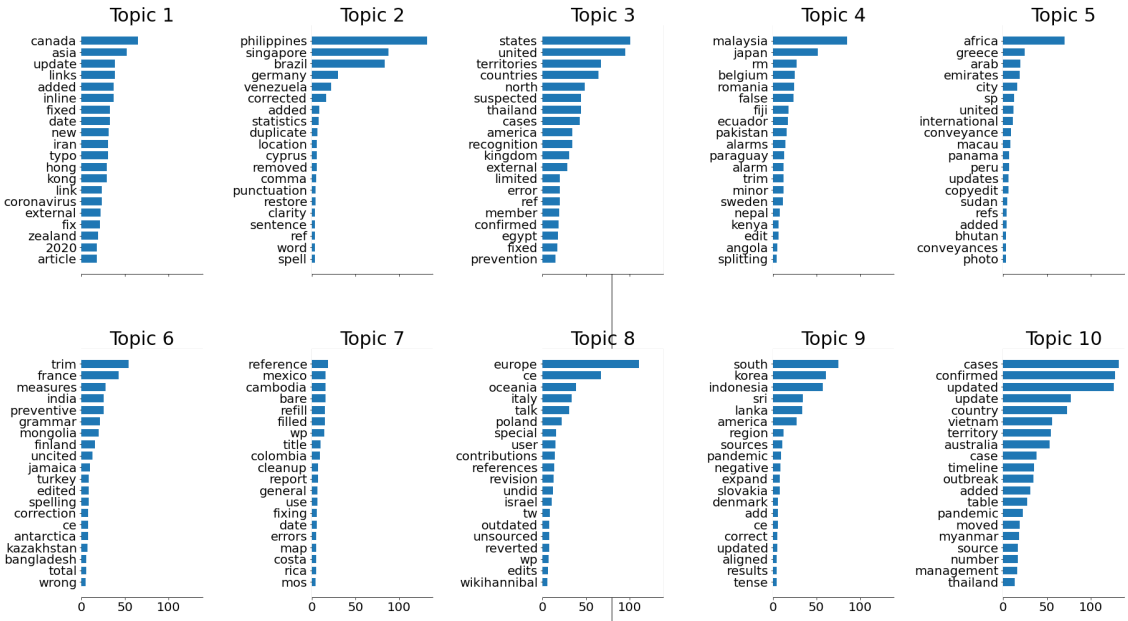
more crowd resourcing. In other words, Wikipedia actually does not apply crowd resourcing well on its COVID-19 project except some big articles like “COVID-19 pandemic”.

Now, to further investigate what editors When they were making edits, we applied the LDA model on those edit comments with 10 topics, and the tables below (Treated as Figure 15,16,17) are our results.

Looking at those editing topic models, we can find several interesting discoveries.

“Coronavirus” article seems to have a lot of





revisions which are not about the contents but the arrangement of the article page, as its topic 4,8,10 are basically talking about some professional terms in editing wikipedia. However, if we look at the “COVID-19 pandemic by country and territory”, we can find that most of their revision works involve certain countries or areas and some revision terms do not show up as important as those countries. This tells us that this article, with more revision records, is focusing more on making quality contents instead of doing some arrangement stuff.

The most meaningful part of the LDA analysis on edit history is the result we get on the “COVID-19 pandemic” article. Due to its significantly large amount of revision records, it has great diversity in topics. They are not only talking about events happening in some major countries/areas in this pandemic, but the editors also care about format and other arrangement issues in this article. But the most important part I think is the Topic 4 this article has, as the editors are working on dealing with “misinformation” and “disinformation”. This is a great difference between a crowd-resourcing platform and a professional website as the former one has a lot more editors who care about fighting with misinformation. However, as we have seen in the whole LDA result, this is not always the case for Wikipedia articles. It seems that the more revisions an article has the more concern on content quality it will get, since the number of edit records is ranked as “Coronavirus”(1500) <”COVID-19 pandemic by

country and territory”(5000)<”COVID-19 pandemic”(23500).

Conclusion

Before we start our research, we assume that Wikipedia will be performing better than traditional websites as it can provide diverse information with crowd-sourcing to keep misinformation from happening. However, as we explored more and more into the actual contents of those websites, we realized that it is difficult to conclude whether Wikipedia is performing better or worse. First of all, Wikipedia does not have a similar level of popularity compared to the traditional websites like the JHU one as the top 1000 popular articles in the Wikipedia coronavirus project. They also have a less overall total pageview than the JHU one. Secondly, the contents provided by the website, according to our word clouds and topic models, are similar however the WHO focuses more on vaccine information and distribution. The last point is, unfortunately, although some of the Wikipedia articles are using crowd-resourcing to find against misinformation in COVID-19 topic, there are still articles who did not utilize this advantage. According to our research on the three major coronavirus pages, only the editors for “COVID-19 pandemic” are collaborating with each other to provide in-time accurate information, and the rest two pages are depending on top active editors or even one significant contributor. Considering that the top 100 editors for those 3 articles are all making up a large portion of their edit works, we cannot say that Wikipedia utilizes crowd-resourcing very well in this coronavirus project. Especially since all of the information found on Wikipedia can be cited from similar credible sources.

Therefore, our final conclusion is that Wikipedia does have real time, reliable covid information, but this doesn’t mean we can generalize it for all information. We cannot conclude that Wikipedia outperforms traditional websites as it does not make use of its advantage well on reporting COVID-19 related information. We can say that there is less advertisement for vaccine distribution compared to the World Health Organization, depending on what kind of information you want to receive about the pandemic you may resort to a different source.

Future studies/projects can look into more Wikipedia articles/sources aside from the ones that this article looked into to provide more in-depth information about how Wikipedia performs in comparison to traditional sites. Also, instead of using just LDA, future projects can utilize the other methods found within topic modeling to see what other results can be gathered. Lastly, looking into past pandemics other than covid-19 could differentiate the difference between Wikipedia and other sources.

Works Cited

https://en.wikipedia.org/wiki/Template:COVID-19_pandemic_data

https://www.mitpressjournals.org/doi/pdf/10.1162/qss_a_00080?fbclid=IwAR0seTLDX5_lFS4wS9y3mIRFqMbIkWfYWtvxoznG6FKZO1RRSHEFWkwxvcQ

https://github.com/CSSEGISandData/COVID-19

Project Proposal