Task 1.1: To understand the depth of this unevenness, we can examine the Oxford Internet Institute's visualizations of global Wikipedia coverage. Click on each of the four maps showing global Wikipedia coverage measured against different properties and read the accompanying text.

Task 1.2: In 3-5 sentences, describe 3 aspects of these findings that you found interesting or surprising.

While it was expected, I found it interesting how the United States was so dominant in Wikipedia (leader in articles written, articles per person, monthly edits, and contribution by locals). The strength of the U.S. in each category means that a large part of Wikipedia knowledge, even Internet knowledge, is shaped by the United States. This could mean that information not pertaining to the United States could be biased, untrue, or inaccurate.

The underrepresentation of information from local Asian, specifically Chinese, writers was very surprising to me. The number of articles written about China was rather high; however, in terms of locally produced information, it remained very low as the graph of China was relatively "pale/white". This suggests that the current information pertaining to China on Wikipedia and the Internet could be false.

The general lack of information on Africa was expected but the degree to which there was a lack of information was surprising. As stated in the article, Africa only had 15 percent of the number of articles as Europe but had almost twice the population and Antarctica had more articles than Africa. This suggests that there is much unknown about Africa and that their technology is in much need of improvement.

Task 2.1: Download the SRC.xml file linked here. Examine its contents, then run your search indexer and querier on it.

Task 2.2: What worked and what broke? With reference to the components of your Indexer, describe 3 assumptions in your code or in the provided src code that make it difficult to search through a corpus written a language whose linguistic rules differ from those of English. [1 paragraph]

First, there was no stemming of Thai words or deletion of Thai stop words. Therefore, one assumption made by our code is that the language will always use Latin alphabet because our current regex only detects English words. We know that it only works for English words because in examining the documents file, the words to document frequencies HashMap only contained English words as keys and had no Thai words.

Second, when trying to query Thai words in search, there were no titles outputted. As a result, it can be said that our code for search assumes that the query word is made of words consisting of the Latin alphabet.

Third, some Thai sentences were very long and contained no spaces. As a result, no Thai words were saved in words to document frequencies. Therefore, we can say that our code assumes that individual words are separated from each other by spaces (" ").

Task 2.3: Suppose that your Search project will be used in Google's next search engine update. Fill in the first row *(Fairness/Inclusivity)* of the social threats framework table from Lecture 18: Identifying Social Threats with regards to your search engine. Don't worry about making your answers in each cell perfect; this task (only Task 2.3) will be graded for completion, as we mainly want to see what your thoughts are so far.

Data - only uses wiki files which could be edited by anyone. This threatens the quality of information as data could be maliciously or unintentionally made inaccurate. Therefore, while it is representative of the population as all users are able to edit wiki files, data trustworthiness is harmed.

Agency – stakeholders are able to not only see data but also edit data. Generally, all stakeholders have similar rights and privileges as they can all edit data assuming they have not been banned. users are influenced by search results

Algorithm – fairness and inclusivity of the search algorithm is harmed as edits to data are mainly represented by European and North American men. The algorithm should generally produce similar outputs for two stakeholders in the class who should be considered similar.

Task 3.3: With reference to the Noble, Srinivadan and what you have learned from Part 1 and Part 2, answer the following: can search algorithms be fair? If so, what should they encompass and why? If not, why? [2-3 paragraphs]

They can be fair but they aren't fair as of now. As stated by Srinivadan, prior to his trip to Cameroon, he Googled Cameroon but all the search results that he received in the first page were not from local authors. As a result, the knowledge pertaining to Cameroon is mainly presented by non-local writers since they appear in the first page of search results. This becomes unfair as search algorithms have effectively made online knowledge regarding Cameroon non-authentic and highly prone to errors (written by non-local writers). To counter this, search algorithms must be able to give some degree of priority to pages which are written by local authors

In addition, as raised by Noble, citations are counted as links to pages regardless of the use of the citation (i.e. agreeing or disagreeing with the cited information). Citations that point to non-credible information still raise the "rank" of the page containing the inaccurate information. This suggests that search algorithms are unfair because pages are ranked in terms of number of citations as opposed to its information credibility. Therefore, to make search algorithms fair and counter this flaw, search algorithms must be able to distinguish the use of the citation and improve a page's rank if it is cited positively or reduce a page's rank if it is cited negatively.

Task 3.4: Copy your table from Task 2.3 and, using what you have learned and the guidance from the questions table, update your answers for the *Fairness/Inclusivity* row for the *Data*, *Agency* and *Algorithm* columns.

Data - only uses wiki files which could be edited by anyone. This threatens the quality of information as data could be maliciously or unintentionally made inaccurate. Therefore, while it is representative of the population as all users are able to edit wiki files, data trustworthiness is harmed. Moreover, as stated by Srinivadan, data is mainly represented by European and North American men so data could be biased towards the thinking of men from Europe or North America.

Agency – stakeholders are able to not only see data but also edit data. Generally, all stakeholders have similar rights and privileges as they can all edit data assuming they have not been banned. As seen by Srinivadan's example with people in Cameroon sharing one computer among many, people in first world countries tend to have more access to the data since they all have their own devices to access the Internet.

Algorithm – fairness and inclusivity of the search algorithm is harmed as edits to data are mainly represented by European and North American men. The algorithm should generally produce similar outputs for two stakeholders in the class who should be considered similar. The algorithm can be gained by making lots of references and citations to a certain page to boost its rank so it appears higher ranked in search results. Moreover, as seen in his example about searching for Cameroon, non-locals are not participating or even represented in data as they don't appear high in search results. Language is exclusive in our search engine; viral pages can be top ranked regardless of accuracy of information meaning that search algorithms do not take account the author and his credibility. A page that is linked to numerous other pages will be given a higher ranking compared to a page that is less linked