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

I found the huge geographic disparities between the number of articles per capita written about countries quite surprising, because I didn't expect it to be so vast — in my mind, I assumed that individuals would write articles about their own countries and their Wikipedia presence would be proportional to their population. I also found the relatively lower number of edits in different countries rather surprising, since this means that most of the information on Wikipedia originated and was edited by individuals from (mostly) wealthier, more economically developed countries. Additionally, the fact that for less wealthy nations, articles were "largely written by outsiders." surprised me, because it has quite significant implications (culturally) for the education of outsiders about a specific country — if the (arguably) most authoritative wiki is by and large written, edited and published by individuals from wealthier countries, how can we be convinced that these articles are free of bias and judgment from an outsider's lens? The article partially discusses Wikipedia's projects to improve "knowledge equity" and systemic bias in its articles, which is quite refreshing.

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]

The things that worked were when running these files were the titles, IDs, page links, and page ranks - since we use xml tags/specific notation like brackets to denote titles, brackets, and IDs, the indexer still outputs the right titles for the document, computes their page ranks accordingly, and figures out which pages link to which other pages.

However, things like getting the word frequencies (as in our method, updateFrequencies) don't work with the non-roman alphabet — the SRCdocs.txt file produced by the indexer only counts the words that are written in English (eg. Brown University) and not the ones in Thai. This is difficult to calculate (and breaks our indexer) because we update our frequencies by splitting on words, which uses the regex "[^A-Za-z0-9]" (ie everything that is not a number/letter), but every character in Thai is not a number/character, so since we split on it, we remove it from consideration in the frequencies.

Since the updating frequencies part doesn't work as expected, so will our maximum frequencies counter in the indexer — it considers only words that were considered from before, so the output in SRCdocs.txt also ignores most of the corpus. Our code here also, unfortunately, assumes that all stop words are included in the list of stop words, and because they are all in English, stop words in Thai are not taken into consideration while we run our indexer. It also assumes words can be stemmed into 'base words' like in English (with prefixes/suffixes), but with character-based words or languages where linguistic rules are different, stemming the words (as we do in the Indexer and Querier) will not work as expected. Additionally, since stemming doesn't work with Thai, unless we search for exact word matches, there will be no matches found in our querier, unfortunately.

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.

For fairness/inclusivity row:

DATA : If our search engine is used by many people (as Google's is), we must be cautious as which search results appear first may influence individuals' decisions. For example, if we mis-judge a page to be 'authoritative' when it is actually spreading misinformation (like wrong voting poll times, for example), it could lead to real-life cultural consequences, like voter suppression/lower voter turnout.

AGENCY : Individuals who work for Google (especially high-ranking executives) could have the power to use their personal agenda to boost certain pages in the algorithm for their personal/business interests. For instance, Google can boost the results of Google-owned/released products above their competitors', even though their product may not be as polished/good. Google/its employees also have access to individuals' browsing data and may use this to either (i) commit identity fraud or (ii) predict/collect private information about a user to impersonate them.

ALGORITHM : Let's say we have two restaurant businesses in the same area, serving similar food, but the owners of restaurant A have quick and easy access to the Internet, and may even know a little bit of Search Engine Optimisation. By changing up their descriptions, page titles, etc, restaurant A could appear significantly ahead of restaurant B in the search engine and may lead to more people visiting restaurant A, even though it's not a reflection of their food quality (just their computer-savviness!).

Task 3.3: With reference to the Noble, Srinivasan 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]

Fairness is very difficult to define, particularly because there is a fine line to tread between equality, equity, or achieving neither. I argue below that it is impossible for a search algorithm to be completely fair because some of its basic principles are inherently biased.

First, search algorithms (and PageRank in particular, as Srinivasan mentions) use mass validation as a key method in pushing results to the front of our devices. As we've learned in Part 1, a disproportionate amount of authors and editors on Wikipedia (and similarly-"authoritative" sites) are homogeneous — men from Europe and North America, and since they were the earliest adopters and have written the most 'mass-validated' articles, their pages become authoritative by PageRank's rules. And this pattern is repeated across the Internet — if most content out there is written by individuals of a specific gender, from a specific socioeconomic background, who share similar cultural and linguistic customs, the results we see will most likely be written by these individuals, reflect their beliefs, and we may (falsely) start to believe that this reflects the beliefs of the majority, when in reality they are just the most vocal minority. In this sense, it is unfair for those who put out content that is buried underneath this vocal minority's, and also unfair for the search user — who is being misled. The example that Srinivasan uses about Cameroon is a particularly important one: when he says that he did not see a single page from the country itself on his first page of results, this remains the harsh reality of how flawed (and unfair) our search engines are: they replicate and push us into our comfort zones, instead of introducing information that may be more useful and 'authoritative', ie. originating from the country itself.

Furthermore, search algorithms magnify a lot of the "worst parts of society", as Noble mentions — it is not an equitable playing field for all sorts of ideas and identities, and although any query is (theoretically) welcomed by a search engine, it does not promote all results equally. As Noble argues, with the rise of Big Tech (especially the Big Tech firms that own search engines) comes the fact that these companies are private enterprises with profiting intent, that may be (perhaps) malicious at times. The search algorithms we use mirror the biases of a very small subset of the world population — the engineers/individuals who create this search engine. And because tech is disproportionately white and privileged, these search engines do not accommodate the experiences of BIPOC folks and may assume the background knowledge that not everyone (outside of tech) may have. Discoverability (and more recently virality) can be created artificially or out of sexist, misogynistic, racist content — and this 'gaming' of the search engine, whether intentional or not, ultimately makes search engines inherently unequal.

Despite search engines (at its current state) being far from perfectly equal, one can imagine how to make improvements to make them more equitable: first to alter the definition of authority, instead of using page-linking (as we do in our Indexer) to determine authority, to make an effort to credential-check page authors: are they promoting biased content? Are they promoting sexist/racist/misogynistic content? Improving the method of determining authority/credibility is one place to start improving search engines. The other way, as Noble mentions, is to improve how we catalog (through metadata), perhaps by giving authors more agency/choice over where they want their page to appear on search results or what search results they wish to be attached to.

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 : If our search engine is used by many people (as Google's is), we must be cautious as which search results appear first may influence individuals' decisions. For example, if we mis-judge a page to be 'authoritative' when it is actually spreading misinformation (like wrong voting poll times, for example), it could lead to real-life cultural consequences, like voter suppression/lower voter turnout. I would additionally be cautious with how authority/credibility of sites are determined, and perhaps consider the impact of having outsider-written content being promoted. There is the potential for spam-sites or low-credibility sites to point towards a specific site that is intentionally created to spread misinformation, and if not caught, this can be used to falsely influence individual decisions. Although the data is mass-validated by many, many users, we cannot be certain that this is representative of the global population — we must look for geographic, socioeconomic, gender, etc diversity within our page authors and opinions to ensure that it does not just reflect the "most popular" pages through pagerank but also ignores irrelevant or misleadingly-popular pages.

AGENCY : Individuals who work for Google (especially high-ranking executives) could have the power to use their personal agenda to boost certain pages in the algorithm for their personal/business interests. For instance, Google can boost the results of Google-owned/released products above their competitors', even though their product may not be as polished/good. Google/its employees also have access to individuals' browsing data, and may use this to either (i) commit identity fraud or (ii) predict/collect private information about a user to impersonate them. The user here does not have much agency over the system. With a few recent exceptions (eg. GDPR), the user often has little control over what their browsing data is used for, and these collection methods are often opt-out as opposed to opt-in, which creates the possibility of inexperienced users (who may not know this is happening) to be taken advantage of. One big flaw of the algorithm in terms of equity is its placement of ads and how they often mirror search results — this is where private interests (Google's need to increase the prices of their ads) conflict with reliability and fairness (these results may be irrelevant, misleading, etc.

ALGORITHM : Let's say we have two restaurant businesses in the same area, serving similar food, but the owners of restaurant A have quick and easy access to the Internet, and may even know a little bit of Search Engine Optimisation. By changing up their descriptions, page titles, etc, restaurant A could appear significantly ahead of restaurant B in the search engine and may lead to more people visiting restaurant A, even though it's not a reflection of their food quality (just their computer-savviness!). The algorithm creates different outcomes for Restaurant A and B: one will appear much higher (and may reap the benefits of being in a particularly good search engine result spot, such as more customers) than the other. Although this is a particularly harmless example, this may have more serious implications in the event of things like local elections, where not all candidates may have similar levels of expertise in search engine optimisation, access to information/monetary resources to 'game the system' and appear earlier/more frequently in resources.