Open Secrets: Investigating systemic biases in Wikipedia, the free online encyclopedia

Eamon Ma, Hisbaan Noorani, Rachel Xie, Philip Harker

March 16, 2021

1 Problem description and research question

As a free online encyclopedia run by volunteer collaborators, edits to Wikipedia articles form the backbone of the website. Without its open nature, Wikipedia would not be what it is today. However, this does not guarantee greatness — for the same reasons it is a symbol of communal knowledge and free information, Wikipedia is inherently biased. Most notably, it is biased towards the countries that articles are edited most in, and the dominant cultures in those countries. "[Our] research does show that most editors to Wikipedia come from the United States and Western Europe. And, as of 2020, our survey data indicate that fewer than 1% of Wikipedia's editor base in the U.S. identify as Black or African American. Considering these data, we can say with certainty that we are missing important perspectives from the world that Wikipedia strives to serve." (Uzzell, 2021). Along with an abundance of information of interest to Western society, there is a distinct lack of information from and for marginalized groups.

The coverage of knowledge is not spread evenly across all the Wikipedia articles. It varies based on what is desired and what is available, and often, the most available information (on English Wikipedia[1]) concerns men from "developed" countries and their interests. Only 17.82% of Wikipedia's biographies are about women. This isn't merely an issue of Wikipedia's editors specifically being unreliable, it is an issue of there not being reliable sources in the wider Internet available for editors to feel confident creating articles about topics like women in science (Erhart, 2018). In this way, Wikipedia can be thought of as a representation of society's general knowledge base.

Our research question is, **"How can we use connections / links between Wikipedia articles to determine areas where our collective knowledge is lacking?**

2 Computational plan

The data to be used is an offline copy of all of Wikipedia's articles. There are a few reasons that this dataset is offline instead of live.

1. Wikipedia request that research not be done on live versions of articles as it puts unnecessary strain on their servers — we would like to respect their wishes.

2. Changes to the dataset may occur midway through computation. This would likely lead to errors in graph generation if references to articles are edited while the computation is taking place.

3. Computational time. Working with an offline dataset is much faster than working with an online one. If the online version were to be used, the user would have to wait for the download of each article, every time that a re-calculation is run, instead of a single time at the beginning and then never again.

4. There will be significantly less bandwidth used with an offline dataset. The download will only need to be performed once, at the very beginning, and not every time the user wishes to recompute the graph. This leads

[1]For the purposes of this project, only English Wikipedia will be considered – demographics and articles of interest will vary between languages, so we will focus on only English articles to avoid confusion.

to approximately 18 GB of data savings compressed, and 78 GB of data savings uncompressed.

Additionally, there will be reduced computational complexity with the offline dataset as fewer operations will need to be performed.

Our computational plan consists of three main stages. *Stage 1, Stage 2,* and *Stage 3.* For now, it will be assumed that every single article on English Wikipedia will be graphed, however this may change later on as the scope of the project is shifted due to computational and time constraints.

2.1 Stage 1: Processing the data

A graph of links between Wikipedia articles will be created. These links are the hyperlinks visible on an article (typically in blue text) that link to other articles. External links will be ignored. A little bit more information about each article such as the word count, number of viewers, number of editors, number of edits, etc. will be stored. This means that a simple `Graph` and `_Node` structure will not suffice. A new `Graph`, `_Node` pair with instance attributes to represent the extra data that is to be collected will be created. The exact data to be collect is subject to change as the viability of collecting each metric and the use that it will provide are considered.

Due to the large size of the dataset, processing the data is a time-consuming endeavour. This means that it should be avoided as much as possible. A possible solution to this issue is to save the graph in plain-text so that a re-computation is not required every time the graph is loaded into the visualization system.

A challenge arises when considering how to store such a graph. A format will be created that saves this graph in parts. First: as a set of nodes in the graph, second: a list of the edges in the graph, and third: a file that will save all other information for each node in the graph. This file may be formatted in a manner similar to or matching a JSON or CSV file.

2.2 Stage 2: Analysis

After the graph is created, nodes with characteristics determined to be found in underrepresented articles will be singled out programmatically. After this, each article will be analyzed on both an individual and a macroscopic level. We will look for trends in the articles that we find to be underrepresented for reasons other than age of article, obscurity of topic (low demand), and other metrics. The trends discovered in this analysis will allow us to find more such articles that may not display the same severity of symptoms, and also allow us to look at the state of the internet as a whole.

2.3 Stage 3: Visualization

After this preliminary analysis is completed, a visualization system will be developed that will allow the user to find underrepresented articles, view them in relation to other articles around them, and visit the actual articles. The python library `networkx` will likely be used to create this visualization and in order to interface with `networkx`, the python library `pandas` will also be used for its `dataframe` objects. This will allow the user to discover why these articles are underrepresented in a visual manner, which many people find will likely find easier to understand than a written report.

3 Datasets

All of the data that will be used from this project is open and available at https://dumps.wikimedia.org/. (Wikimedia, 2021b) The first of the two main sources of data will be the Wikipedia pages, available here. (Wikimedia, 2021a) The most recent dataset at the time of writing is the one that will be used (2021-01-01) . This dataset is simply every single English Wikipedia page accurate to 2021-01-01, compiled into one XML file. A smaller, sample dataset is available here. All of the page datasets must be downloaded via torrent due to their large size, to help reduce server load. Due to the nature of this dataset, it is not possible to provide an example.

The second main source of data will be analytical data on the uses of Wikipedia pages, such as page view counts. This is available here. The following is an example of some of the data found in the dataset:

```
en.wikipedia 12_Songs_(album) desktop 1 J1
```

This can be better represented as a table:

subproject.project	Article	Page id	Daily Total	Hourly Counts
en.wikipedia	12_Songs_(album)	desktop	1	J1

Explanation of Data Table

- `subproject.project`

  This is the domain that the article is hosted on.

- `Article`

  This is the title of the article.

- `Page id`

  This distinguishes between the platform of access.

- `Daily Total`

  This is the total number of requests that the page receives for a given day. This is not the same thing as unique visitors.

- `Hourly Counts`

  This is a representation of when the clicks happened in the day. It can be deciphered as follows. The hours of the day go from 0 to 23. In this encoding scheme, each letter represents its direct count in hours. This means that A represents 0, B represents 1, ..., W represents 22, X represents 23. The number following the letter is the count of requests in that hour. This means that the extracted data row says that there was 1 request for the page `12_Songs_(album)` at the 9^{th} hour of the day on the first of December, 2011 (from the file name, each file represents one day).

References

Erhart, E. (2018). Why didn't Wikipedia have an article on Donna Strickland, winner of a Nobel Prize? *Medium*. https://medium.com/freely-sharing-the-sum-of-all-knowledge/why-didnt-wikipedia-have-an-article-on-donna-strickland-winner-of-a-nobel-prize-dff8e518daaa

Uzzell, J. (2021). Who tells your story on Wikipedia. *Medium*. https://medium.com/freely-sharing-the-sum-of-all-knowledge/who-tells-your-story-on-wikipedia-6d6c29f45028

Wikimedia. (2021a). Data dump torrents [Data set]. *Wikimedia Meta-Wiki*. https://meta.wikimedia.org/wiki/Data_dump_torrents#English_Wikipedia

Wikimedia. (2021b). Wikimedia Downloads [Data set]. *Wikimedia*. https://dumps.wikimedia.org/