Analyzing AO3 Tag Data with an Interactive Dashboard

Etienne Bauer

Introduction:

Archive of Our Own (AO3) is a collection of user-submitted fanfictions and other fanworks, containing over 14.8 million works in more than 71 thousand fandoms. In order for users to sort through the millions of works, creators can add tags specifying the content within their work, such as fandom, characters, common tropes and plotlines, and more. Although AO3 contains a thorough search function to sort through works containing tags, there's little in terms of analysis of the tags themselves. For example, a statistics-enjoying user wanting to see how a certain tag's usage has varied over time would have to sort by time and record the number of results for each year. This project solves that problem by giving the user an interactive dashboard to look up certain statistics for any tag on AO3, such as tag usage by year, word count, and completion status. The intended usage will allow the user to enter a tag, automatically sort to find all works involving that tag, then produce visual graphics of those works.

Data:

In February 2021 AO3 released a data dump on a post titled "Selective Data Dump for Fan Statisticians." This data was split into two CSV files – works-20210226 and tags-20210226. The works file contains information about the works published to AO3 up to the time of the data dump: creation date, language, word count, restricted or not, complete or not, and associated tag IDs. The last column contains all tags used by the author to describe the work, separated by "+". Currently, over 7.2 million works exist in the dataset. Work IDs have not been included, so there is no primary key. To remedy this, an SQL table

called “works” is created from the works-20210226 CSV file, with works assigned a unique ID in the work_id column by using autoincrementing (*Figure 1*). The “creation date” column is also renamed to creation_date, to keep column names consistent and easy to handle.

work_id	creation_date	language	restricted	complete	word_count	tags
1	2021-02-26	en	False	True	388.0	10+414093+1001939+4577144+…
2	2021-02-26	en	False	True	1638.0	10+20350917+34816907+…
3	2021-02-26	en	False	True	1502.0	10+10613413+9780526+…
4	2021-02-26	en	False	True	100.0	10+15322+54862755+20595867+…
5	2021-02-26	en	False	True	994.0	11+721553+54604+1439500+…
…	…	…	…	…	…	…

Figure 1: *Example of works SQL Table*

The tags file contains information about the tags used by authors on AO3, with six columns: tag ID, tag type (Character, Fandom, Relationship, Freeform, etc.), tag name, canonical or not, number of uses, then merger ID. Merger ID indicates the tag’s canonical version, if one exists. The tag name was only included if the tag had more than 5 works that included it, so all tags with fewer than 5 uses (cached_count < 5) were removed, leaving only named tags. Once non-viable tags were dropped, the number of tags dropped from over 14 million to just under 870,000 tags.

Id	type	name	canonical	cached_count	merger_id
1	Media	TV Shows	True	910	NaN
2	Media	Movies	True	1164	NaN
3	Media	Books & Literature	True	134	NaN
4	Media	Cartoons & Comics & Graphic Novels	True	166	NaN
5	Media	Anime & Manga	True	501	NaN
…	…	…	…	…	…

Figure 2: tags SQL Table

In order to efficiently find which works contain certain tags, the tags column in the works table is split by tag IDs on “+” and pivoted, creating separate rows for each tag ID and

work ID pair. This will allow for much faster sorting of the data in future steps. These new pairs of work_id and tag_id are saved on the work_tag_pairs SQL table (*Figure 3*).

work_id	tag_id
1	10
1	414093
1	1001939
1	4577144
…	…

Figure 3: Example of work_tag_pair SQL Table

The work_tag_pairs SQL table references the works SQL table with work_id and tags SQL table with tag_id. This allows the project to receive the name of the tag from user input, find the corresponding tag ID in the tags table, then find all works paired with the tag in the work_tag_pairs SQL table. Next, the data for the works is collected via the work IDs in the works table. Once all works associated with user-input tag are collected, any works with NaN values in any column are removed, and a creation_year column is created from creation_date. This finalized table is then saved as an SQL table called selected_works. This selected_works SQL table is used to produce the graphics seen on the dashboard.

Conclusion:

The finalized project is an interactive Dash dashboard that will allow the user to look up any known AO3 tag included in the final tags dataset and get a series of visualizations summarizing the statistics of works involving that tag. Users can use this dashboard to see the popularity and change over time of their tag, work distribution by word count, and other useful information they may be interested in. Because the user may not know the specific tag they are looking for, any input not found in the database will return suggested potential

tags similar to the input that the user may have been looking for, based on partial matches. An example of the graphics produced by the dashboard, showing results for a found tag, is shown in *Figure 4, 5, and 6*.

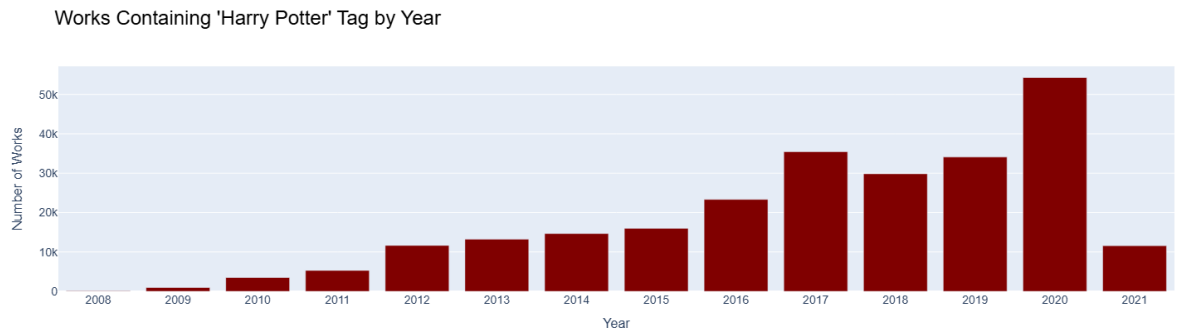

Figure 4: Year-sorted graphic result. Note that data dump was released in February 2021.

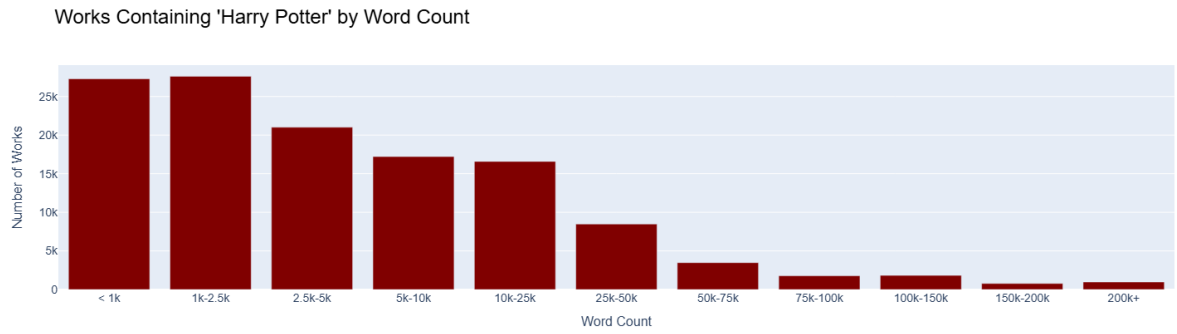

Figure 5: Example image of word count graphic result.

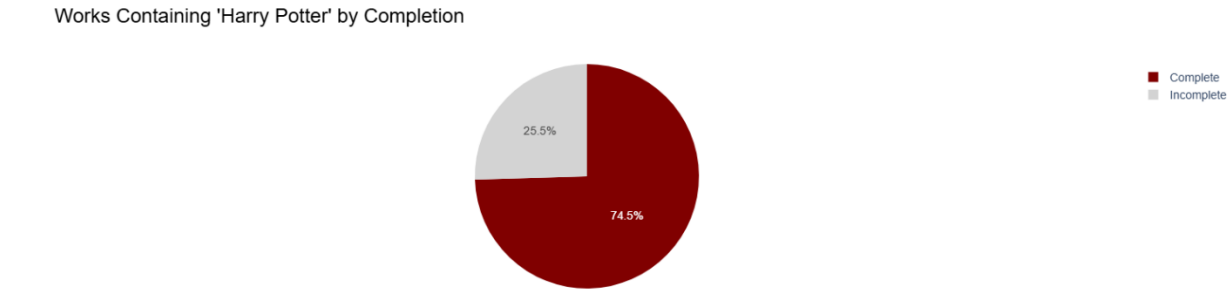

Figure 6: Example image of completion status graphic result.

Future Plans:

Future plans for project improvement involve adding a more complex search option, for example, one that may allow the user to look up two tags at once and view works containing both tags. I would also have liked to have a better autocorrecting function for user tag lookup, for any typos or similar tags. I explored using fuzzy matching, but that did not seem to be effective for this project. Further research is required. Lastly, finding a more current version of the AO3 data would be useful, however, AO3 hasn't released a data dump since, so the 2021 version is the most recent available dataset.