**ANLYTC3: Analytics 3 - Project**

**Midterm Project Report**

**Due 17 January 2023**

• **Executive Summary**

As one of the most widely used platforms for hosting transformative creative works, AO3 boasts a vast collection of more than 12 million published content. Another reason for the website’s popularity is the elaborate tagging system it employs. The tagging system allows creators to freely use a set standard of tags which can be classified into eight official types, which are media, rating, warning, category, character, and fandom. However, despite the rich data that can be garnered from the tagging system, AO3 does not implement a recommendation algorithm unlike other similar websites. For that reason, the decision that this project aims to improve is “Which works can be suggested based on a work’s tags”. Improving this decision will improve the users’ experience with the site. To improve the decision, some descriptive and associative analysis that focus on the tags will be used. Due to constraints with the personal device used for this project, the data used will be limited to only data from original works in the dataset. After analyzing the data and visualizing it in stories, there are some information that were gathered. One is that despite being website created for the purpose of hosting fan works for media, original works are ranked 11th in the fandoms. For the time series analysis, it is shown that there is an upwards curve for the amount of works being uploaded by the month for original works. The last two years have also seen a rapid rise in number. For the percentage of category tags, the composition of category types is also similar compared to the percentages for all works with M/M fiction being the highest percentage. Lastly, for the association of freeform tags, the tags that display high correlation are Fluff and Angst, Original Fiction and Original Character, Hurt/Comfort and Angst, Magic, and Fantasy, and lastly, Oral Sex with Anal Sex.

* **Problem framing**:

Archive of Our Own (AO3) is one of the most popularly used platform for posting fan-made written fiction, also known as fanfiction. The website is a project of the Organization for Transformative works, a nonprofit organization established by fans to serve the interests of media enthusiasts by providing access to and preserving the history of fan-made works and culture [1]. It has one of the most elaborate tagging systems compared to other similar platforms like Wattpad or Fanfiction.net, which is why the reason why the website appeals to a very large audience.

Currently, there are 63,270 fandoms, 6,642,000 users, and around 12,370,000 works hosted on the website and these numbers rapidly grow by the day. To give an example, by the end of 2020, there were only around 7 million works on the website but by the end of 2023, that number has increased to 12 million. Three years garnered a growth of 5 million additional fanfiction works, which is roughly a 70% increase.

Part of the reason for AO3’s popularity among fanfiction writers is the tagging system. The website allows for creators to categorize their works based on Each work published on the system can be tagged with eight types of tags: rating, warning, category, character, relationship, additional tags, and freeform tags.

Despite the elaborate tagging system that AO3 used, the website does not implement a recommendation system. While the website does have an advanced searching and filtering system to accompany the tagging system, users still have to manually search for works by the tags. This deficiency presents an opportunity for a recommendation system that will leverage the website’s existing tags using association techniques. For this reason, this project aims to answer the question of “Which works can be suggested based on a work’s tags”. It has been a trend for social media platforms to employ algorithms to improve user experience and such system would be beneficial to readers in the platform in finding works based on their preferences. Another possible beneficiary for this project is the creators and authors on the website, as the project may give them insights on the trends within the website.

• **Initial choices**:

The datasets used in this project are the AO3 data dump that the official AO3 website provided back in 2021. The data dump is comprised of two CSV files, which are named “works-20210226.csv” and the “tags-20210226.csv”. The works dataset contains the data for the existing works published on the website while the tags dataset contains the information on the tags used in the website’s tagging system.

The data dump that AO3 released was due to a great demand from users, scholars, and hobbyists that wish to analyze the website’s massive amount of works. The AO3 staff are clear against data scrapers that intend to use the website’s contents as a source material for AI writing. This is reflected in the works dataset that they released, as the works dataset does not contain any creative content from individual works.

The works dataset contains a total of 7,269,693 records. It has 6 columns, which are: the creation date, language, restricted, complete, word count, and tags. The creation date refers to the date of which the work was created. It is different from the publication date which can be set by the creator. Language is the language in which the work was written in with a total of 85 languages in this category. Restricted refers to whether the work is available only to registered users or available publicly, denoted by “TRUE” or “FALSE”. Completed is if the work has been flagged by the creator as completed or not; this variable is also denoted by either “TRUE” or “FALSE”. Word count is the total number of words used in that work. Lastly, tags are a congregation of that work’s associated tags. This last variable is steeped in interesting information as there can be as little as three and as many as a hundred individual tags used in a work. The tags are written as numerical tag ids and separated by ‘+’ signs.

On the other hand, the accompanying dataset to the works dataset, the tags dataset, has 14,467,138 records. The second dataset has six columns: id, type, name, canonical, cached\_count, and the merger\_id. The type is how the tags are categorized. There are nine unique types of tags, which are Media, Rating, ArchiveWarning, Category, Character, Fandom, Relationship, Freeform, and UnsortedTag. Cached\_count is the total number of usages of that tag. Lastly, the merger\_id is the id of the canonical version of that tag if it exists.

Both datasets contain records of works and tags that have been uploaded on the website from 2008 up to 2021. Due to constraints regarding my resources for this project, particularly my machine’s memory capacity and power, this project will have to greatly reduce the amount of works to be processed. Even with limiting the data to only the year 2020, it’s still 1,622,090 records. For this reason, the project will focus on a decently sized fandom to be able to simulate an environment that would be applicable to association-based techniques on the tags. The reason why fandom was chosen instead of other categories is because of the nature of the website itself that primarily puts emphasis on fandom content. The website itself puts fandom first in the menu bar because users primarily start with that when looking for something to read.

After much deliberation and testing, the original works fandom was chosen to be analyzed for this project. It is relatively neutral when it comes to the other types of tags and has existed since the beginning of the website. It is also one of the bigger fandom in the current dataset and is decently sized at 105,960 total records while still being reasonable to the capacity of the device being used for the analysis. In the grand scheme of things, this will not be able to represent the entirety of the AO3 library, but it will still be able to give insight into the relationships between tags.

In the data pre-process, the datasets were checked for null values and duplicates. Merged\_id on the tags dataset was ignored for checking null values because it is meant to either be empty or contain a tag identifier. The dataset did not have any duplicates or null values, so the first step was finished easily enough. The second part of prepping the data comes in the tag column from the work dataset. The original works were first filtered from the works dataset by searching for records with the tag id ‘2692’, which is the identifier for original works. After taking the original works, the contents of the tags column were separated and replaced with the names from the tags dataset. This is to make it easier to read the work’s tags and gain insight later. After renaming the tags on the works dataset, the records were recombined to keep all tags together but are now placed into separate columns.

A screenshot of a graph

Description automatically generated A graph with blue lines

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

Lastly, for the association of freeform tags, the tags that display high correlation are Fluff and Angst, Original Fiction and Original Character, Hurt/Comfort and Angst, Magic, and Fantasy, and lastly, Oral Sex with Anal Sex.

• **Results and recommendations**:

For future phases of this project, it is recommended that the team use a more powerful and capable machine. Data of this nature tends to increase rapidly as time passes. To gain more accurate and complete analyses on the subject, it is vital to use an updated version of the dataset used in this project. It has been years since an official dataset was given out by AO3 and while scraping is possible, it will take an incredible amount of time due to the website’s rate limiting.