COSC421 Project Proposal

Yerdana Maulenbay (24831786) Nat Scott (31533524)
Sara Srinivasan (10801751) An Tran (79499364)

October 17, 2025

1 Problem solved

In this project, we intend to evaluate ways for users of Archive Of Our Own (AO3) to find works of fiction they are likely to enjoy.

AO3 is a user-generated media platform typically used to host works of fan fiction, although it hosts a wide variety of fictional works. For example, some fans of Harry Potter may want to write their own stories set in the same universe, and stories like these are frequently uploaded to online repositories. AO3 is one of the older and more popular of these websites. Notably, the website does not have a built-in recommender algorithm. This is the problem we hope to remedy.

2 Data

For this project, we intend to use a dataset of AO3 stories and their tags that was scraped from the website in 2020. This dataset includes all publicly accessible stories from before July 17, 2020, or 6 million stories. This dataset is available freely online, uploaded to Google Drive by its creator.

This is a very large dataset and much of the information is not useful for our research questions. We cleaned the data and filtered out story content, taking only a 8GB vertical subset.

3 Nodes and Edges

This dataset constitutes a bipartite network, connecting stories to descriptive tags. For our analysis, we will be taking a one-mode projection onto the tags, such that if two tags appear in the same story, they will be connected by an edge. Our network will be weighted by the frequency of these connections.

In this undirected network, each node will constitute a tag, and an edge implies at least two stories share the tag.

4 Research questions

We will explore four research questions in this project:

1. How can we assess the importance of a tag using its centrality?
2. By analyzing clustering, can we identify communities of tags that are highly related?
3. By using analysis of similarity, can we make recommendations by connecting high-degree tags?
4. How can we systematically analyze changes in popularity over time to make recommendations?

5 Metrics

We will use several metrics in our analysis of the AO3 network:

- Centrality measures:
  - Degree: a high degree implies there are many related tags
  - Closeness: A closeness of 1 implies two tags are on the same story; in general, closeness implies they are on closely related stories
  - Betweenness: a high betweenness implies the tag is common among fandoms/groups of stories
  - Eigenvector centrality: a high centrality implies that the tag is popular across the whole network. This metric may not be possible/computationally reasonable to compute for the network, in which case it will be excluded
- Clustering: if a community of tags has a high local clustering coefficient, then tags in that community are likely to appear together in a story
- Structural similarity: If two tags are similar, then there's overlap in how they're used. In other words, there are tags closely related to both the tags being analyzed

6 Analysis

We will analyze the relationships between tags in view of assessing strategies for making content recommendations for AO3 users. By answering our research questions, we expect to be able to describe a possible recommender algorithm for the platform. Each research question explores a different method of relating stories by tags, and our combined analysis will reveal which methods work well, which ones work poorly, and which can be combined for a better recommender algorithm.

7 Timeline

- Week 2: Choosing a topic (done)
- Week 3: Data collection (done)
- Week 4: Data cleaning (done)
- Week 5: Finalizing the proposal (done)
- Weeks 6-10: Independent analysis of research questions
- Weeks 10-11: Analysis discussion and revision
- Week 12: Report and Presentation

8 Work distribution

- Sara: Research question 1, data collection
- Nat: Research question 2, proposal and video collation
- Yerdana: Research question 3, data cleaning
- An: Research question 4, final report collation, slide show creation