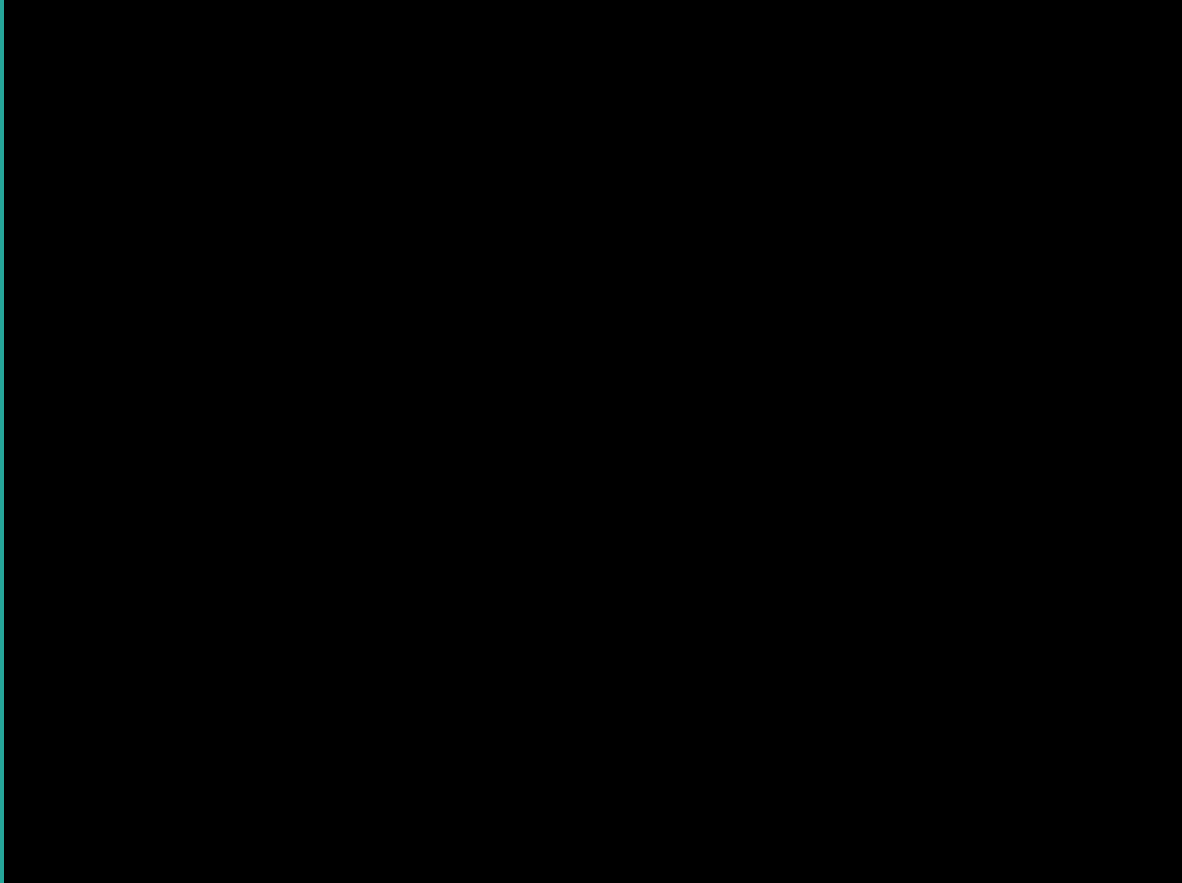


A03, Fanfiction, and Popularity: Tag Analysis in Genre Literature



DH 100 Theory and Methods | Instructor: Adam Anderson | Student:
Yvonne Gonzales |
28 May 2021 | yvonnegonzales@berkeley.edu

Watch me present this slideshow!

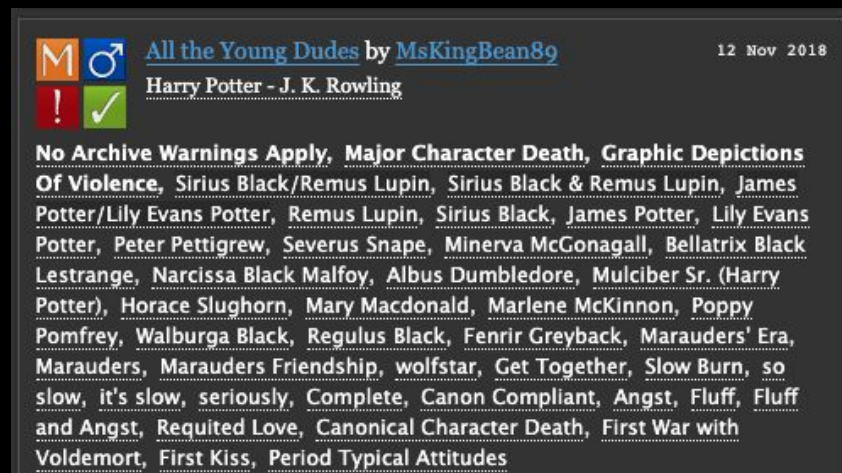
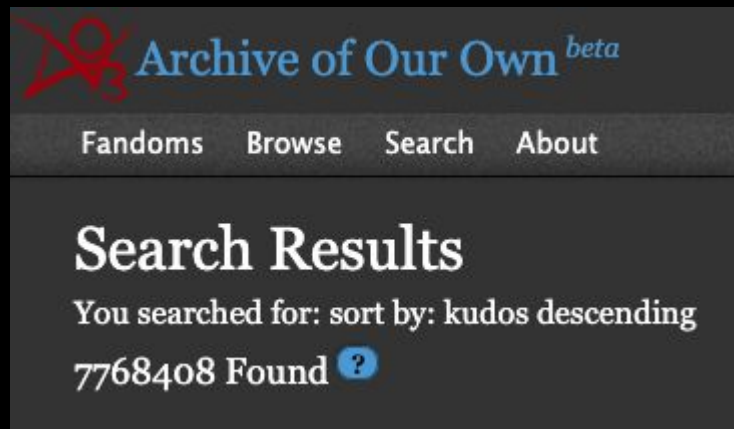


The Process



Research Questions

1. **[Primary Research Question.]** How is fanfiction tagged, and what do these subgenres say about fanfiction as a whole?
 - a. Are there any specific names or relationships that show up disproportionately?
 - b. What are the percentages of M/M stories, F/F stories, F/M stories, and other gender combinations? What might that say about the genre?
 - c. How popular is fanfiction? How popular are specific fandoms? Does the amount of clicks line up with the amount of likes?

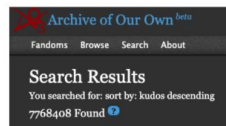


AO3, Fanfiction, and Popularity: Linguistic Analysis in Genre Literature

DH 100 Theory and Methods | Instructor: Adam Anderson | Student: Yvonne Gonzales |
28 May 2021 | yvonnegonzales@berkeley.edu

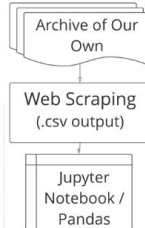
Introduction.

Reading fanfiction feels different than reading traditionally published literature. While fanfiction might come with popular connotations of being romantic or sexual in nature, fanfiction is more of a mode of writing than it is one specific genre. Using The user generated tagging systems and topic modelling, I want to figure out what makes fanfiction what it is. What topics are most popular, and how do people categorize them? In a future project, I intend to do a similar analysis of popular published literature, but this serves as a starting point for comparing fanfiction to standardized literary forms. Fanfiction has often been studied as a cultural, psychological object, but lacks research as a form of literature, and I want to investigate why it is or why it might not be.



Rank	Fandom	Words	Kudos	Comments	Bookmarks
1	Harry Potter	1,234,567	12,345	1,234	567
2	Star Wars	987,654	9,876	987	456
3	Doctor Who	876,543	8,765	876	345
4	Game of Thrones	765,432	7,654	765	234
5	Marvel Cinematic Universe	654,321	6,543	654	123

Workflow.



Dataset.

I will be focusing on a single dataset, sourced from Archive of Our Own. Sarah Sterman, UC Berkeley graduate student and one of the co-writers of the scraper I used, has been an incredible resource.

- The top 3,500 works on Archive of Our Own
 - Creator/author generated tagging
 - Hyperspecific content tags
 - Data from webscraper, generated 2021 (<https://github.com/radiolarian/AO3Scraper>)
 - .csv output
 - Full body text, over 233,000,000 words to analyze
 - Includes numbers of likes, page hits, authors, relationships, and more

Tools, Methods, and Problems.

The primary tools and libraries I will be utilizing in this project for are: pandas, spaCy, Natural Language Toolkit (NLTK), GenSim. The core models will be made through simply counting the consistency and frequency with which certain tags (user-generated sub-genres) show up. Hopefully I will then be able to see if there's any correlation between the tags and a fic's popularity, and use them to figure out how users categorize within fanfiction communities

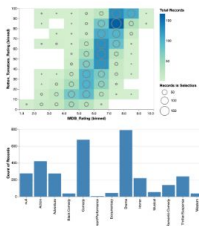
For the visualizations, I hope to use: matplotlib, seaborn, and Altair Viz for barplots, scatterplots, heat maps, and other visualizations. If I am able to create an edge list using Doc2Vec I will use NetworkX or Gephi for network visualization. My Word2Vec model is meant to be used in code for specific associations I plan to investigate, particularly in character associations

My main issues so far have been in my lack of experience with NLP and visualizing networks, but since the networks won't answer my questions most effectively I am focusing on genre tagging and barplots for now.

Research Questions.

1. [Primary Research Question.] How is fanfiction tagged, and what do these subgenres say about fanfiction as a whole?

- Are there any specific names or relationships that show up disproportionately?
- What are the percentages of M/M stories, F/F stories, F/M stories, and other gender combinations? What might that say about the genre?
- How popular is fanfiction? How popular are specific fandoms? Does the amount of clicks line up with the amount of likes?



Altair Visualizations: "Interactive Chart with Cross-Highlight." 2016-2019. https://altair-viz.github.io/gallery/interactive_cross_highlight.html

Matplotlib/ Altair Viz/ Seaborn

```
model.LstmSimilarity("harry", "harry")
[0.9462676]
model.LstmSimilarity("harry", "potter")
[0.8932435]
model.LstmSimilarity("positive", "steve")
[0.894558327182661, 0.9579649608085821, 0.988035135131367, 0.988035135131367, 0.988035135131367, 0.988035135131367, 0.988035135131367, 0.988035135131367, 0.988035135131367, 0.988035135131367]
```

Visualizations for placeholder purposes only

Searchable Algorithm emphasizing the importance of specific relationships (ships) in fanfiction

Tags on works

Full Body of Texts

TRIDF and focus on user created labelling -- Tag, rating, and genre counts in relationship to popularity

(optional route) Natural Language Processing (NLP): spaCy, NLTK, GenSim

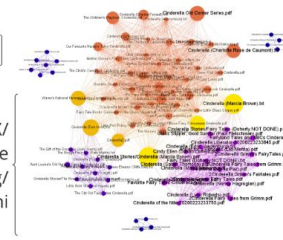
Doc2Vec Model

NetworkX/ Genre Mapping/ Gephi

Anderson, Adam. "99_Problems.png" (Gephi visualization of Cinderella stories). 2021. https://drive.google.com/file/d/1g_exna3axNyIM2BnXcbxKWN7enQpZq/view?usp=sharing

Interpretation and Results.

This is part of a much longer period of study. I have been researching and writing about non-normative literature for years, and based on extant scholarship on the subject, I expect the writings on Archive of Our Own to show very specific tropes, writing styles, ideas, and concepts. To sum up existing research on fanfiction, it is often viewed as a queer, feminine, romantic reclamation of popular culture, but I want to be able to show what that really means. Ideally, this data will become a part of my year long Honors English thesis project, where I intend to investigate gatekeeping in the publishing industry and how fanfiction pushes back.



References.

Anderson, Adam. "99_Problems.png" (Gephi visualization of Cinderella stories). 2021. https://drive.google.com/file/d/1g_exna3axNyIM2BnXcbxKWN7enQpZq/view?usp=sharing

Altair Visualizations. "Interactive Chart with Cross-Highlight." 2016-2019. https://altair-viz.github.io/gallery/interactive_cross_highlight.html

Li Jiyang and Sterman, Sarah. "AO3Scraper." 2017. <https://github.com/radiolarian/AO3Scraper>



AO3, Fanfiction, and Popularity: Linguistic Analysis in Genre Literature

DH 100 Theory and Methods | Instructor: Adam Anderson | Student: Yvonne Gonzales |
28 May 2021 | yvonnegonzales@berkeley.edu

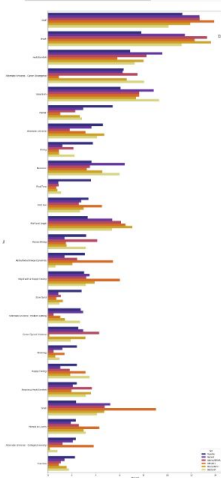
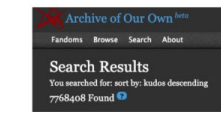
Introduction.

Reading fanfiction feels different than reading traditionally published literature. While fanfiction might come with popular connotations of being romantic or sexual in nature, fanfiction is more of a mode of writing than it is one specific genre. Using the user-generated tagging systems and topic modeling, I want to figure out what makes fanfiction what it is. What topics are most popular, and how do people categorize them? In a future project, I intend to do a similar analysis of popular published literature, but this serves as a starting point for comparing fanfiction to standardized literary forms. Fanfiction has often been studied as a cultural, psychological object, but lacks research as a form of literature, and I want to investigate why it is or why it might not be.

Research Questions.

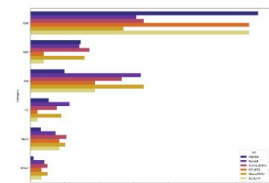
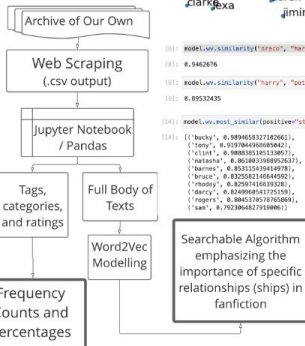
1. [Primary Research Question.] How is fanfiction tagged, and what does this categorization say about fanfiction as a whole?

- Are there any specific names or relationships that show up disproportionately?
- What are the percentages of M/M stories, F/F stories, F/M stories, and other gender combinations? What might that say about the genre?
- How popular is fanfiction? How popular are specific fandoms? Does the amount of clicks line up with the amount of likes?

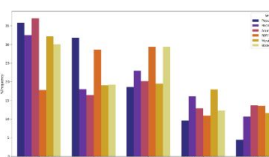


Graph 1: Tag Comparisons

Workflow.



Graph 2: Relationship Comparisons



Graph 3: Rating Comparisons

Dataset.

I will be focusing on a single dataset, sourced from Archive of Our Own. Sarah Sterman, UC Berkeley graduate student and one of the co-writers of the scraper I used, has been an incredible resource.

- The top 3,500 works on Archive of Our Own**
 - Creator/author generated tagging
 - Hyperspecific content tags
 - Data from webscraper, generated 2021 (<https://github.com/radiolarian/AO3Scraper>)
 - .csv output
 - Full body text, over 233,000,000 words to analyze
 - Includes numbers of likes, page hits, authors, relationships, and more
- Most Recent 500 works on AO3**
 - Serves as a sort of control for popular (most read) vs. what is most written
- Most Recent 500 works in representative fandoms of different media types, chosen based on popularity**
 - Anime - **My Hero Academia** (BNHA)
 - RPF ("real person fic") - **BTS** (K-pop band)
 - Movie - **Marvel Cinematic Universe** (MCU)
 - Book - **Harry Potter** (HP)

Interpretation and Results.

This is part of a much longer period of study. I have been researching and writing about non-normative literature for years, and based on extant scholarship on the subject, I expected the writings on Archive of Our Own to show very specific tropes, writing styles, ideas, and concepts. For the most part, that was true. All barplots are sorted in the order of most to least frequent tag within my core dataset of popular works.

All the most popular tags are emotionally driven, with a focus on how a work may make the reader feel. "Fluff," the most popular tag, is a term fairly unique to fanfiction, and means that the writer feels the work as something light-hearted and intends to bring no negative feeling. "Angst," the second most popular, similarly describes a feeling, rather than a theme. While many of my individual fandom samples followed the same trends as the "Popular" works, in some cases, individual fandoms or my small sample of "Recent" works had stark differences. For example, in the Marvel Cinematic Universe fandom there are more works with F/M relationships than M/M, whereas in popular works, over 66% of the relationship tags are M/M.

My initial research question on what tagging says about fanfiction, though, is that it is driven by emotions. There is no "science fiction" or "fantasy" tag, but instead "Angst," "Fluff," and "Hurt/Comfort." Fanfiction is not about themes, it is about making the reader feel a specific way.

Further Research.

That outlier spike in the lower half of Graph 1 indicates that the BTS fandom has almost twice as many works tagged "Smut" than any other fandom. In further investigation, I want to know if that trend continues across other RPF, and why this fandom is such an outlier. I also want to give the scraper more time and get bigger data sets from individual fandoms. 500 is a relatively small sample and I could get so much more information from these and other fandoms. I would also like to analyze topic models in more depth; my Word2Vec model didn't get me anything useful other than ship-based character associations but perhaps TF-IDF or Doc2Vec might

Tools, Methods, and Problems.

The primary tools and libraries I utilized for this project are: pandas, Natural Language Toolkit (NLTK), and GenSim. My main visualizations were created through simple frequency counts and comparisons between my different datasets.

For the visualizations, I used matplotlib, seaborn, and PCA with matplotlib for the Word2Vec model. My Word2Vec model is meant to be used in code for specific associations I plan to investigate, particularly in character association within fandoms.

This research and that datasets I have used are more of a proof of concept than an in-depth analysis of fandom categorization and trends. Particularly with my comparison data, I only pulled the most recent 500 in fandoms that I think would be representative of their respective media types. This is a very small sample, and there is plenty of room for further analysis there. For example, my most interesting results came out of the overwhelming presence of explicit and sexually-tagged works in the BTS fandom, and I need to do further research on how that relates to other RPF fandom, or if K-pop fandoms are outliers.

References.

- Li Jijiang and Sterman, Sarah. "AO3Scraper." 2017. <https://github.com/radiolarian/AO3Scraper>
- Kavita Ganesan. "GenSim Word2Vec Tutorial." 2020. <https://kavita-ganesan.com/genSim-word2vec-tutorial-starter-code/#YNgc5dZuc-R>

Dataset

Compiled into a .csv using a data scraper made by Jingyi Li of Stanford and Sarah Sterman of UC Berkeley.

- **The top 3,500** works on Archive of Our Own
 - Creator/author generated tagging
 - Hyperspecific content tags
 - Data from webscraper, generated 2021
(<https://github.com/radiolarian/AO3Scraper>)
 - .csv output
 - Full body text, over 233,000,000 words to analyze
 - Includes numbers of likes, page hits, authors,

- **Most Recent 500** works on Ao3
 - Serves as a sort of control for popular (most read) vs. what is most written
- **Most Recent 500** works in representative fandoms of different media types, chosen based on popularity
 - Anime - **My Hero Academia** (BNHA)
 - RPF ("real person fic") - **BTS** (K-pop band)
 - Movie - **Marvel Cinematic Universe** (MCU)
 - Book - **Harry Potter** (HP)


work_id	title	author	rating	category	fandom	relationship	character	additional tags	language	...	status date	words	chapters	comments	kudos	bookmarks	
0	2080878	I Am Groot	[sherlocksmth]	Explicit	NaN	Guardians of the Galaxy - All Media Types, Mar...	NaN	Groot (Marvel)	NSFW	English	...	2014-08-04	1308	1/1	2931	98099	38
1	8337607	Yesterday Upon The Stair	[PivViperOfDoom]	Teen And Up Audiences	Gen	Pu nohiroakademia Boku no Hero Academia My...	Midoriya Izuku & Yagi Toshinori All Might, M...	Midoriya Izuku, Yagi Toshinori All Might, M...	Alternate Universe - Canon Divergence, BAMF M...	English	...	2019-10-12	424070	60/60	19940	62276	148
2	5096105	In Another Life	[LittleLuxray]	Teen And Up Audiences	M/M	Haikyuu!!	Akaashi Keiji/Bokuto Koutarou	Akaashi Keiji, Bokuto Koutarou, Kozume Kenma, ...	Angst, Fluff and Angst, Alternate Universe - H...	English	...	2015-10-28	22995	1/1	5103	56228	9
3	10057010	All the Young Dudes	[MaskingBean89]	Mature	M/M	Harry Potter - J. K. Rowling	Sirius Black/Remus Lupin, Sirius Black & Remus...	Remus Lupin, Sirius Black, James Potter, Lily ...	Marauders' Era, Marauders, Marauders Friendsh...	English	...	2018-11-12	526969	188/188	16214	48865	107
4	5457191	a reylo fic	[treezie]	General Audiences	F/M	Star Wars - All Media Types, Star Wars Episode...	Kylo Ren/Rey (Star Wars)	Kylo Ren, Rey (Star Wars)	stan loona, #ReleaseTheRian,JohnsonCut	English	...	2015-12-19	12	1/1	8093	47051	11

Data Analysis




Jupyter Notebook (access on GitHub!)

main AO3TagAnalysis / Gonzales Ao3 Stats.ipynb Go to file ...


 **yvonnegonzales** Add files via upload

Latest commit 98ed956 19 hours ago [History](#)

 1 contributor

3348 lines (3348 sloc) | 509 KB

<> 📄 Raw Blame 🖨 ✎ 🗑

 **Archive of Our Own** *beta*

[Fandoms](#) [Browse](#) [Search](#) [About](#)

Fandom Categorization and Genre Writing

Yvonne Gonzales

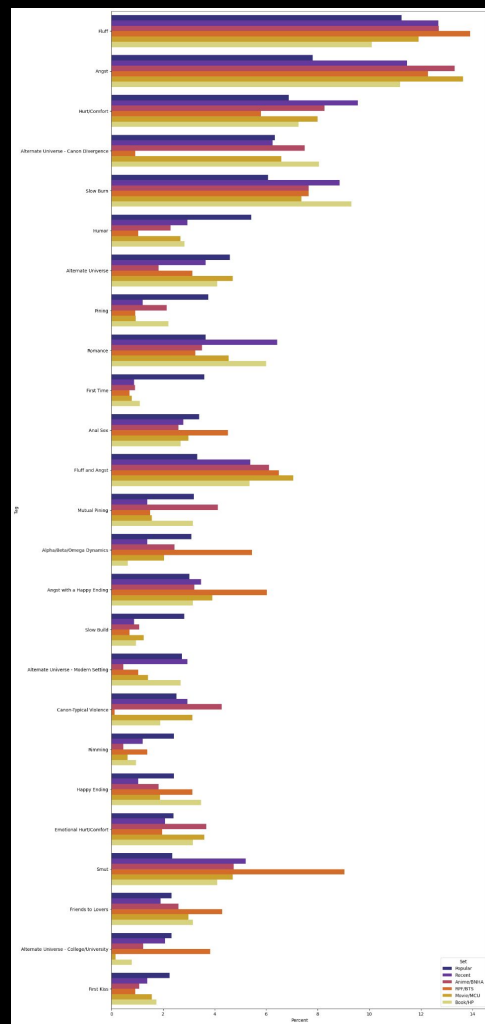
Course: Digital Humanities 100 and 101; Summer 2021

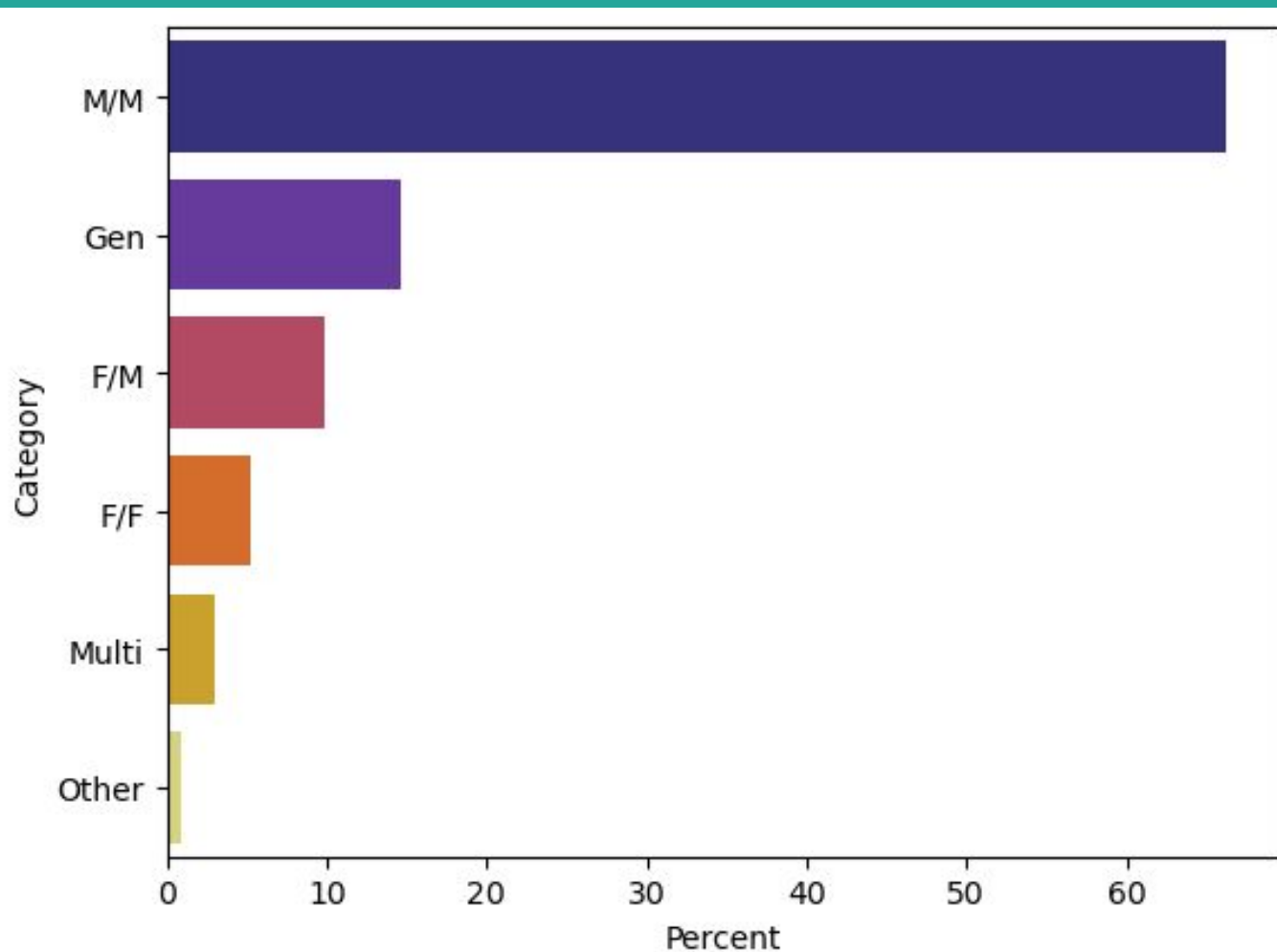
Instructors: Adam Anderson and Brooks Jessup

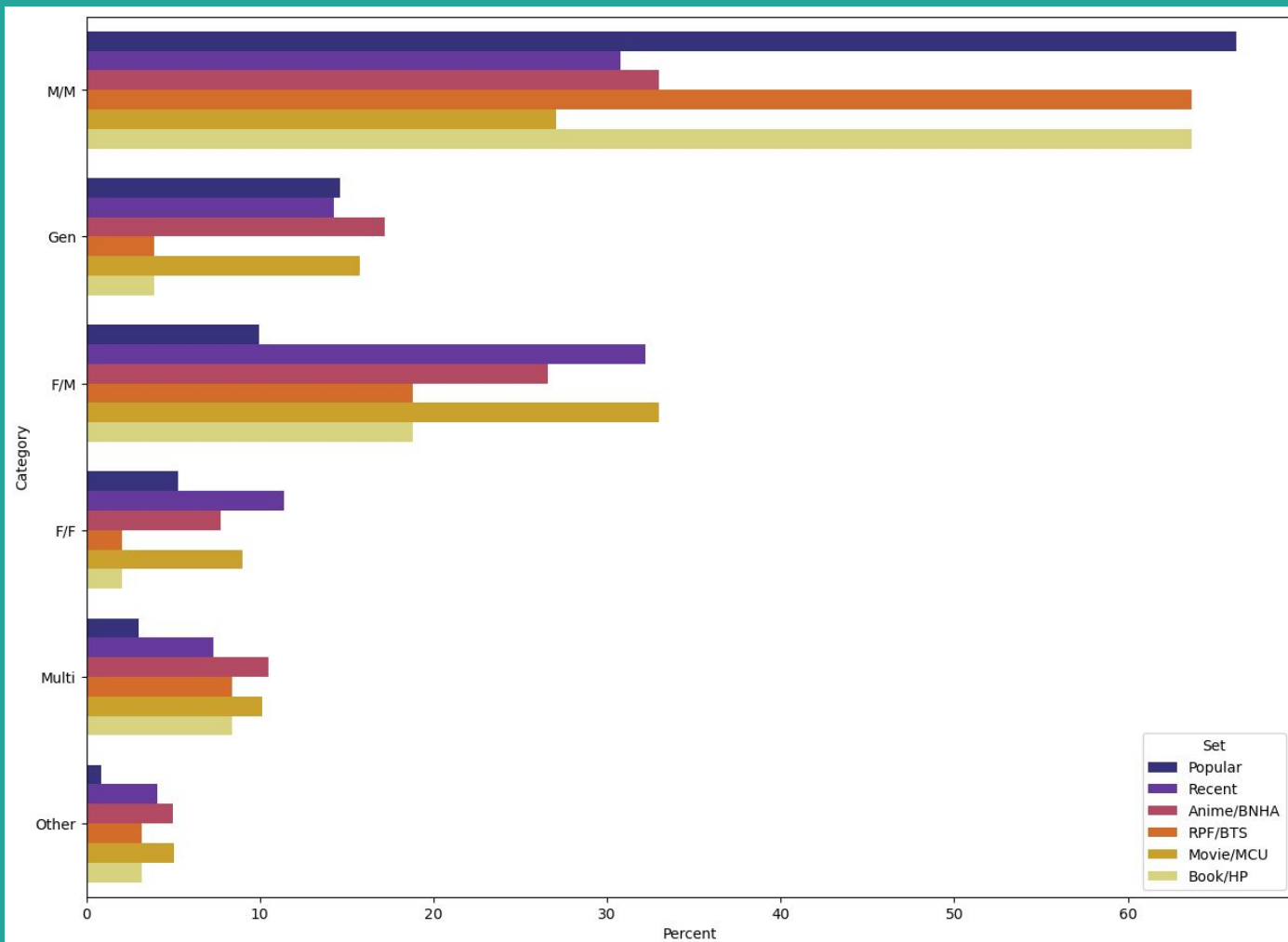
Abstract

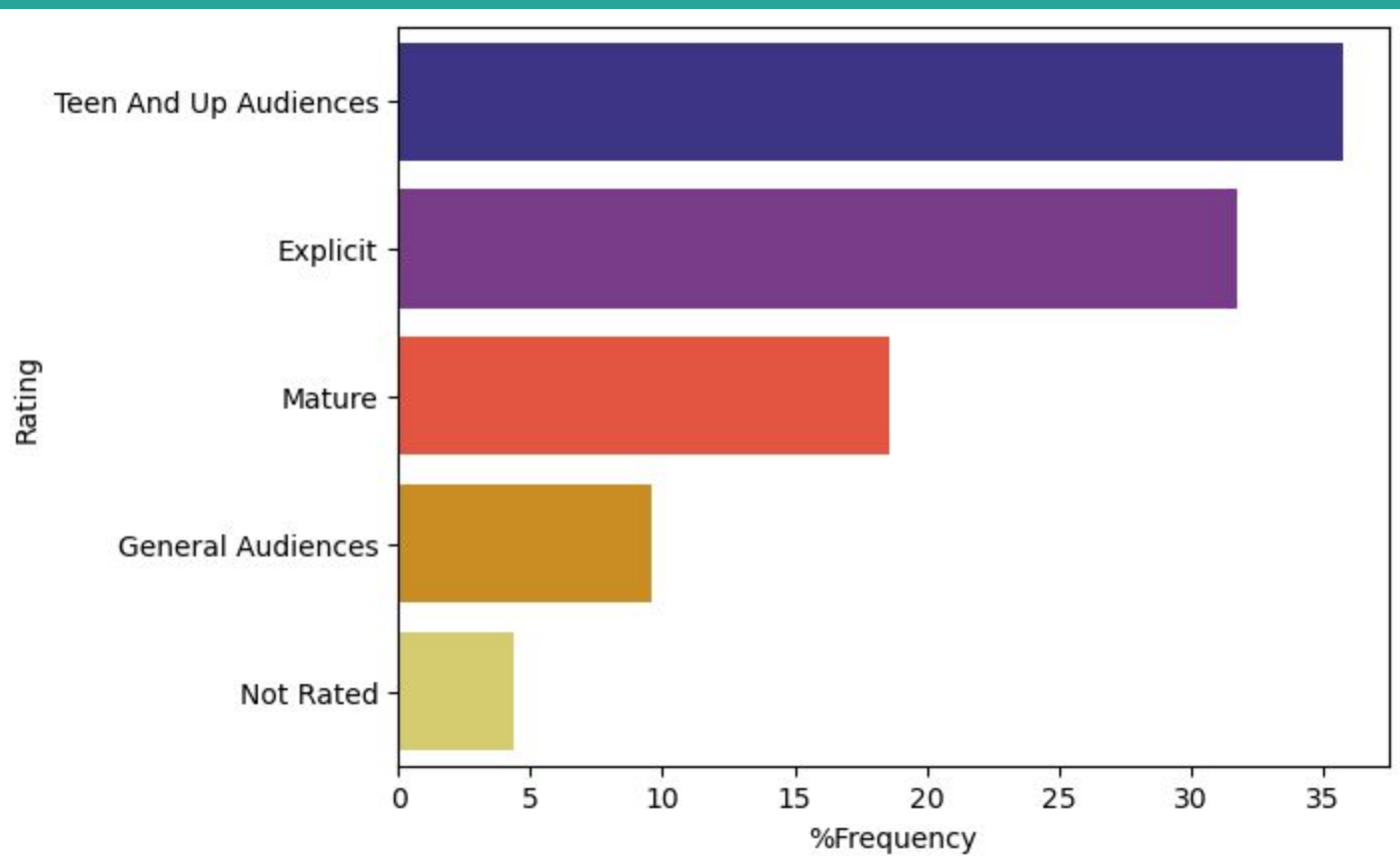
Frequency Counts

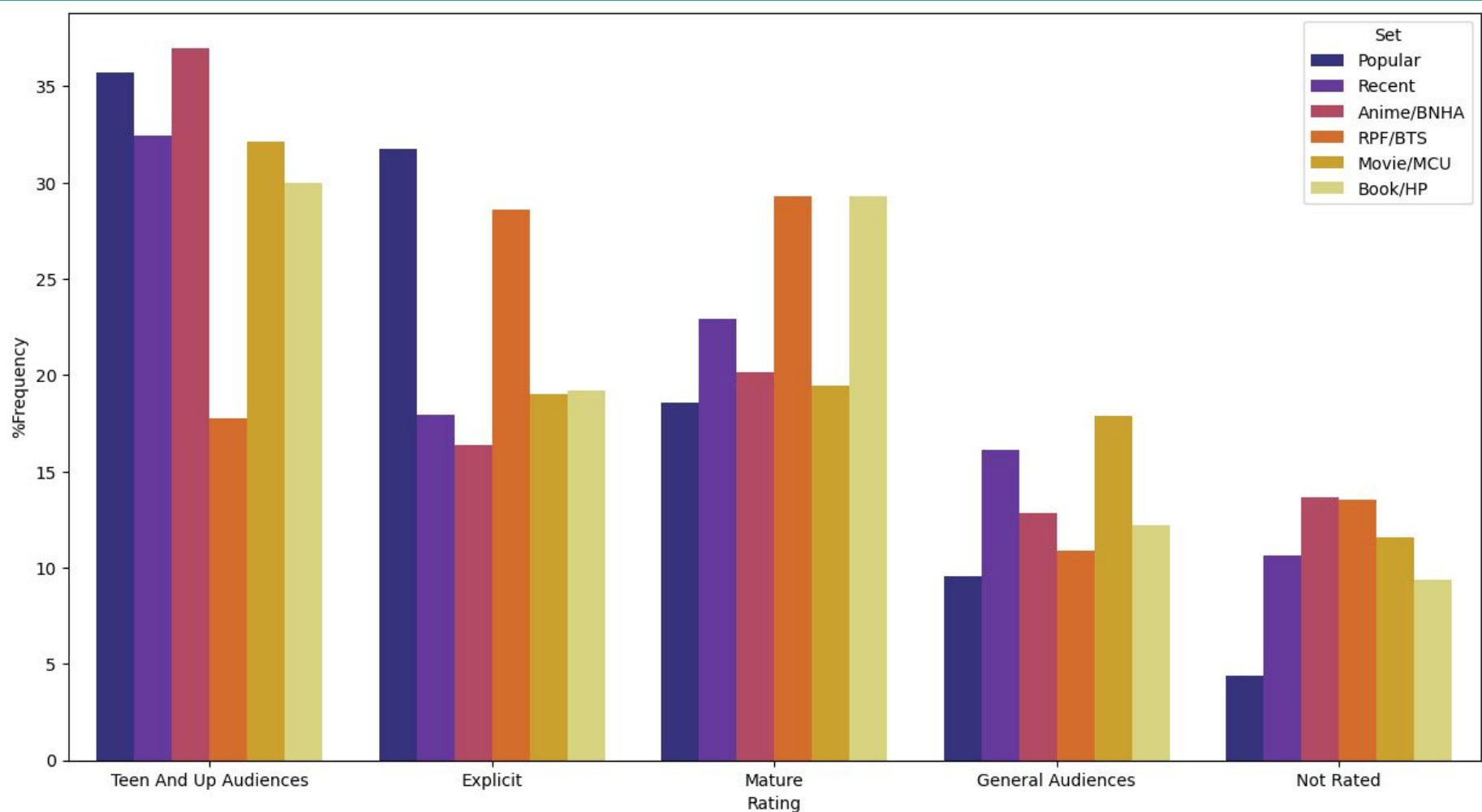
Using frequency counts within the tagging system, I will investigate the most popular subgenres within fanfiction, and see if their tags are in any way reflective of their popularity.

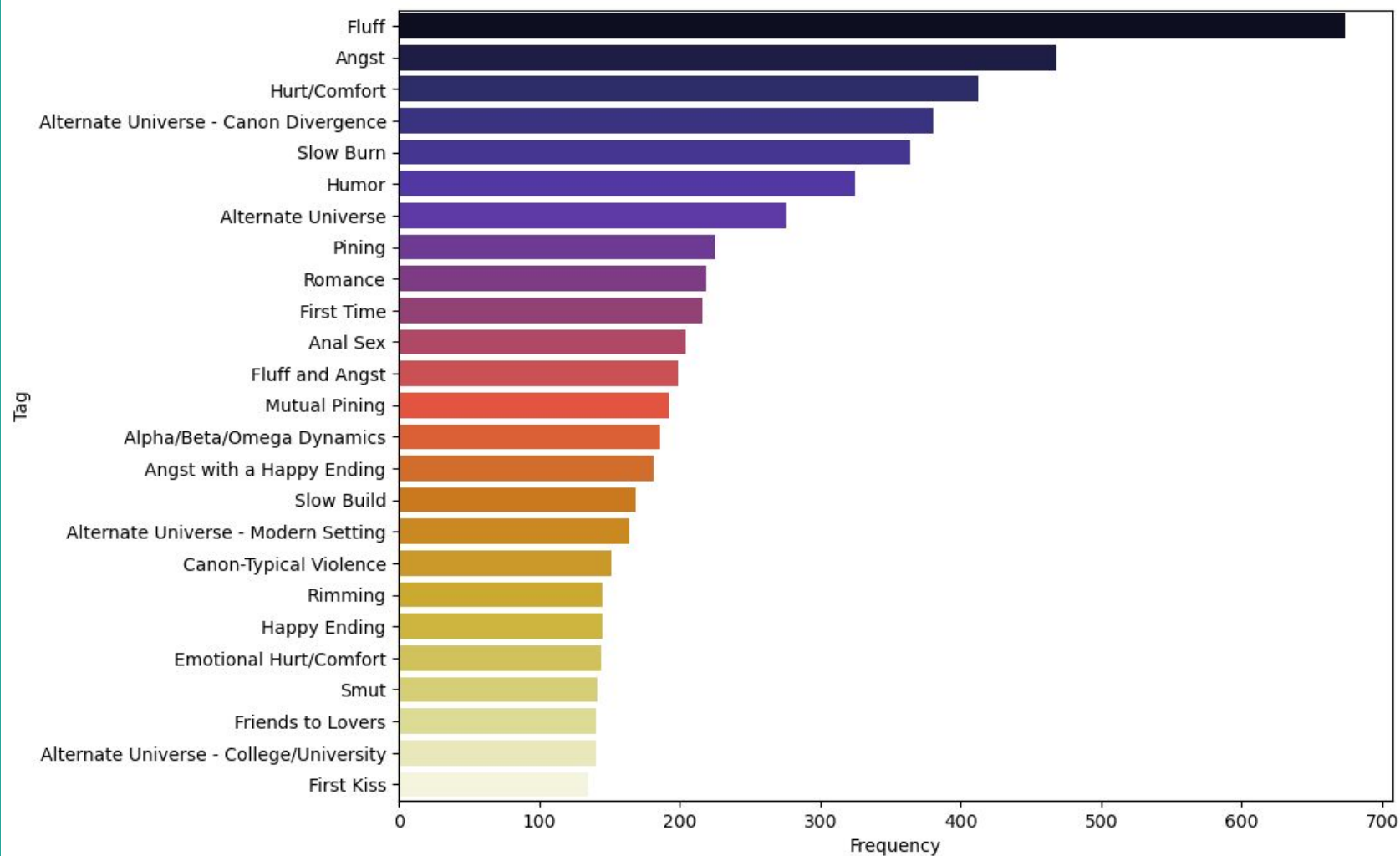




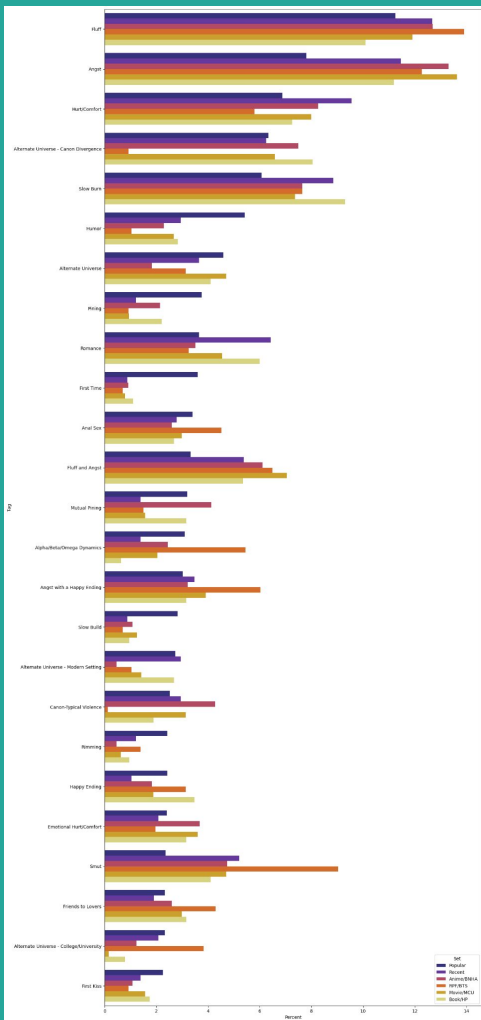








This one is... a lot.
We have to look
closer!



Emotional Hurt/Comfort



Smut

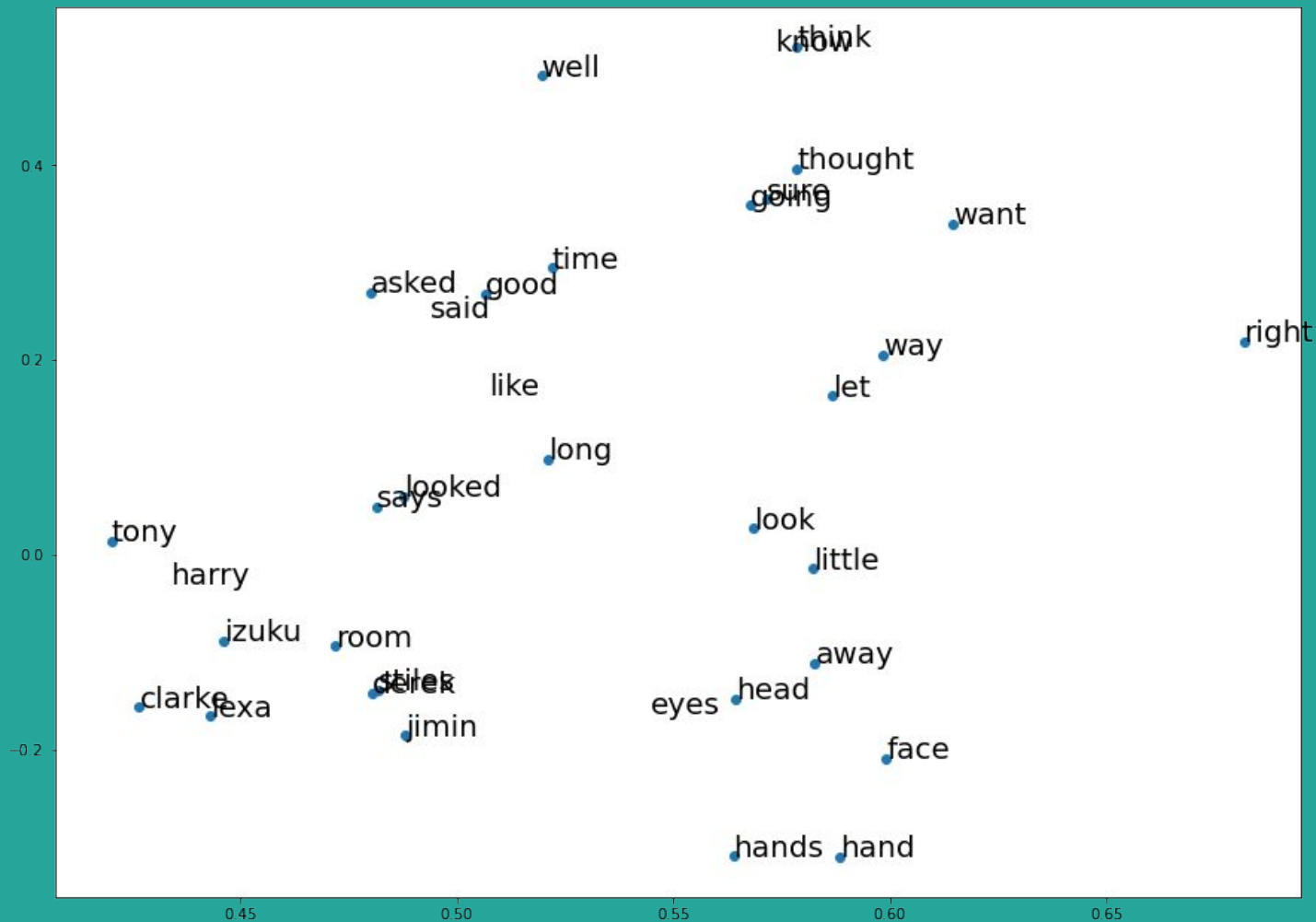


Friends to Lovers



Word2Vec Model

—



Conclusions

—

Takeaways

This is part of a much longer period of study. I have been researching and writing about non-normative literature for years, and based on extant scholarship on the subject, I expected the writings on Archive of Our Own to show very specific tropes, writing styles, ideas, and concepts. For the most part, that was true. All barplots are sorted in the order of most to least frequent tag within my core dataset of popular works.

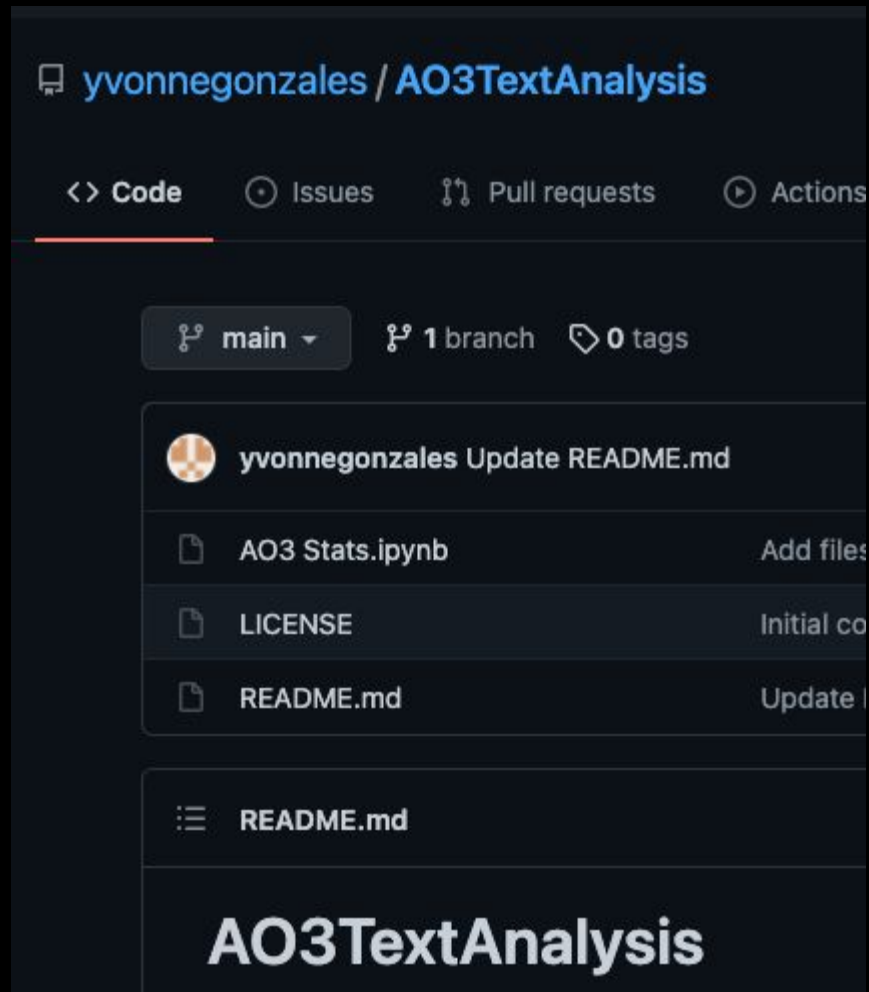
All the most popular tags are emotionally driven, with a focus on how a work may make the reader feel. "Fluff," the most popular tag, is a term fairly unique to fanfiction, and means that the writer feels the work as something light-hearted and intends to bring no negative feeling. "Angst," the second most popular, similarly describes a feeling, rather than a theme. While many of my individual fandom samples followed the same trends as the "Popular" works, in some cases, individual fandoms or my small sample of "Recent" works had stark differences. For example, in the Marvel Cinematic Universe fandom there are more works with F/M relationships than M/M, whereas in popular works, over 66% of the relationship tags are M/M.

My initial research question on what tagging says about fanfiction, though, is that it is driven by emotions. There is no "science fiction" or "fantasy" tag, but instead "Angst," "Fluff," and "Hurt/Comfort." Fanfiction is not about themes, it is about making the reader feel a specific way.

GitHub Repository

Try it out yourself!

<https://github.com/yvonnegonzales/AO3TagAnalysis>



Works Cited

Li Jiying and Sterman, Sarah.
"AO3Scraper." 2017.

<https://github.com/radiolarian/AO3Scraper>

Kavita Ganesan. "Gensim
Word2Vec Tutorial." 2020.

<https://kavita-ganesan.com/gensim-word2vec-tutorial-starter-code/#.YNgScDZuc-R>

Acknowledgments

Many thanks to my classmates
Ethan Yen and Hannah
Shlesinger, the incredible Ash
Tan, Dr. Brooks Jessup, and Dr.
Adam Anderson.