Fanfiction Classification Analysis

A Project by Rohan Bruce

Background Information

Why Fanfiction?

- I wanted to build a project around something that I’m passionate about - fiction writing
- Published fiction is not very freely available for use (unless public domain)
- I decided that I would collect my own dataset from freely available fanfiction and build my project around it

Warning! A lot of the material in my dataset is explicit. Some of my findings will reflect this.

The Original Plan

The Data

Or, why the original plan didn’t work out

My project was built around my dataset, and it has turned out to be not quite as simple to analyze as I anticipated.

Data Collection

- I used scraPy to scrape and download over 10,000 html files of individual fanfics (or more specifically, chapters) from Archive of Our Own
- I then used BeautifulSoup to parse them into searchable html files

Data Cleaning

- I got lucky here (or I was lulled into a false sense of security) with ao3's in-house tagging system.
- BeautifulSoup allows you to search html tags, and using that functionality, I was able to turn each tag (pictured) into a column of the DataFrame I created

```
<dd class="rating tags">
  <ul class="commas">
    <li><a class="tag" href="/tags/Mature/works">Mature</a></li>
  </ul>
</dd>
<dt class="warning tags">

    <a href="/tos_faq#tags">Archive Warning</a>:
</dt>

<dd class="warning tags">
  <ul class="commas">
    <li><a class="tag" href="/tags/Choose%20Not%20To%20Use%20Archive%20Warnings/works">Creator Chose Not To Use Archive Warnings</a></li>
  </ul>
</dd>
<dt class="category tags">

    Category:
</dt>

<dd class="category tags">
  <ul class="commas">
    <li><a class="tag" href="/tags/M*s*M/works">M/M</a></li>
  </ul>
</dd>
<dt class="fandom tags">

    Fandom:
</dt>

<dd class="fandom tags">
  <ul class="commas">
    <li><a class="tag" href="/tags/Sally%20Face%20(Video%20Games)/works">Sally Face (Video Games)</a></li>
  </ul>
</dd>
<dt class="relationship tags">
```

	Filename	Rating	Warning	Category	Fandom	Relationships	Characters	Additional	Text
0	1005380.html	Explicit	Creator Chose Not To Use Archive Warnings	M/M	Harry Potter - J. K. Rowling	Sirius Black/Remus Lupin, James/Lily, Peter Pe...	Remus Lupin, Sirius Black, James Potter, Lily ...	Humor, Angst, First War with Voldemort	While we've done our best to make the core fun...
1	10057010.html	Mature	No Archive Warnings Apply, Major Character Dea...	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 Friendshi...	While we've done our best to make the core fun...
2	10074443.html	Explicit	No Archive Warnings Apply	M/M	Batman (Movies - Nolan), Dark Knight Rises - F...	Bane (DCU)/John Blake	Bane (DCU), John Blake	nightwing!blake, dub-con, Fanart, molesting ag...	While we've done our best to make the core fun...
3	1008747.html	Explicit	Underage	M/M	Harry Potter - J. K. Rowling	Harry Potter/Voldemort, Harry Potter/Tom Riddl...	Harry Potter, Voldemort, Tom Riddle | Voldemor...	Angst, First Time, Work In Progress, Alternate...	While we've done our best to make the core fun...
4	10159223.html	Explicit	Rape/Non-Con	F/M, Multi, F/F	Spider-Man (Comicverse)	Peter Parker/Calypso, Betty Brant/Peter Parker...	Peter Parker, Calypso, Betty Brant, Felicia Ha...	Adultery, Femdom, Seduction, Office Sex, Three...	While we've done our best to make the core fun...

My data in its final form

As I was exploring my data, I came to realize it would be extremely difficult to build a simple classifier for it.

In [15]:

```
fan_freq = nltk.FreqDist(fans_list)
fan_freq.most_common(20)
```

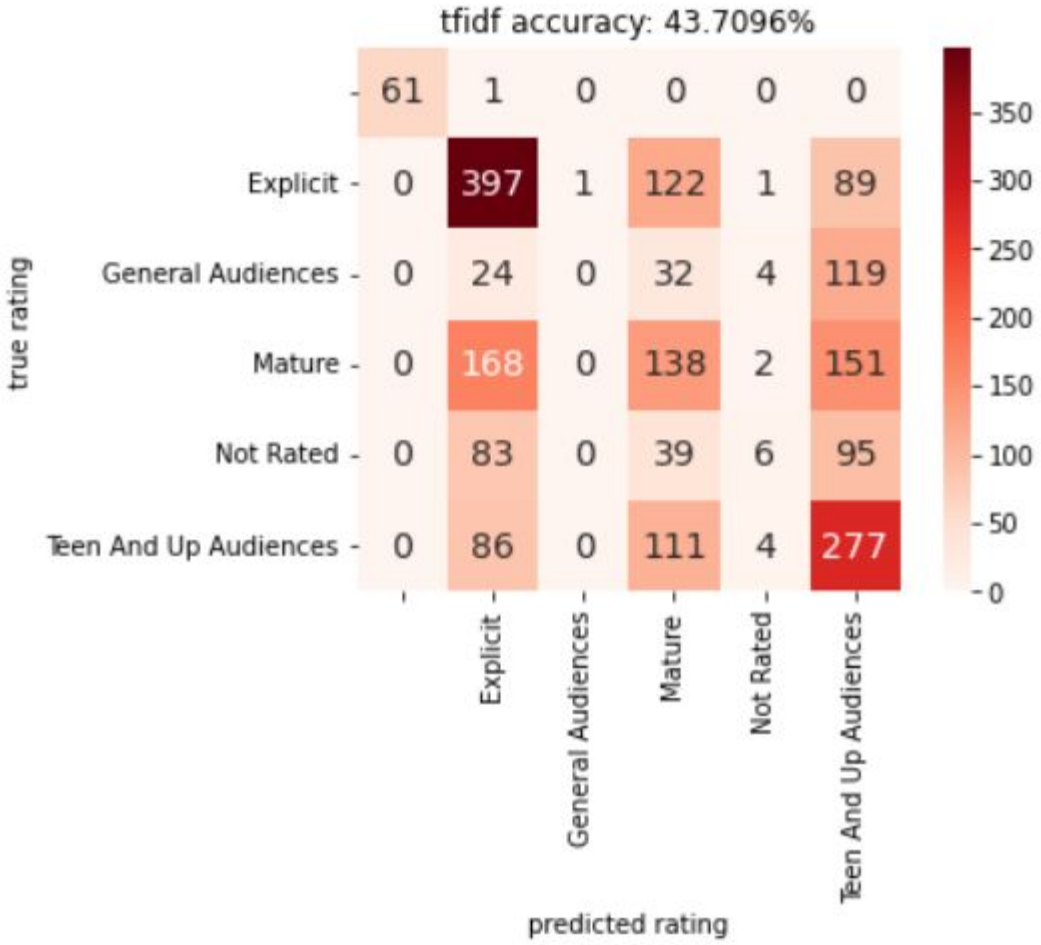
Out[15]:

```
[('Harry Potter - J. K. Rowling', 753), ('僕のヒーローアカデミア | Boku no Hero Academia | My Hero Academia', 575), ('Marvel Cinematic Universe', 304),
('Minecraft (Video Game)', 300), ('Naruto', 241), ('Video Blogging RPF', 232), ('原神 | Genshin Impact (Video Game)', 205), ('Star Wars - All Media Type
s', 203), ('Encanto (2021)', 194), ('The Avengers (Marvel Movies)', 181), ('방탄소년단 | Bangtan Boys | BTS', 167), ("Five Nights at Freddy's", 167), ('B
atman - All Media Types', 165), ('Shingeki no Kyojin | Attack on Titan', 132), ('鬼滅の刃 | Demon Slayer: Kimetsu no Yaiba (Anime)', 130), ('Biohazard |
Resident Evil (Gameverse)', 129), ('Game of Thrones (TV)', 126), ('A Song of Ice and Fire - George R. R. Martin', 118), ('Miraculous Ladybug', 116), ('H
aikyuu!!', 115)]
```

Of over 2000 fandoms, even the top 20 were not very evenly represented in my data

However, I tried anyway

Pictured here is my attempt at building a text-based classifier for ratings (the only category for which there were a reasonable number of possible tags). It… did not go very well.



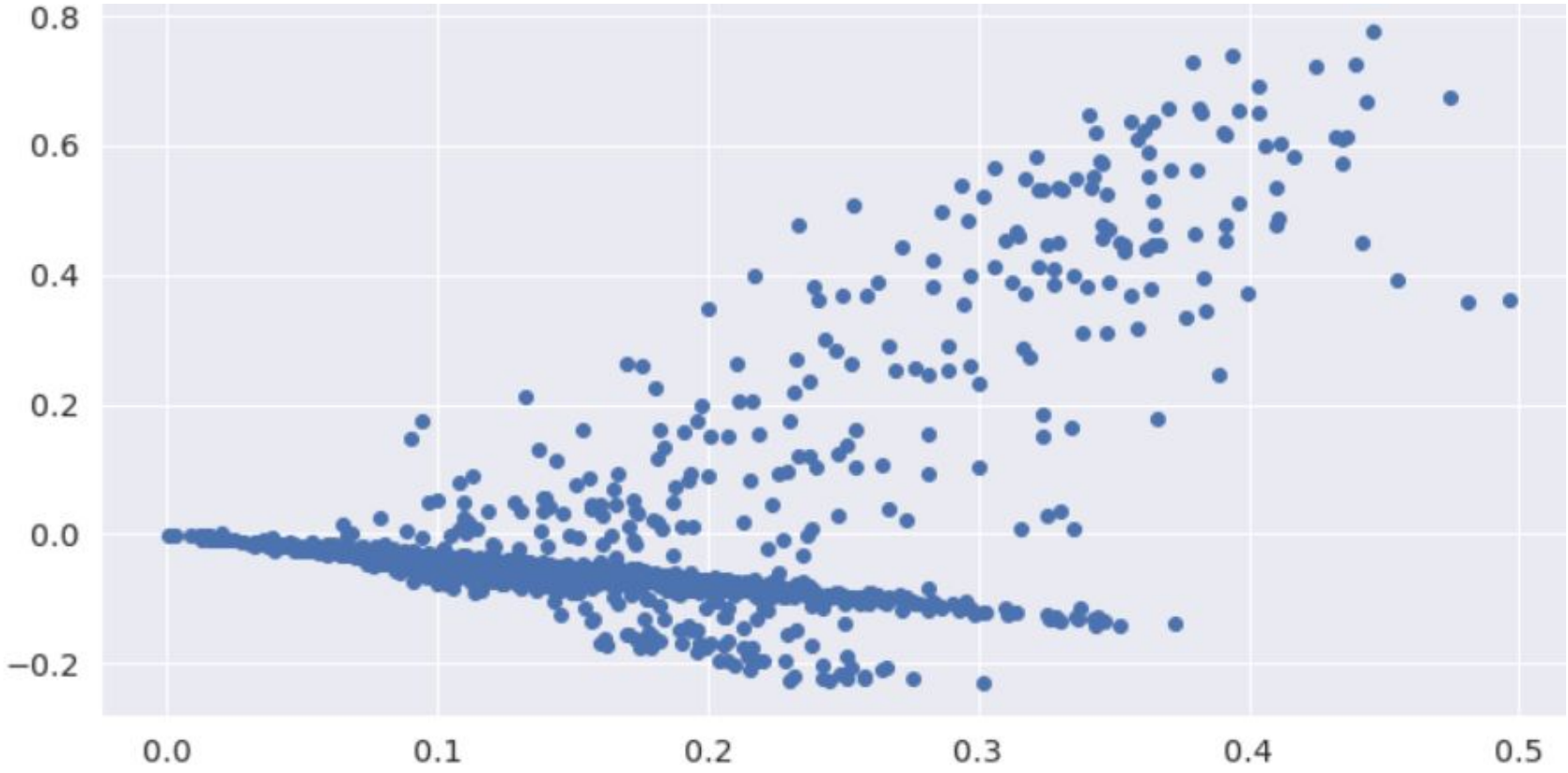
The Change of Plans

Unsupervised machine learning with the CRC

Because of my (fairly last-minute) change of plans, my analysis is very much a work in progress. Here's what I have so far.

Clustering

- Since supervised classification went… extremely poorly, I decided to attempt topic modeling on the CRC OnDemand. First, I looked at how well my data clustered together.
- I would then attempt topic modeling, deciding the number of topics based on what the clusters showed.



My clustering scatter plot

On to Topic Modeling!

- I started with four topics, based on what I saw in the graph.
- This was not particularly informative, so I moved on to 20 (pictured). This is the part where it gets, shall we say, not safe for work.
- This did actually identify fandoms reasonably well!

```
Topic 0:
said just like didn know don asked time going really
Topic 1:
harry tom potter ron magic draco lord sirius lily boy
Topic 2:
chapter 10 12 11 13 14 16 15 17 18
Topic 3:
izuku katsuki midoriya hero bakugou boy students class kid school
Topic 4:
says like doesn just asks looks don feels know takes
Topic 5:
cock cum pussy hips mouth fuck ass lips fingers tongue
Topic 6:
peter tony stiles derek kid man mr boy okay apartment
Topic 7:
tommy wilbur techno dream george fucking sam boy arthur man
Topic 8:
jungkook jimin taehyung namjoon yoongi baby kim younger car eyes
Topic 9:
hermione draco ron potter george said year father magic harry
Topic 10:
eyes hand like face head man time felt away didn
Topic 11:
mirabel family town sister room mother children vision years daughter
Topic 12:
remus sirius james lily potter year black hermione boy magic
Topic 13:
omega alpha scent heat mate smell stiles katsuki mark derek
Topic 14:
naruto kakashi sasuke 10 team blonde san training boy son
Topic 15:
jason tim dick percy kid brother red father family users
Topic 16:
que la su en el se san son ya um
Topic 17:
lexi shit like girl party fucking sister car fuck house
Topic 18:
tony steve bucky sam james team mr man room kid
Topic 19:
obi wan anakin master force luke ship war order sir
```

Moving Forward

- I separated my fanfics by rating, then did topic modeling on them again - I haven't broken that information down yet, but I'm hoping to do some comparison between topic modeling and actual fandom value counts soon.
- If I have time - I'd like to analyze how swear words break down by rating
- More graphs! I'm really sorry I don't have them yet.

References

- Data - https://archiveofourown.org/media
- Data collection - https://docs.scrapy.org/en/latest/intro/overview.html
- Data cleaning - https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Questions?