

A Linear Discriminant Analysis of Data from
Archive of Our Own

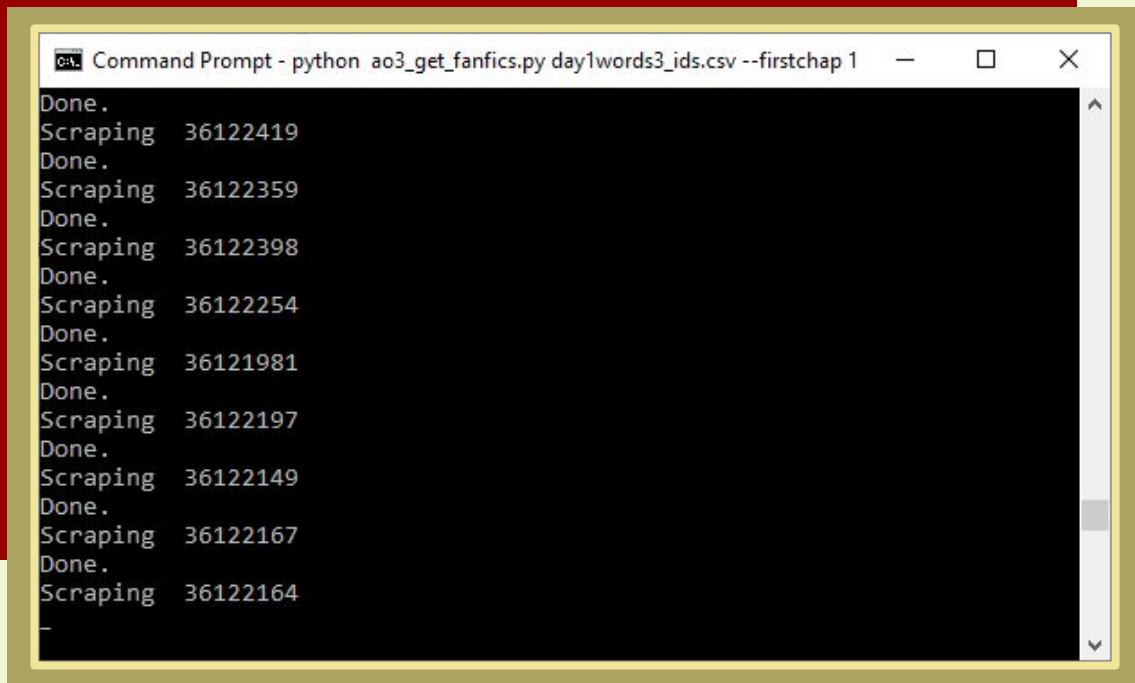
Alice Hankin for STATS 767



The Data

I used a Python web scraper to download metadata from every work that was:

- Completed in the first week of 2022
- Not locked (for registered users only)
- Written in English



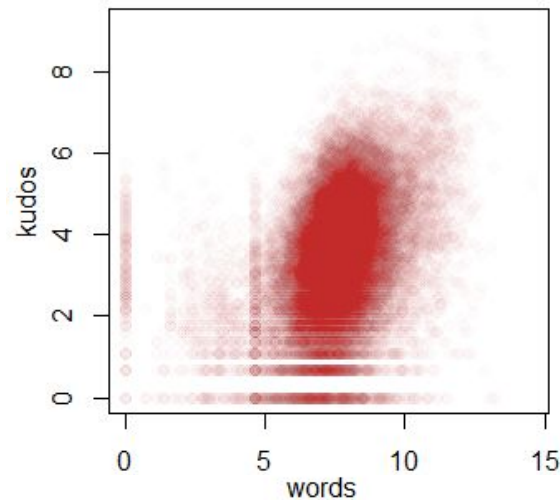
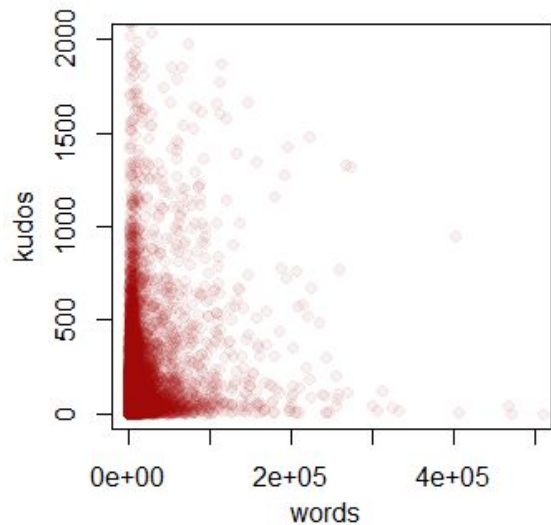
```
Command Prompt - python ao3_get_fanfics.py day1words3_ids.csv --firstchap 1
Done.
Scraping 36122419
Done.
Scraping 36122359
Done.
Scraping 36122398
Done.
Scraping 36122254
Done.
Scraping 36121981
Done.
Scraping 36122197
Done.
Scraping 36122149
Done.
Scraping 36122167
Done.
Scraping 36122164
-
```

The Data

- ~29000 works
- 8 numeric variables plus 3 categorical variables

rating	num_tags	daystocomplete	words	chapters	comments	kudos	bookmarks	hits
General Audiences	3	1	883	1/1	2	7	1	163
Mature	4	12	17182	2/2	5	19	0	412

Can we determine the **rating** from these **numeric variables** ?



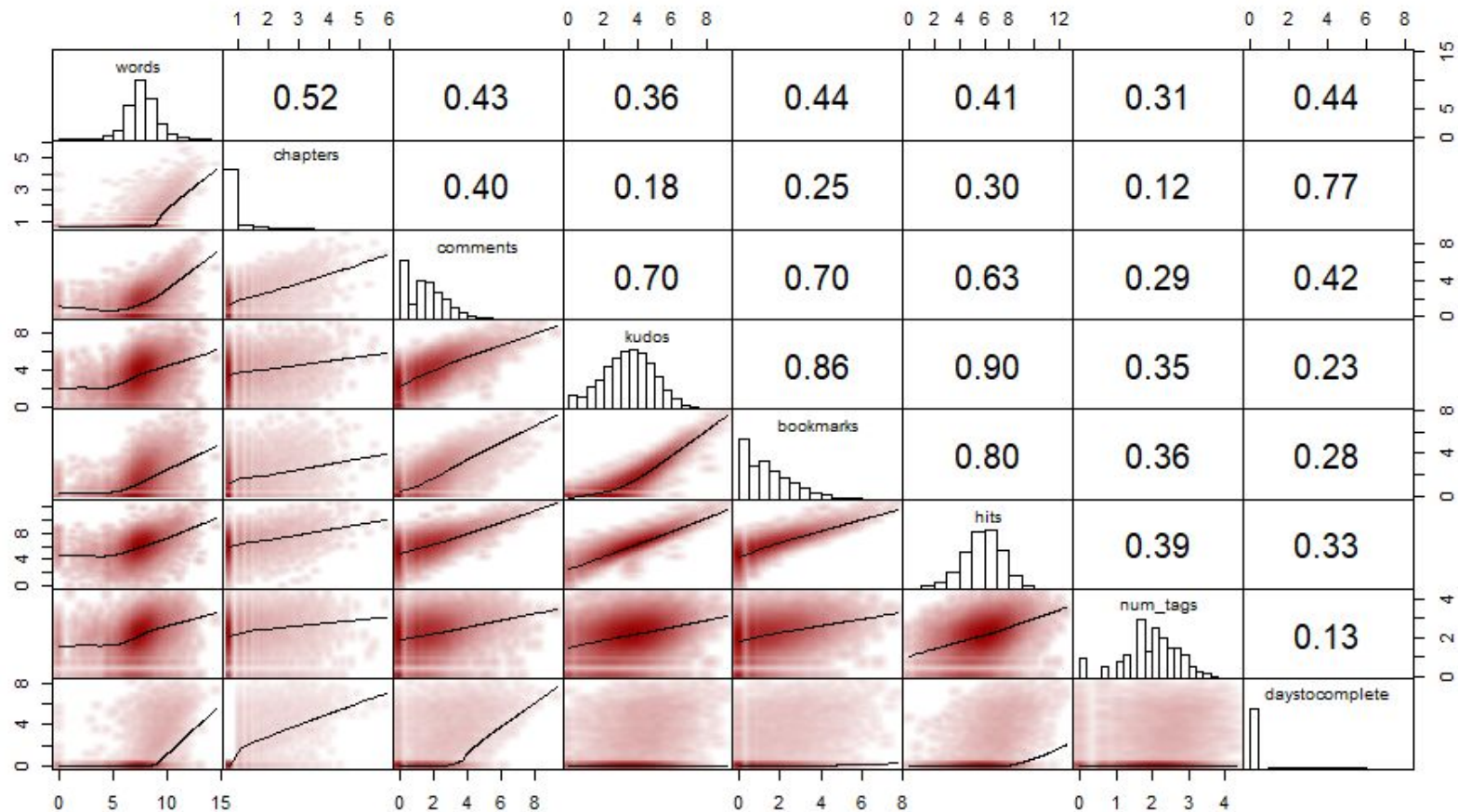
- Why do we need to log the data?
- What are these strange vertical lines?

Drabble

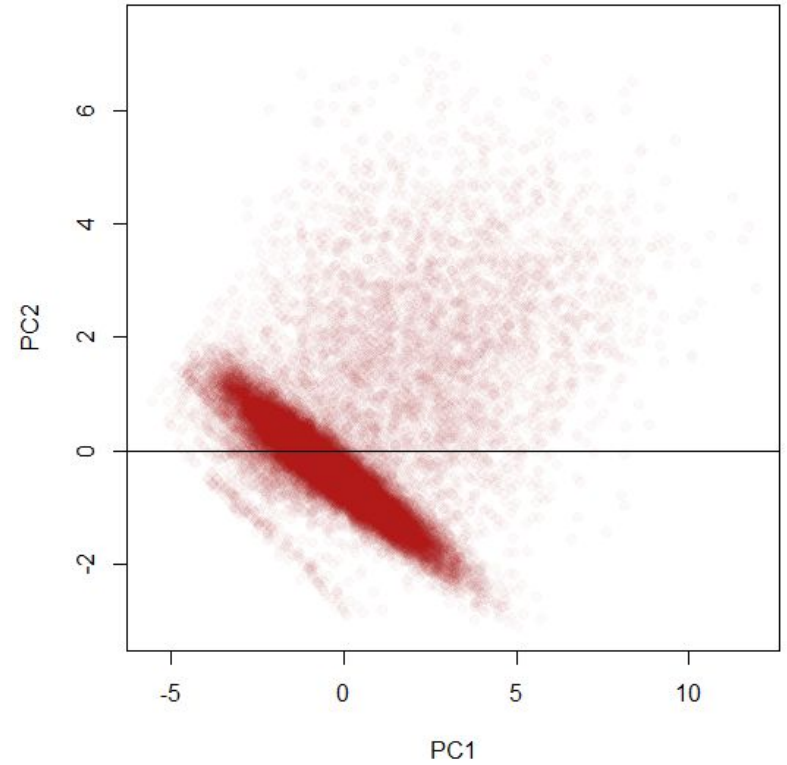
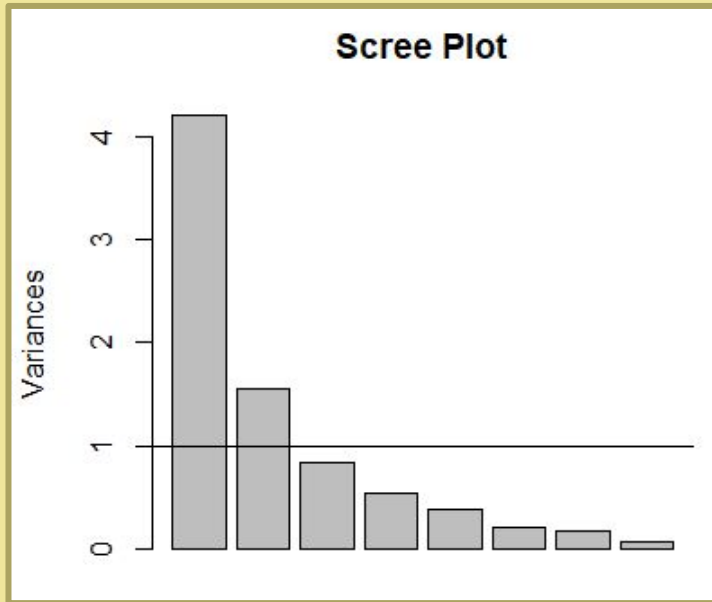
From Wikipedia, the free encyclopedia

For other uses, see [Drabble \(disambiguation\)](#).

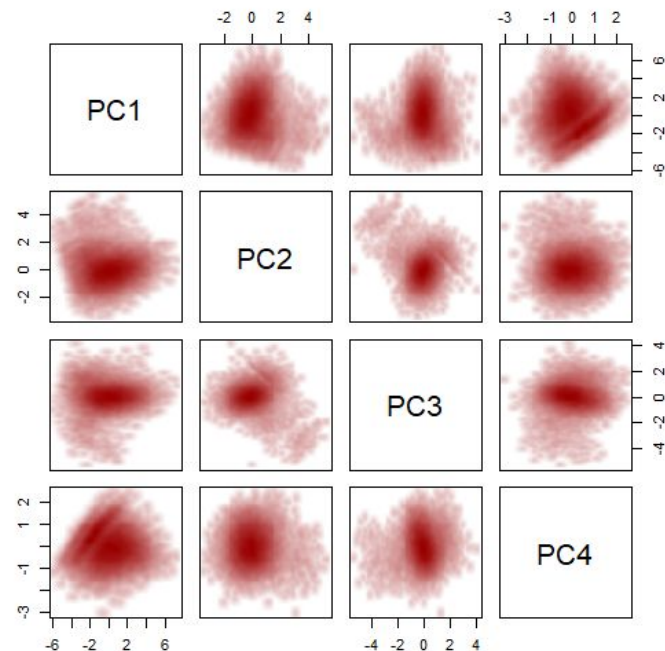
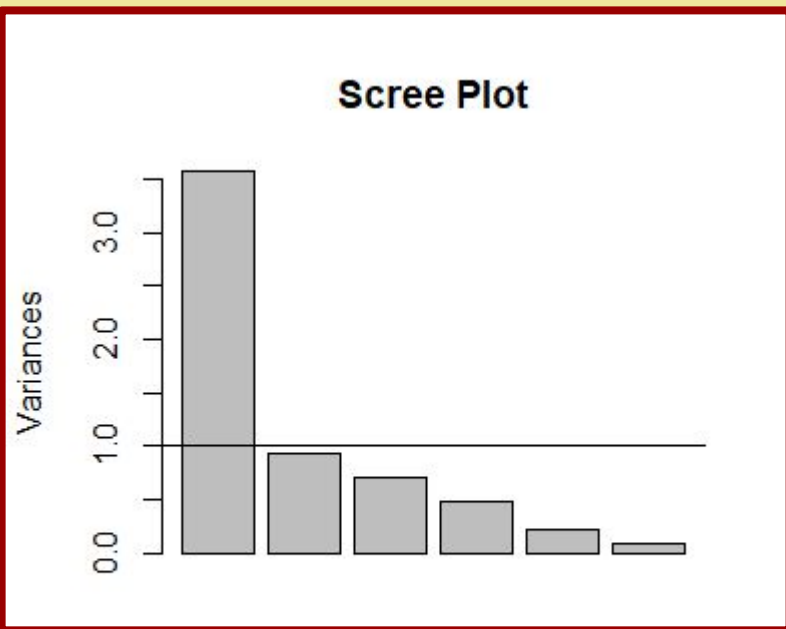
A **drabble** is a short work of fiction of precisely **one hundred words in length**.^{[1][2][3][4]} The purpose of the drabble is brevity, testing the author's ability to express interesting and meaningful ideas in a confined space.



PCA - attempt #1



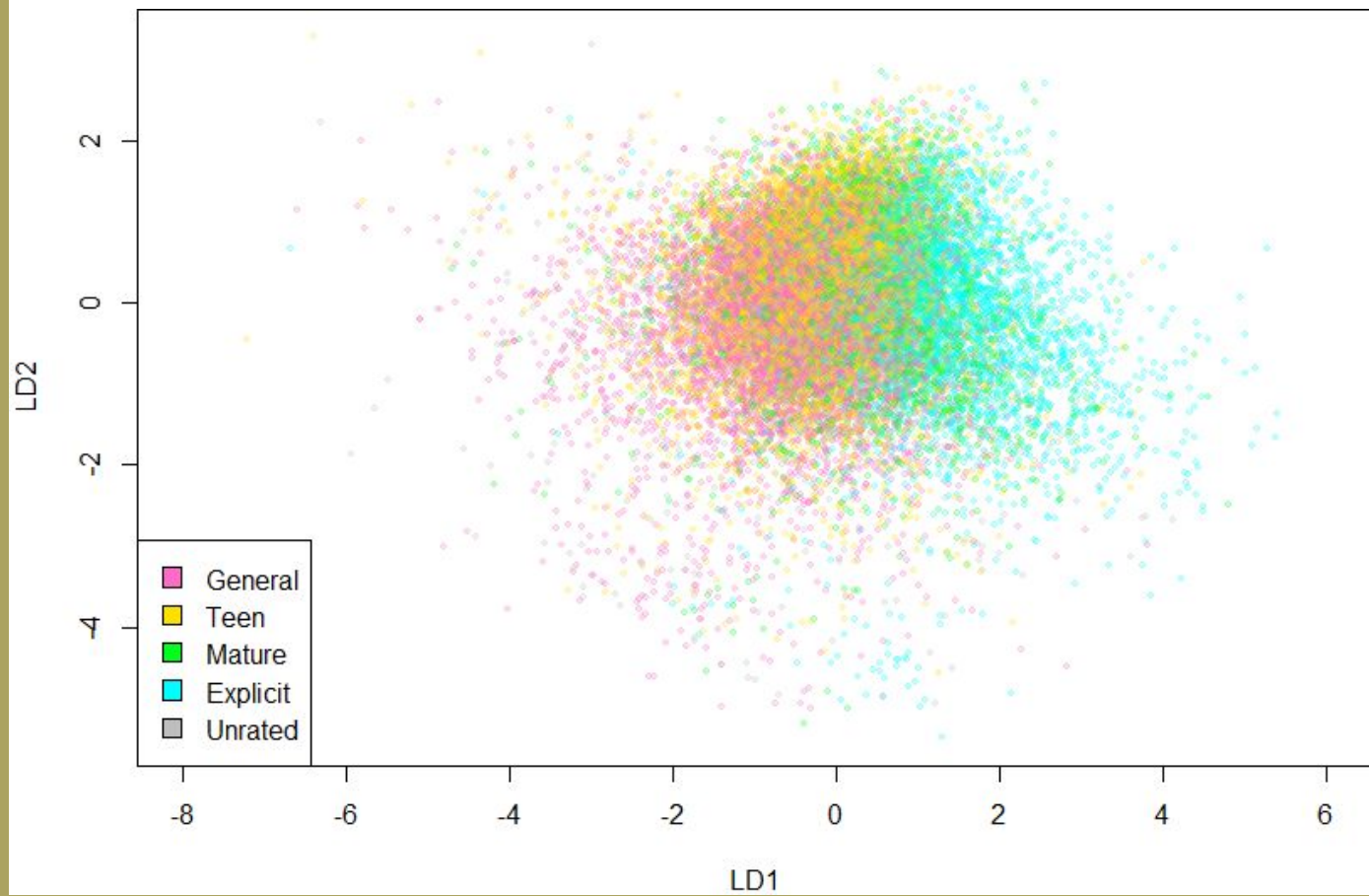
PCA - attempt #2

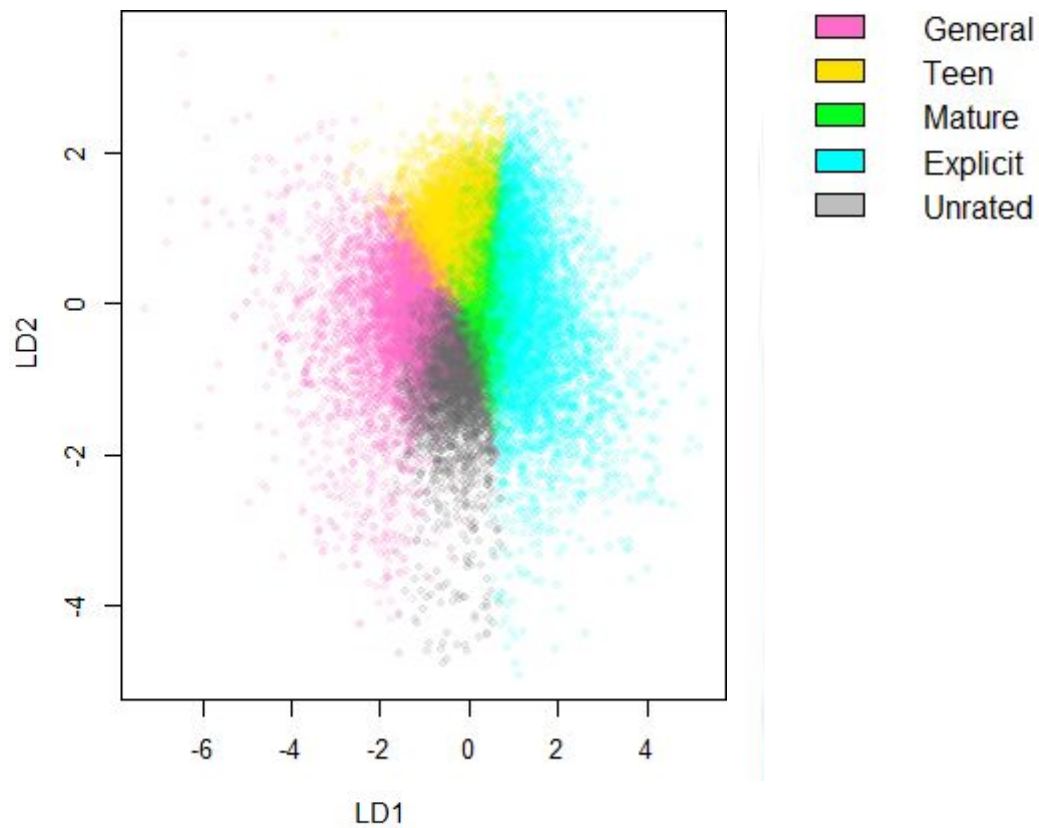
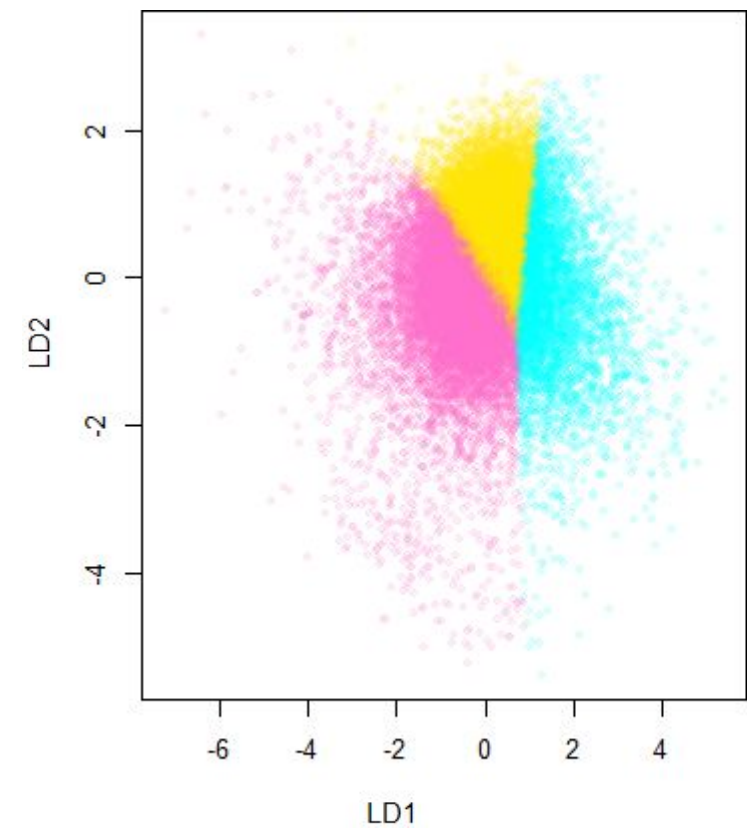


Loadings:

	[,1]	[,2]	[,3]	[,4]
words	0.506	-0.614	0.601	
comments	0.764			-0.584
kudos	0.934			
bookmarks	0.906			
hits	0.888			
num_tags	0.499	-0.645	-0.564	

	[,1]	[,2]	[,3]	[,4]
ss loadings	3.570	0.931	0.712	0.486
Proportion var	0.595	0.155	0.119	0.081
Cumulative var	0.595	0.750	0.869	0.950





	Predicted Explicit	Predicted General Audiences	Predicted Teen And Up Audiences
Actual Explicit	2545	517	1126
Actual General Audiences	451	5055	2254
Actual Mature	656	904	998
Actual Not Rated	285	1237	564
Actual Teen And Up Audiences	629	3205	3164

Misclassification rate of 55%

Loadings:

	[,1]	[,2]
words	0.435	0.816
comments		0.505
kudos	0.291	0.297
bookmarks	0.343	0.346
hits	0.664	
num_tags	0.466	0.357

	[,1]	[,2]
ss loadings	1.076	1.262
Proportion var	0.179	0.210
Cumulative var	0.179	0.390

- Hard to interpret the loadings!
 - First component - how appealing the work is?
 - Second component - how long the work is?

Limitations

- Data could not be taken all at once
- Monthly/yearly effects
- Locked works

Future Work

- Sentiment analysis of the tags or the content of the work
- QDA / test of multivariate normality
- Comparison of fandoms

Thank you!