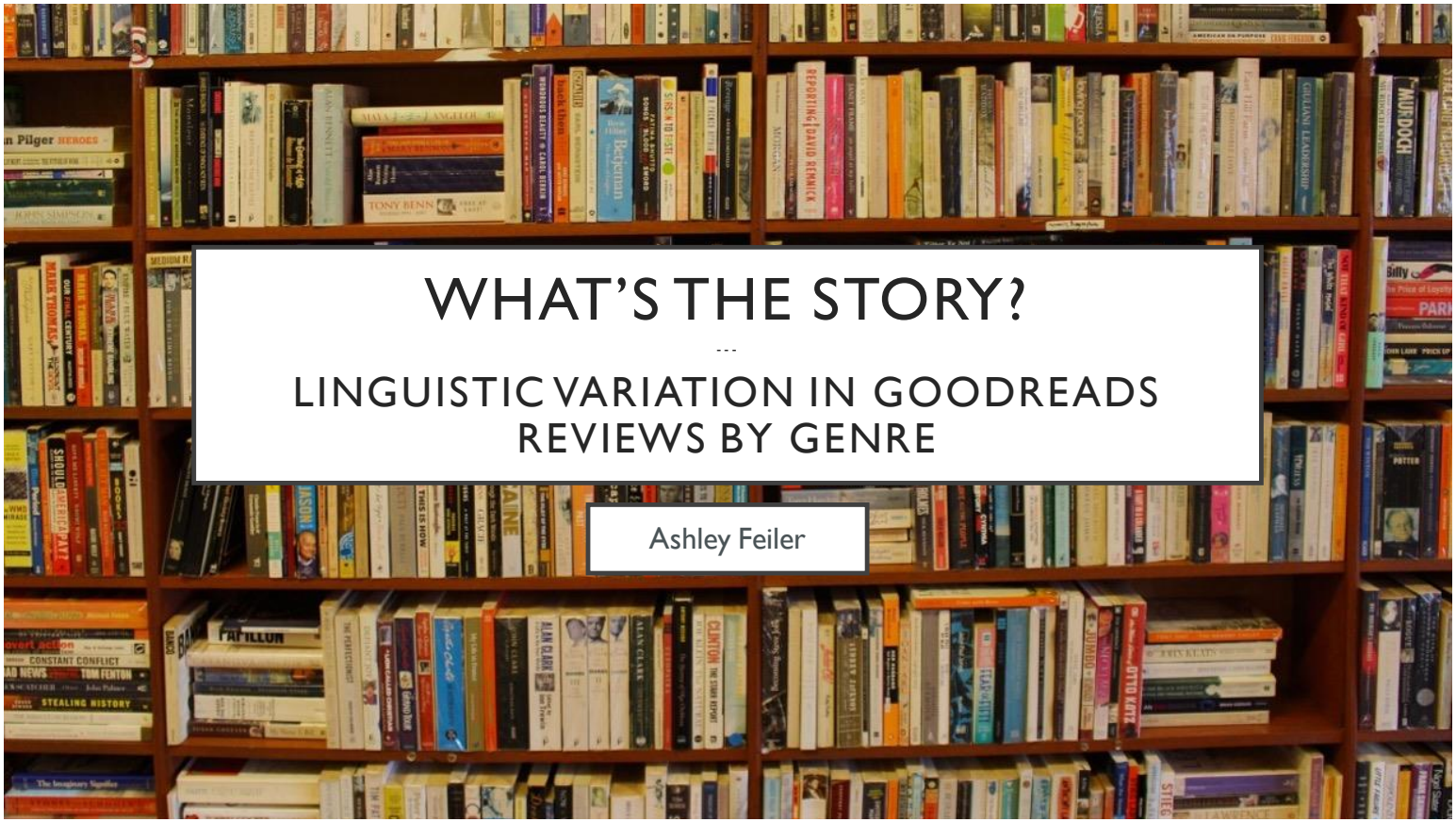


WHAT'S THE STORY?

LINGUISTIC VARIATION IN GOODREADS REVIEWS BY GENRE

Ashley Feiler

BACKGROUND INFO

- Genre most important factor for book-buyers (especially Gen Z)
- Reader engagement with genres: (Noorda & Berens, 2021)
  - Adult Fiction 71.5% (Top sub-genres: Mystery, Thriller, Classics, Romance, Historical Fiction)
  - Adult Nonfiction 55.6%
  - Young Adult Fiction 26.2% (Top sub-genres: Fantasy, Mystery, Romance, Sci-Fi, Comics)
- Genre identification is an NLP research focus (linguistic profiling) (Mendhakar, 2022)
  - Can also apply to identifying authors, languages, sociolinguistic register, etc.

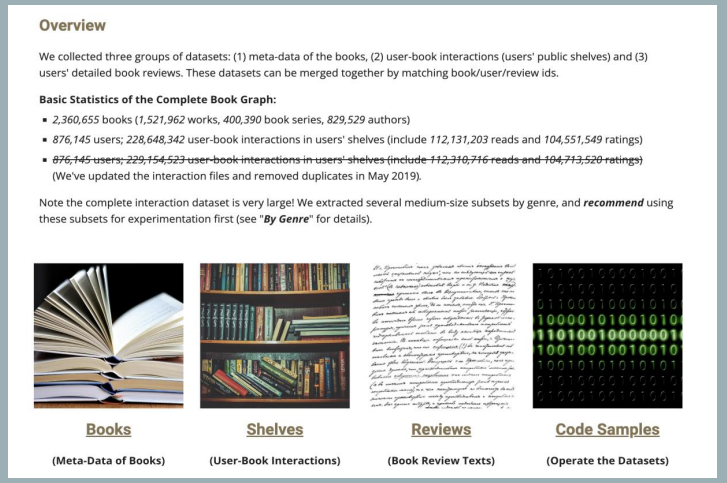
RESEARCH QUESTION(S)

Overall: How does the language used by reviewers of different book genres differ?

1. Which features of reviews are the most significantly different between genres?
2. What different kinds of adjectives are used to describe/review different book genres?
3. Which genre use the most distinct vocabulary?

UCSD GOODREADS DATASET

- Collected in 2017, updated in 2019
- Original form = JSON
- Separate datasets for review text, book info, author info, and genres
- MASSIVE AMOUNT OF DATA!!!
  - 15.7 million reviews of 2.36 million books
  - Loading the data was a challenge



Overview

We collected three groups of datasets: (1) meta-data of the books, (2) user-book interactions (users' public shelves) and (3) users' detailed book reviews. These datasets can be merged together by matching book/user/review ids.

Basic Statistics of the Complete Book Graph:

- *2,360,655* books (*1,521,962* works, *400,390* book series, *829,529* authors)
- *876,145* users; *228,648,342* user-book interactions in users' shelves (include *112,131,203* reads and *104,551,549* ratings)
- ~~*876,145* users; *229,154,523* user-book interactions in users' shelves (include *112,310,716* reads and *104,713,520* ratings)~~ (We've updated the interaction files and removed duplicates in May 2019).

Note the complete interaction dataset is very large! We extracted several medium-size subsets by genre, and ***recommend*** using these subsets for experimentation first (see "***By Genre***" for details).

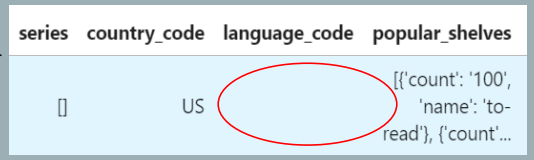
DATA PROCESSING

LARGE DATA SIZE

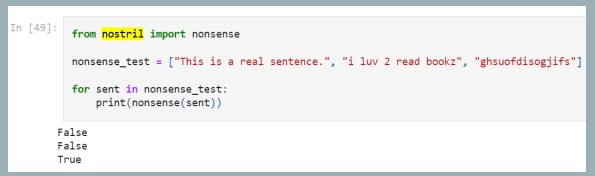
- Had to load whole files in order to join reviews with metadata
- Tried using multiple JNBs, chunksize, manually calling garbage collection – nothing worked ☹
- Solution: sample 5000 shuffled reviews per genre through Command Line (40000 total)

DATA PREPARATION

- Merged review data with metadata
- Eliminated unnecessary info
- Removed empty text rows
- Removed non-English books
- Remove nonsense text with NOSTRIL? (No)
- Added sentiment/POS tags with NLTK



series	country_code	language_code	popular_shelves
[]	US		[{'count': '100', 'name': 'to-read'}, {'count'...



```
In [49]:
from nostril import nonsense

nonsense_test = ["This is a real sentence.", "i luv 2 read bookz", "ghsuofdisogjifs"]

for sent in nonsense_test:
    print(nonsense(sent))

False
False
True
```

FINAL DATA

- 28274 total reviews from 8 genres
- Reviews of 17774 different books

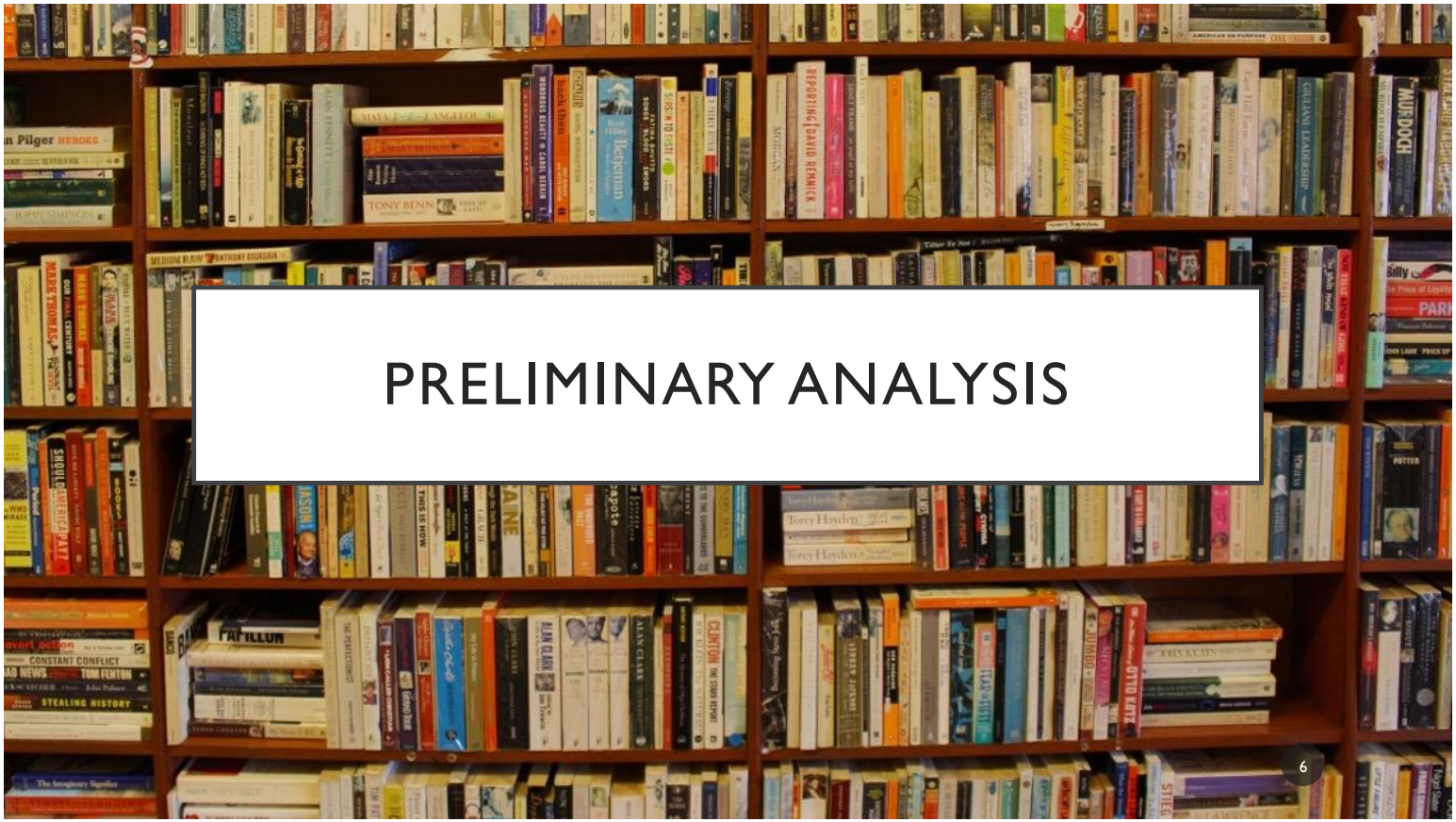
```
In [15]: total_df.Category.value_counts()

Out[15]: ya                        4334
         fantasy_paranormal        4323
         romance                   3918
         mystery_thriller_crime    3789
         comics_graphic            3505
         history_bio               3362
         children                  2858
         poetry                    2185
         Name: Category, dtype: int64
```

Top 15 Most-Reviewed Books

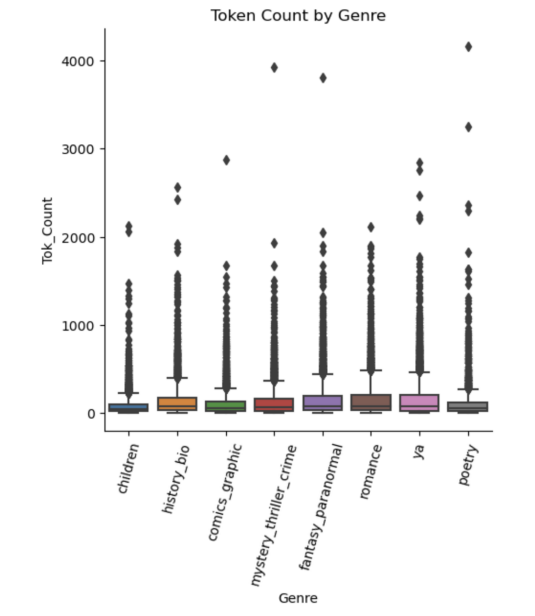
```
In [18]: total_df.Title.value_counts()[:15]

Out[18]: Milk and Honey                                                                  113
         Hamlet                                                                           50
         The Giver (The Giver, #1)                                                        50
         The Hunger Games (The Hunger Games, #1)                                          49
         Cinder (The Lunar Chronicles, #1)                                                49
         The Girl on the Train                                                            47
         Brown Girl Dreaming                                                              44
         Wonder (Wonder #1)                                                               43
         Miss Peregrine's Home for Peculiar Children (Miss Peregrine's Peculiar Children, #1)  42
         Divergent (Divergent, #1)                                                        40
         Where the Sidewalk Ends                                                          40
         Gone Girl                                                                        37
         City of Bones (The Mortal Instruments, #1)                                       37
         Throne of Glass (Throne of Glass, #1)                                            37
         The Fault in Our Stars                                                           35
         Name: Title, dtype: int64
```

Final DataFrame Columns

```
df.columns = ['Text', 'Rating', 'Title', 'Author', 'Genre', 'Lang',
       'Pages', 'Avg_Rating', 'Ratings_Count', 'Toks',
       'Toks_Low', 'Tok_Count', 'Avg_Word_Len', 'Sents_Count',
       'Avg_Sent_Len', 'Nonsense', 'Sentiment_Num', 'Sentiment_Tag', 'Adjs',
       'Adjs_Count']
```


PRELIMINARY ANALYSIS

TOKEN COUNT



ANOVA Results:

- F-Value: 62.71
- p-Value: 5.45e-90

Observations:

- Shortest Reviews: Children's Literature (45), Comics/Graphic Novels (52), Poetry (53)
- Longest Reviews: Romance (80), History/Biography (75)
- Reflective of reviewed text length?

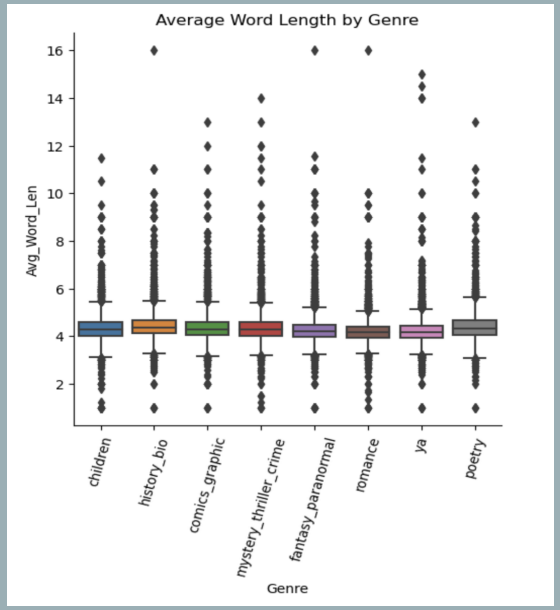
AVERAGE WORD LENGTH

ANOVA Results:

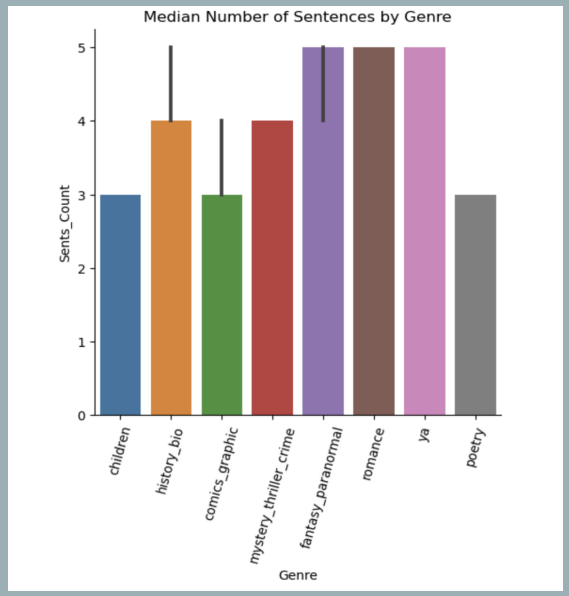
- F-Value: NaN
- p-Value: NaN

Observations:

- Not much variation overall
- Highest Avg. Word Length: History/Biography (4.46) and Poetry (4.45)
- Lowest Avg. Word Length: Romance (4.21) and Young Adult (4.23)
- Reflective of vocabulary level?



SENTENCE COUNT



ANOVA Results:

- F-Value: 89.83
- p-Value: 4.50e-130

Observations:

- Fewest sentences same 3 genres as fewest tokens (Children's, Comics, Poetry)
- Most sentences some different genres than most tokens (Romance, YA, Fantasy/Paranormal)
- Indicates History/Biography may have longer sentences (many tokens, fewer sentences)

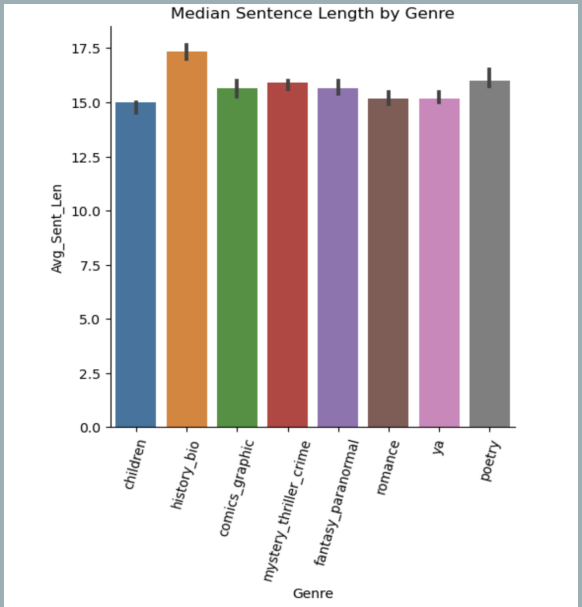
SENTENCE LENGTH

ANOVA Results:

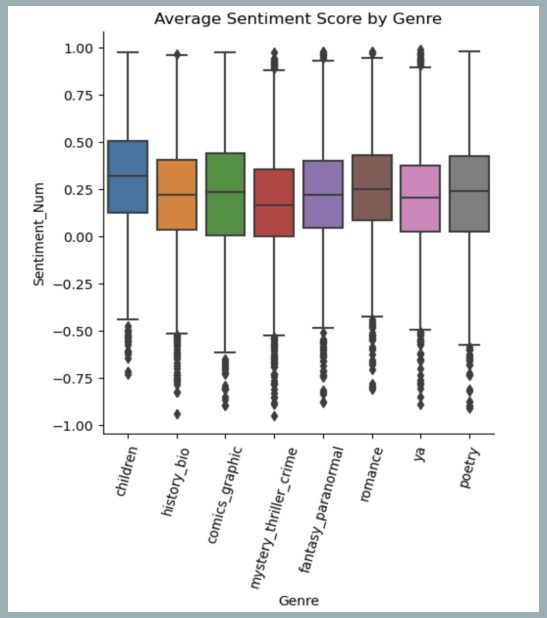
- F-Value: 21.57
- p-Value: 3.03e-29

Observations:

- Correct! History/Biography reviews = highest sentence length (17.33 tokens)
- Lowest sentence lengths: Children's Literature (15), Romance (15.16), Young Adult (15.17)
- Reflective of syntactic complexity / age of readers?



SENTIMENT



ANOVA Results:

- F-Value: 65.67
- p-Value: 2.27e-94

Observations:

- Background: Based on NLTK's sentiment tagger (#>0 = positive, #<0 = negative)
- Most Positive: Children's Literature (0.32) (2nd highest star ratings)
- Most Negative: Mystery/Thriller/Crime (0.17) (lowest star ratings)

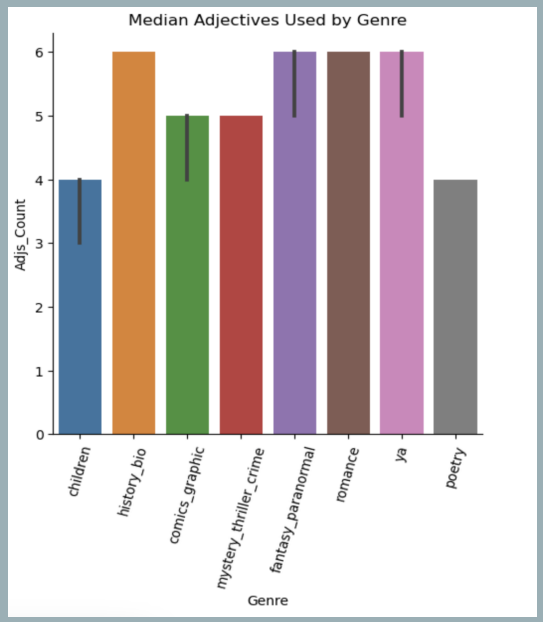
ADJECTIVES

ANOVA Results:

- F-Value: 49.12
- p-Value: 7.13e-70

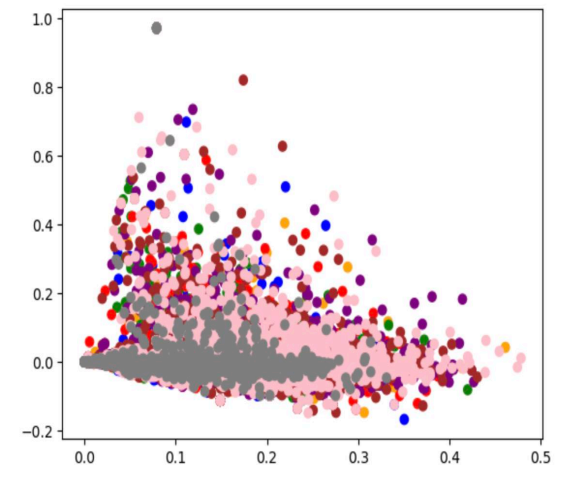
Observations:

- Lowest Median Adjective Use: Children's Literature and Poetry
- Maximum Adjective Counts: Fantasy/ Paranormal (290), and Poetry (254)




CLUSTERING!!!

CLUSTERING




```
Topic 0:
story like just really characters didn time way don know
Topic 1:
book reading great enjoyed amazing wait recommend second review written
Topic 2:
love books heart story romance fall beautiful absolutely just poems
Topic 3:
read great books fun ve quick easy wait best time
Topic 4:
loved wait absolutely story characters amazing great ending beautiful heart
Topic 5:
stars review rating come actual gave half giving liked given
Topic 6:
series books favorite wait characters great far best new reading
Topic 7:
good pretty really ending story bad liked mystery stuff bit
```

```
cdict = {'children':'blue', 'history_bio':'orange', 'comics_graphic':'green', 'mystery_thriller_crime':'red',
        'fantasy_paranormal':'purple', 'romance':'brown', 'ya':'pink', 'poetry':'gray'}
```

CONCLUSIONS

SIGNIFICANT FINDINGS

- Mystery and romance/poetry most distinct vocabulary
- History/biography has unique features as primary nonfiction genre
- Sentence count, sentiment, and token count too 3 most significant features
- Overall, evidence to suggest relationship between linguistic style of books and linguistic style of their reviewers

FUTURE ANALYSIS

- Multinomial Naïve-Bayes on whole text (TF-IDF, Numeric Features)
- Multinomial Naïve-Bayes on only adjectives (JJ)
- Adjust clustering parameters
- Replace NTLK POS tags with spaCy

CITATIONS

Fine-Grained Spoiler Detection from Large-Scale Review Corpora (Wan et al., ACL 2019)

Hutto, C.J. & Gilbert, E.E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

Mendhakar, A. (2022). Linguistic profiling of text genres: An exploration of fictional vs. non-fictional texts. *Information*, *13*(8), 357. https://doi.org/10.3390/info13080357

Mengting Wan and Julian McAuley. (2018). Item recommendation on monotonic behavior chains. In Proceedings of the 12th ACM Conference on Recommender Systems (RecSys '18). Association for Computing Machinery, New York, NY, USA, 86–94. https://doi.org/10.1145/3240323.3240369

Noorda, R., & Berens, K. I. (2021). *Immersive media and books 2020: Consumer behavior and experience with multiple media forms.* Panorama Project. https://drive.google.com/drive/folders/10DlCPSvcVnHEIcAsD7pJJRc_Tsf3Fodu