Investigating Concept Drift in Reading Trends during the COVID Pandemic:

A Novel Project

Preston Harry

Leora Rosenberg

Sabrina Sheu

Kathleen Young

December 5, 2020

NYU Courant Center for Data Science

DS-GA 1001: Introduction to Data Science

Business Understanding

In the midst of the COVID pandemic, as people change their habits, hobbies, and lifestyles, their preferences and tastes likely change in turn. The goods they consume, the media they observe, and the way they spend their days may no longer align with their pre-COVID habits. With COVID leading to potentially large upheavals in preferences, existing heuristics and models may no longer apply to the people they attempt to understand. To examine the possibility of concept drift within the pandemic, we focus on a particular case—change in reading habits. In such difficult and isolating times, a reader may tend towards joyful fiction to cheer themselves up or philosophy to really work out that existential dread. Back in September, a *Washington Post* article noted that in the beginning of the pandemic, adult nonfiction sales decreased while children's books sales increased (Merry, Johnson 2020). In May, sales in audiobooks and e-books leapt, and in June, books about race became more popular coinciding with the Black Lives Matter movement. On the other hand, as mentioned in *Observer,* publishing companies have pushed back titles intended for spring and summer in the hopes of releasing them in a better climate (Harvey 2020). Several other articles have also noted similar behaviors.

For companies like *Goodreads*, *Amazon*, and others that likely rely on recommender systems to suggest books to users, understanding how reading trends change in response to the COVID pandemic can be helpful in identifying whether or not models need to be retrained or entirely rethought during unexpected and difficult events. The relevance of this analysis depends on the cost of investing in it, the size of the effect, the cost of redeveloping a model, and the marginal benefit in improving that model.

However, the results of this paper speak less towards business concerns and more towards a broader understanding of social and behavioral impacts of the COVID pandemic.[1] In particular, this paper focuses on the COVID pandemic's impact on reading habits through the lens of concept drift. Additionally, conclusions here may be extended to other areas as well—if people are reading different books during a pandemic, they may also watch different movies, consume different internet content, or listen to different music.

To capture a shift in reading habits, we use book review data from *Goodreads.com* to predict the number of reviews a given book receives and, by extension, whether or not that book is being read at the time of the review. First, we train a model to predict the number of reviews a book receives in a time period just before COVID in late 2019. Second, we observe how the performance of that fitted model changes when predicting the number of reviews during COVID time periods. A decrease in the performance of the model will provide evidence towards concept drift in terms of people's reading habits.

Data Understanding

To build a model to predict readership we use data from *Goodreads*, a social networking site which maintains an extensive book database from which its users can maintain reading lists and share their reactions to books. Unlike public library data which only shows what is currently checked out, *Goodreads* contains historical data including

[1] Brian approved a non-business problem by email on 9/25, saying that we had to demonstrate only that the model would bring value to an imagined set of stakeholders.

past reviews and when those reviews were made. Note that *Goodreads* also allows us to observe people reading any book, rather than best-seller lists which only reference recent releases' sales.

However, our data can still contain selection bias. All readers do not use *Goodreads*, and it is reasonable to think that people who read more are more likely to use the site. As such, we do not capture data about readers who have begun to read recently, and it is these new readers that could affect our models the most.

In order to collect data from *Goodreads*, we built a distributed web scraper, which we ran on Amazon Web Services (AWS). The scraper simultaneously built two databases: the first database consisted of a randomized sample of reviews, which represented users' interactions with specific books, and the second database contained details on books for which we had scraped five or more reviews. Refer to Appendix A for full details on the contents of each database.

Goodreads data is collected through a distributed web scraper in which several smaller scripts collect and communicate data via a REST API to be aggregated by a main script (Appendix B). Distributing the scraper with AWS allows multiple scripts to run simultaneously to increase the rate at which data is pulled (Appendix C). In particular, our distributed system ran ten instances of scripts to query *Goodreads* for data and one instance of a script to aggregate those data. This sped up data collection approximately six-fold, ultimately capturing about 1 million review observations and 16,000 book observations. See Appendix D for more detail.

In order to get genre data for each book scraped from *Goodreads*, we use data from *Open Library*, an open source database of general book information. Using a python

package to query Open Library's API, we collect both "genre" and "subject" information. Genre information includes expected terms such as "fiction" or "fantasy," whereas subject information can range from terms like "fiction" to more complicated terms like "Hogwarts School of Witchcraft and Wizardry (Imaginary organization)." Since subject data is far more represented than genre information, we combine genre and subject terms in one subject vector for each book.

Data Preparation

Our target variable is the number of book reviews in both pre-COVID and COVID time periods where book reviews specifically refer to an interaction with a book by a *Goodreads* user. For most books, there are relatively fewer recorded reviews whereas some books tend to be very popular with an extremely high number of reviews (Figure 1).[2] Note that these trends occur in each month within the data as well.

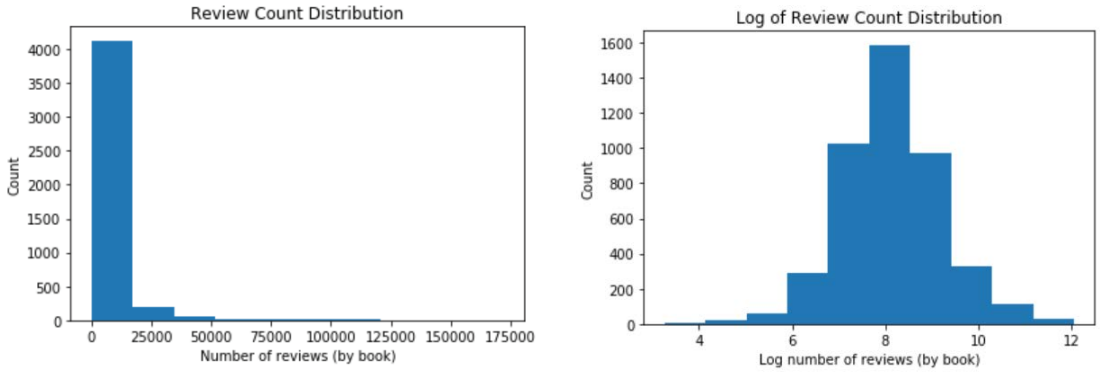


Figure 1: Distributions of number of reviews and log of number of reviews by book

[2] We considered the possibility that because monthly review counts were so heavily skewed to the left, our target values would require additional transformations before being fit to a linear model. However, because early tests showed that logarithmic transformations had no impact on model performance, we reverted back to a linear approach.

With both our scraped *Goodreads* data and associated *Open Library* data, we format each for use in our models and merge them into one dataset.

The *Goodreads* data generally did not require additional cleaning beyond the process used to parse *Goodreads* web pages. However, author data was inconsistent for books with multiple contributors past the author. For example, *Harry Potter* books are associated with "J.K. Rowling" and "J.K. Rowling Olly Moss (Illustrator)" among other author values. For any book with multiple author values, we use the shortest author value, excluding some unusually short author values. We then transformed each unique author into binary features. Additional binary features created from categorical features include book language and book series. Refer to Appendix A for more information.

For subject information from *Open Library*, for each subject term in each book, we removed non-alphabetic characters and stemmed each term using a Lancaster stemming algorithm. Doing so consolidates subject vectors by removing any numeric terms and combining similar terms—terms like “child” and “children” are treated as the same word for the sake of subjects, for instance. Using count vectorization, we then simultaneously removed stop words from each subject vector and converted each unique subject term to individual binary features. Lastly, any subject that occurs in less than three books are dropped to reduce the total number of features from 3,265 to 1,225. These subject features are then merged to the cleaned *Goodreads* data on each book’s unique ISBN number for use in the fully specified model. Of each unique ISBN, approximately 45% contained subject information with Open Library.

Modeling & Evaluation

We build a model using a linear regression methodology for three reasons. First, our target variable is the number of reviews per book, a continuous rather than a binary value. Second, linear regression outputs interpretable results which provide insight into the reading landscape over time. Third, with tens of thousands of potential values for authors, series, and subjects, each of which would be transformed into many features, linear regression allows for an implicit feature selection process. To improve the interpretability of the model output, we normalize the features as part of the regression modeling process. As such, we use the coefficients generated by each model as the primary indicator as to what books people chose to read in each period and how those selections changed between periods.

To gain an initial understanding of reading trends during COVID, we construct a baseline linear regression model to understand the relationship between our features and the target variable. Our process is as follows:

1. Train a linear regression model on a subset of book-level data from March 15, 2019 to June 15, 2019 using the following features: total number of reviews, total number of ratings, and average rating. Test the model on the remaining subset of book-level data from the same period.
2. Test the model again on a dataset from March 15, 2020 to June 15, 2020.
3. Compare performance for the two time periods. A decrease in model performance offers evidence for concept drift regarding reading habits during COVID.

The date ranges for each test set cover the same months to account for possible seasonal effects associated with March through June. This date range also captures the effects of the early pandemic when habits and preferences are probably more likely to shift. Table 1 shows results of the baseline model.

Test data date range	MSE (Test)	R^2 (Train)	R^2 (Test)
March 15, 2019 – June 15, 2019	27.477	0.276	0.197
March 15, 2020 – June 15, 2020	83.748	--	0.131

Table 1: Baseline model MSE and R^2

Relative to the pre-COVID test data, R^2 values for test data during COVID decrease from 0.197 to 0.131. This does reveal a decrease in the model's effectiveness during COVID. While this may offer some small evidence towards concept drift resulting from the pandemic, the initial R^2 value is fairly low. These results question the effectiveness of this baseline model, and it becomes difficult to make claims about the pandemic's relationship to reading trends. The proportionally small decrease in R^2 may have been due to an expected amount of concept drift that would exist independent of any substantial event, or it could have simply been due to random chance.

The above model trains and tests on a fairly wide range of months. If readership across months differs considerably, the model may have difficulty predicting number reviews over such a wide range. To accommodate this issue, we separate the test data into monthly groups. In particular, we train a model using a book-level data from 2018 to 2019, test that model separately on each month of 2020 data from January to September, and compare the outcome of each month.

We simultaneously build this monthly baseline model alongside a fully specified "kitchen sink" model including additional features. Table 2 displays features in each model type.

Feature Group	Baseline Model	Kitchen Sink	Feature Type	Feature Count (Total)	Feature Count (Sparsity Filter)
2018-2019 Scraped Monthly Review Counts	YES	YES	Numerical	24	24
Number of Reviews	YES	YES	Numerical	1	1
Number of Ratings	YES	YES	Numerical	1	1
Average Rating	YES	YES	Numerical	1	1
Book Language	NO	YES	Binary	22	36
Book Series	NO	YES	Binary	3,966	789
Book Author (Clean)	NO	YES	Binary	1,463	7,035
Book Subject	NO	YES	Binary	1,223	1,218

Table 2: Feature type and inclusion by model

Feature Selection R^2 results	Split 1	Split 2	Split 3	Split 4	Split 5
Baseline	0.604	0.654	0.660	0.652	0.709
Baseline & Language	0.603	0.654	0.659	0.651	0.708
Baseline	0.587	0.645	0.679	0.695	0.656
Baseline & Series	0.577	0.631	0.666	0.678	0.642
Baseline	0.684	0.664	0.553	0.684	0.690
Baseline & Author (Raw)	0.665	0.658	0.541	0.672	0.675
Baseline	0.666	0.605	0.664	0.636	0.578
Baseline & Author (Clean)	0.641	0.589	0.648	0.617	0.569
Baseline	0.675	0.676	0.639	0.694	0.599
Baseline & Subject	0.652	0.653	0.627	0.667	0.584

Table 3: R^2 values for baseline models and models including additional features on different train-test splits

For the monthly baseline model, we first train on 75% of books in 2018 and 2019 and test on the remaining 25% of books in those years. Table 3 shows mean R^2 values across all months for different samples of train-test splits. We use R^2 values as the metric for the for the feature selection process to more easily compare among models built on different splits. If adding features had *improved* model performance, adjusted R^2 may have been a better metric, but given that adding features degraded model performance, the out of the box R^2 seems a fair metric. Each "baseline" row represents mean R^2 values for 5 randomly sampled train-test splits on the monthly baseline. We find notably improved R^2 values of about 0.55 to 0.7 relative to the initial baseline.

With a now better performing model, we then test separately each feature type's impact on performance. For each "baseline" row in Table 3, we compare the mean R^2 values of the baseline to the baseline with an additional feature type on the same train-test split samples. Rather than iterating over every potential combination of features to determine feature importance which would have been infeasible with over 6,000 features, we tested features by adding like features in groups (author features, series features, etc.). For each feature type, relative to the baseline, we observe a very slight decrease in mean R^2 values. Though the decreases are not large enough to suggest that any one model is definitively better than another, if included together, the fully specified model's performance may suffer.

As such, we choose additional features to the monthly baseline model based on explanatory power of each feature. Language does not seem like it would explain large scale change in concept drift associated with COVID. The magnitude of series coefficients was fairly small and did not seem to impact the target variable much.

Similarly, subject features had low magnitudes and there appeared to be a decent amount of noise in the subjects themselves with vague features like “causes,” “manipulation,” or “trove.” Authors, however, appeared to have both larger and more interpretable coefficients. Ultimately, to preserve performance as much as possible while seeking some additional explanatory power, our final model includes all features included in the monthly baseline as well as the (cleaned) author features.[3]

Target Month	Regression Type	Alpha	MSE (Test)	R^2 (Train)	R^2 (Test)
01/2020	Ridge	1.0	2.961	0.801	0.774
02/2020	Ridge	1.0	1.785	0.724	0.685
03/2020	Ridge	1.0	1.978	0.719	0.657
04/2020	Ridge	1.0	3.500	0.777	0.688
05/2020	Ridge	1.0	5.775	0.728	0.574
06/2020	Ridge	1.0	5.912	0.655	0.465
07/2020	Ridge	1.585	4.301	0.644	0.509
08/2020	Ridge	1.0	2.804	0.656	0.55
09/2020	Ridge	2.0	2.32	0.598	0.494

Table 4: Final model (baseline + author features) output by month

To choose appropriate hyperparameters, for each month, we used the 75-25 train-test split to build twenty-three models: one simple linear regression and both ridge-regularized and lasso-regularized models built for eleven potential alpha values.[4] The model with the lowest MSE on test data would be chosen as the model for that time

[3] We include the clean author rather than the raw values both because that view corresponded with less degradation and because we felt it better reflected the true author of each book.

[4] In early tests, we used the integers from one to ten as alpha values in order to cover a broad range of amount of regularization. However, we never observed an alpha value larger than three being selected, so we shifted to the following values: 1.0, 1.585, 2.0, 2.322, 2.585, 2.807, 3.0, 3.17, 3.322, 5.0, 10.0.

period. Table 4 shows for each month's model, the selected hyperparameters, the mean squared error, and the R^2 values for both training and test sets. Figure 2 shows the change in R^2 across each month.

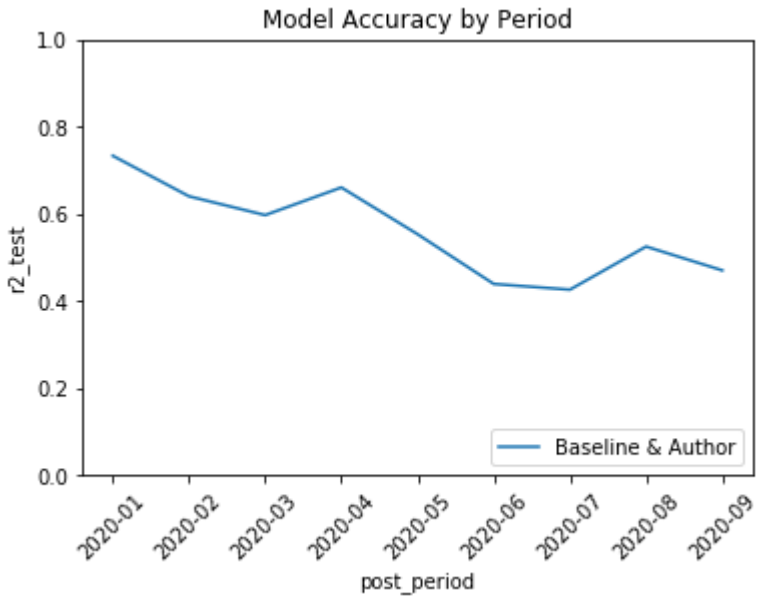


Figure 2: R^2 by month for final model (baseline + author features) in 2020

As expected, the R^2 values of the models did decline in most months, but the decline did not appear to accelerate in the spring of 2020, indicating that while later models suffered from concept drift, the pandemic had not added more concept drift than normal. We also observe a spike in mean squared error in May and June. To gain a better understanding of what factors may be driving concept drift in 2020, we look towards the features with the largest coefficient magnitude.[5] The most prevalent trends were consistent with the Black Lives Matter movement. Table 5 shows coefficients associated with features related to racial issues. For each author, the highest coefficient (highlighted) coincides with 06/2020, the peak of the Black Lives Matter protests.

[5] We did not consider the p-values associated with each feature, primarily because we did not realize until late in the modeling process that scikit-learn does not output p-values. Because models were selected based on a cross validation process, we have assumed that their coefficients are meaningful irrespective of true statistical significance.

Ultimately, though this model seeks to quantify the social and behavioral impacts of the pandemic on reading behavior, it does not indicate that the pandemic has drastically upset reading behavior. However, the model does illuminate some possible explanations for changes in reading trends throughout 2020 by highlighting the impact of particular authors on reading trends during the summer of the Black Lives Matter movement.

Author Feature	01/2020	02/2020	03/2020	04/2020	05/2020	06/2020	07/2020	08/2020	09/2020
Angie Thomas	-0.507	-1.821	-1.222	-2.030	-0.463	1.917	0.796	-0.070	-0.705
Chimamanda Ngozi Adichie	-0.540	0.017	-0.027	-0.072	0.355	0.756	0.012	-0.200	0.280
Jeanine Cummins	7.947	6.350	5.099	6.537	7.220	10.212	7.740	6.339	3.595
Toni Morrison	0.709	0.095	-0.360	0.098	-0.418	1.703	0.321	0.523	-0.134
Ta-Nehisi Coates	0.499	-0.410	0.336	-0.624	-0.600	1.407	-0.506	-0.397	-0.053
Zora Neale Hurston	-1.215	-0.615	0.346	0.832	-0.382	2.358	0.455	-0.057	0.644
Ibram X. Kendi	1.765	0.513	0.234	-0.121	2.577	18.610	7.138	5.242	1.204
Trevor Noah	-0.325	1.129	0.191	1.023	-0.523	5.157	2.396	0.302	0.969

Table 5: Model coefficients for select author features

Deployment

In most cases, our model does not need to be deployed in a long-term manner. Our model may be expanded upon, however, for research purposes to understand how COVID has impacted reading habits and to contribute to a greater literature on the social impacts of pandemics. To do so, the distributed scraping system would need to be expanded to gather more data over a wider time period. Using data from periods even earlier than 2018 would allow for a difference-in-difference style model which would allow us to test whether or not any reduction in model accuracy during COVID meets an expected standard amount of concept drift or if the COVID period really does correspond

to especially large concept drift. Continuing to collect data into the future would also allow us to test whether or not model accuracy starts to improve again as COVID begins to subside—if accuracy "bounces back," any reductions in accuracy during the COVID period may indeed be due to the pressures of the pandemic.

Note that there exist some risks with expanding the distributed scraping system to collect more data. Maintaining the scraper on AWS costs money and could become expensive to maintain over long periods. Furthermore, when monitoring data gathering, any testing (notably hypothesis testing) should be done after a predetermined amount of data has been gathered. Otherwise, one risks spuriously confirming a potentially incorrect hypothesis during the data gathering process.

Additionally, a better source of content information for each book would likely be necessary. *Open Library*, being open source, is unreliable and inconsistent with books' subject information. A more complete source with standardized subject information, would serve to improve understanding of how certain types of books shift during the pandemic.

Ultimately, though there exist few ethical risks with using publicly available data to make conclusions about large scale social trends, there exist some risks when attempting to extend any conclusions to a general population. Goodreads book review data represents only people who enjoy publicizing their reading habits and thoughts on books. As such, any conclusions may not apply to other readers.

Bibliography

Harvey, Gillian. “What You’ll Be Reading in Fall 2020 and Beyond, According to Publishing Industry Insiders.” *Observer*, July 26, 2020. https://observer.com/2020/07/2020-fall-book-trends-publishing-industry-barnes-and-noble-strand-bookstore/

Merry, Stephanie, and Steven Johnson. “What the Country Is Reading during the Pandemic: Dystopias, Social Justice and Steamy Romance.” *Washington Post*, September 2, 2020. https://www.washingtonpost.com/entertainment/books/2020-book-trends/2020/09/02/6a835caa-e863-11ea-bc79-834454439a44_story.html.

Appendix

Contribution

Entirely building web-scraper and collecting Goodreads data – Leora

Setting up Amazon Web Services distributed system with scraper – Kathleen

Collecting Open Library merging and merging datasets – Preston

Modeling research – Sabrina

Feature selection – Leora; Preston

Model testing, coding, fine-tuning – Kathleen; Leora

General topic research – Sabrina

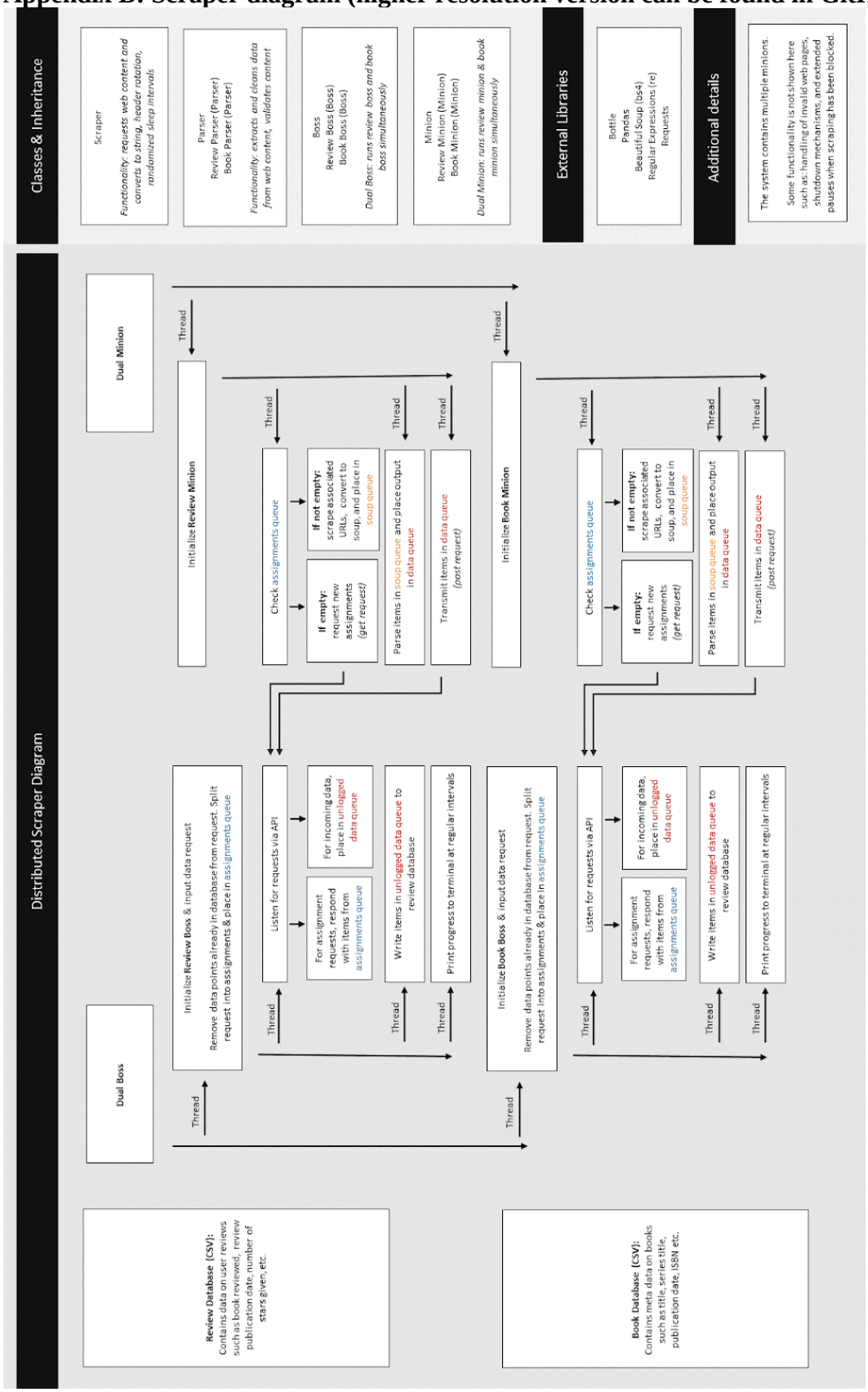
Paper writing and editing – All

Appendix A: Scraped Database Contents

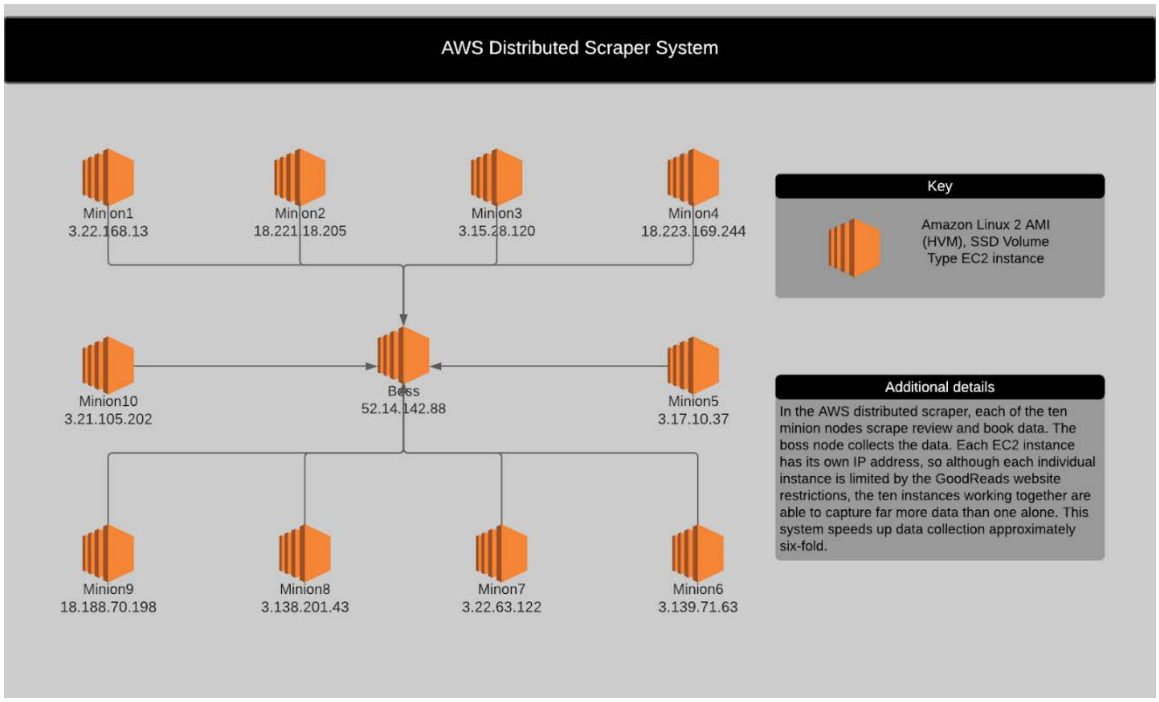
Database	Field	Details
Review	Review ID	A unique identifier for each review, found in its URL.
Review	URL Validity	Because the scraper generated potential URLs at random, it visited many pages which did not contain valid reviews. We hypothesize that these are associated with reviews written by users who subsequently deleted their Goodreads accounts.
Review	Rating	The review's numerical rating (1-5) of the book being reviewed.
Review	Reviewer URL	The URL associated with the user who wrote each review, containing both a unique identifier and a username.
Review	Book Title	The title of the book being reviewed.
Review	Publication Date	The date associated with a user's most recent interaction with the reviewed book, generally the date at which they marked the book as complete and submitted a numerical rating or written review.
Review	Other Dates	We also collected the dates at which users added the book to their shelf (i.e., added it to a reading list), started reading each book, and completed the book. Most users did not appear to populate these fields.
Review & Book	Book ID	A unique identifier, specific to *Goodreads*, associated with each book. This field was used to match observations from the review and book databases.
Book	Book Author	The author or authors of each book. This field required additional cleaning as part of the feature engineering process.
Book	Language	The language in which each book was published. It was common for books to have multiple editions in different languages.
Book	Number of Ratings	The total number of numerical ratings that a given book had received on *Goodreads* at the time of scraping. Because users generally submit a rating when they have completed a book, this value represents the best estimate of the total number of times a book has been read.
Book	Number of Reviews	The total number of written reviews that a given book had received on Goodreads at the time of scraping.
Book	Average Rating	The average numerical rating that a given book had received at the time of scraping.
Book	ISBN	An industry-standard unique identifier associated with each book. This field was used to match observations from the scraped data and *Open Library*.

Book	Editions URL	A web page which listed other editions of a given book.
Book	First Publication Date	The publication date for the book's first edition.
Book	Publication Date	The publication date for the given book.
Book	Series	The series (i.e., *Harry Potter*) to which a book belongs

Appendix B: Scraper diagram (higher resolution version can be found in GitHub)



Appendix C: Distributed Scraper System



Appendix D: Discussion of Scraped Database Size

In total, our scraped databases consist of about 1.4 million review observations and 16,000 book observations. From these, we remove invalid reviews, duplicate observations, observations from outside our time period of interest, and all observations associated with books which were not present in both databases (ie, reviews for books for which we had not scraped book data). This process results in about 800,000 review observations and 16,000 book observations.

We then aggregate the data to show the number of monthly scraped review observations per book from January of 2018 through September of 2020. While the exact distribution of monthly reviews varied by month, the overall shape was generally heavily skewed to the right as over 50% of books had no scraped reviews in any given month. Please see key statistics on the number of monthly reviews per book below. We show

statistics only from the first six months of 2020; these months are a fair representation of the distribution throughout the thirty-three month time period.

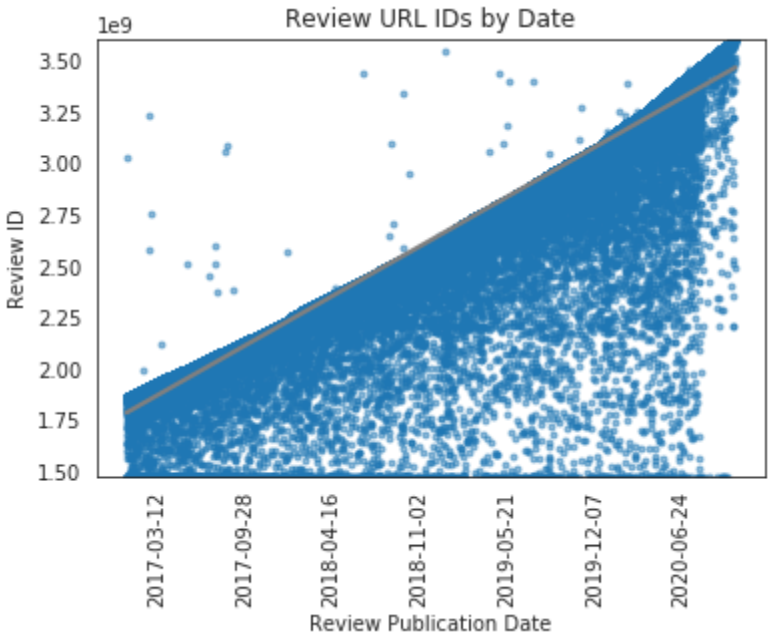
Reviews Per Book	**01/2020**	**02/2020**	**03/2020**	**04/2020**	**05/2020**	**06/2020**
Mean	1.2	0.82	0.88	1.1	1.19	1.1
25th Percentile	0	0	0	0	0	0
50th Percentile	0	0	0	0	0	0
75th Percentile	1	1	1	1	1	1
Max	102	57	80	122	93	79

Appendix E: Discussion of Time Periods & Scraping Mechanics

To scrape reviews which were published in the thirty-three month period from January 2018 through September 2020, we first identify URLS of reviews which were published in that time period. We conduct that analysis in two stages:

First, we conduct a preliminary scrape of roughly 6,000 reviews. Analyzing the relationship between the unique identifier in each review URL (review ID) and the date at which the corresponding review was published, we observe that the two seemed to be roughly correlated. Based on the minimum review ID associated with 2017, we then scrape a larger database of reviews published in the period starting January 2017 onward.

We then use the larger dataset to conduct a more rigorous analysis. As a first step, we confirm that among review IDs associated with valid reviews, the review ID and the review publication date (as an ordinal value) is almost perfectly correlated ($r = 0.97$, $p = 0.0$). Furthermore, we observe that the majority of the error in the correlation consisted of reviews which were published later than would be predicted by a linear regression; these reviews would still fall into the time period of interest.



We then train a decision tree model to classify reviews based on whether they fell into the time period of interest (2018-2020), limiting the depth of the decision tree in order to identify a single cutoff. Subsequently, we set our scraper to only scrape reviews whose review IDs fell above that cutoff value.

Code used in this project can be found in a GitHub repository here.