Analyzing Book Checkout Trends at the Seattle Public Library

Sarvesh Fotedar, Ekrem Kizilkaya, Abbas Shaikh, Cody VanZandt, Andy Wang

Contents

1 Introduction

From classic literature to contemporary bestsellers, the Seattle Public Library's book collection offers a rich window into its surrounding community's literary tastes and interests. By analyzing checkout trends and book movement, we can uncover which novels and authors are most beloved among library patrons, giving us remarkable insight into the community's unique cultural and social context.

This report conducts a preliminary examination of approximately 41 million checkouts from the Seattle Public Library system between 2005 and 2022. We intend to combine this information with Goodreads review data from the UCSD Book Graph to answer general questions about the contours of contemporary American readership and even perform specific analyses on the competitive dynamics of the book authorship and publishing industry.

From this point forward, the following questions to guide our exploration and report:

1. Why do some books enjoy widespread acclaim while others fizzle?
2. Do popular books explode onto the scene or accumulate readers more gradually?
3. How do the popular success and critical acclaim of novels differ?
4. How has recent conglomeration of publishers changed the literary marketplace?
5. And, ultimately, to what degree can the success or failure of a book be predicted?

From the most popular genres and titles to the unique factors influencing readers' choices, book checkout data offers a rich window into the complex relationship between readers, libraries, and the larger cultural context. In this paper, we will examine the book checkout trends at the Seattle Public Library in detail, drawing on both statistical and literary analysis to shed light on the unique reading culture of this vibrant community.

2 Datasets

2.1 Seattle Public Library

Our primary dataset comes directly from the Seattle Public Library and has approximately twelve million rows for book checkouts spanning the five years between 2018 and 2022. Specifically, each row corresponds to a monthly count of checkouts for the physical or electronic version of an item. Items are not just limited to books; the Seattle Public Library also makes video discs, e-books, and sound discs available to patrons, to name a few. Each row is described by eleven variables, which are described in Table 1.

Variable	**Description**
UsageClass	Whether the item was physical or digital
CheckoutType	The tool or vendor that was used for checkout
MaterialType	The item type (ex: book, video disc, etc)
CheckoutMonth	The four digit checkout year
CheckoutYear	The month of checkout
Checkouts	The number of times that the item was checked out within the checkout month
Title	The full title and subtitle
Creator	The author or entity responsible for creating
Subjects	The subjects as they appear in the library catalog
Publisher	The publisher of the title
PublicationYear	The year that the item was published, printed, or copyrighted

Table 1: Seattle Public Library Variables

2.2 UCSD Book Graph

Our supplementary dataset comes from the UCSD Book Graph initiative and contains over fifteen million reviews for approximately two million books from 465,000 users. Each row of the dataset is a JSON object which represents a single review and its associated metadata. The nine attributes of each JSONified review are described in Table 2

Attribute	**Description**
user_id	The UUID of the reviewing user
book_id	The unique numerical id for the reviewed book
review_id	The UUID of the review itself
rating	The numerical rating of the book out of 5
review_text	The text content of the review
date_added	The date the review was published
date_updated	The date the review was updated, if applicable
n_votes	The number of votes endorsing the review
n_comments	The number of comments for the review

Table 2: UCSD Book Review Object Attributes

3 Exploratory Analysis

3.1 Total Checkouts per Month

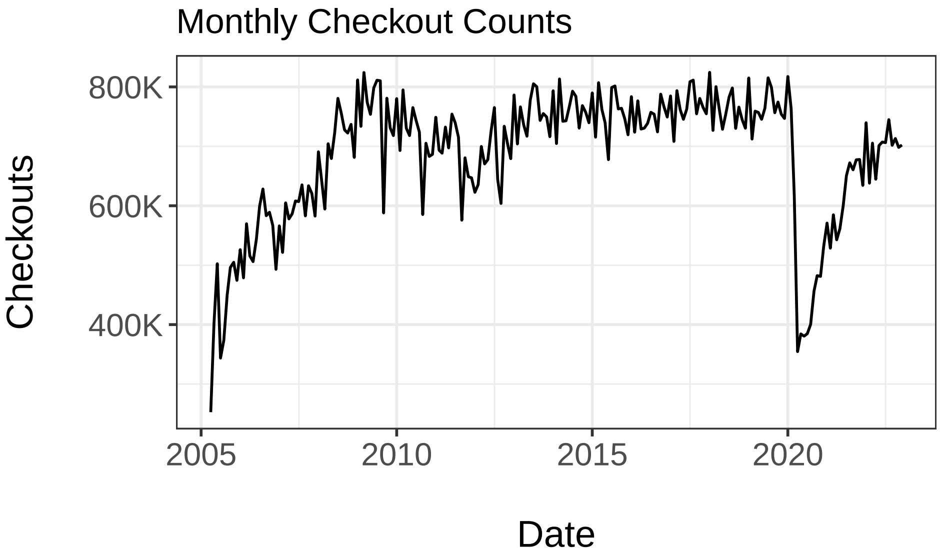

Figure 1: Total Checkouts per Month

Figure 1 shows the change in the total number of checkouts per month over the entire period from 2005 to 2022. The initial increase in checkouts from around 2005 to 2009 is most likely due to the adoption of the online checkout system which enabled data collection. There is also a steep drop in checkouts during 2020, which we are certain is due to the COVID-19 pandemic. Since then, the plot shows that checkouts have been recovering quickly and are almost at pre-pandemic numbers.

3.2 Total Checkouts per Month, by Medium

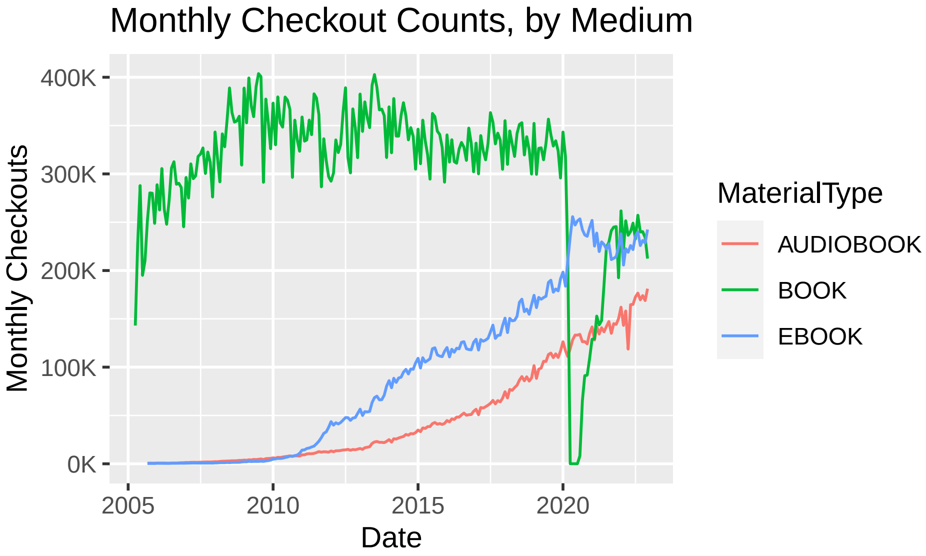

Figure 2: Total Checkouts per Month

Figure 2 is similar to Figure 1, but is broken down by the three most popular mediums: physical books, audiobooks, and e-books. The variation in checkout counts of physical books is almost identical to the pattern exhibited in Figure 1, including the steep drop in 2020. Interestingly, the checkouts of audiobooks and e-books were completely unaffected by the pandemic, most likely due to the ability to check them out online without physically visiting the library.

3.3 Relative Popularity of Mediums

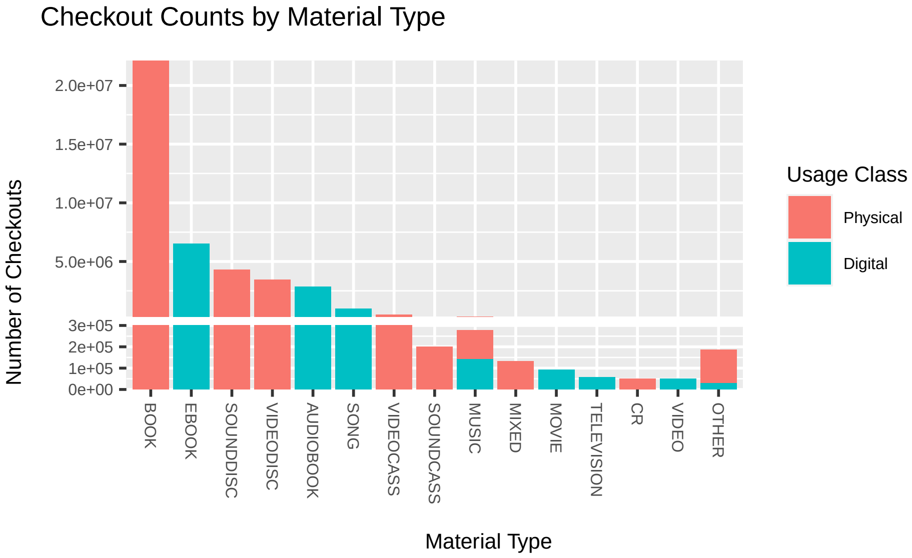

Figure 3: Checkout Counts by Medium

Figure 3 shows the relative popularity of the types of items that the Seattle Public Library makes available for checkout. Physical books are by far the highest, the checkouts of the next highest two categories (ebooks and sound discs) being around a quarter of the checkouts of physical books. This explains why in Figure 2, the pattern of physical book checkouts follows the total number so closely.

3.4 Genre Popularity Among Physical Books

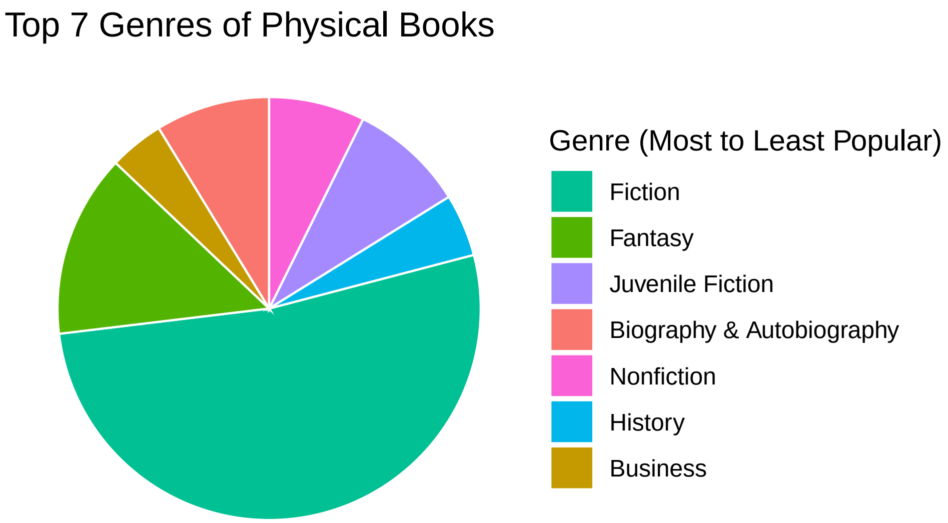

Figure 4: Genre Popularity

Figure 4 shows the relative popularity of the genres of physical books at the Seattle Public Library. Fiction is overwhelmingly popular, accounting for the genre of over 50% of the books that were checked out over the period.

4 Data-Driven Insights

4.1 Publisher Conglomeration

We aim to explore how the conglomeration of the publishing industry affects the literary marketplace, particularly in the microcosm of the Seattle Public Library. As such, we investigate the number of checkouts attributed to the top 5 percent of publishers every month by number of checkouts. We hypothesize that his metric will represent conglomeration within the publishing industry since a small portion of publishers controlling a disproportionately large part of the industry indicates greater conglomeration and less diversification of publishers used by the library. Figure 5 displays the number of checkouts per month accounted for from the top 5 percent of publishers overall, as well as separated by book, e-book, and audiobook publishers. What we find is that the proportion of checkouts controlled by top book publishers has been steadily decreasing, indicating a possible greater diversification in the publishers for physical books, whereas the number of checkouts for audio books and e-books from top publishers has increased consistently to around 90 percent currently. This suggests that large conglomerate publishers have grown to control the digital publishing industry as digitization of library content has become more common and more profitable in recent years.

```
## ‘summarise()‘ has grouped output by ’Date’.  You can override using the## ‘.groups‘
argument.
```

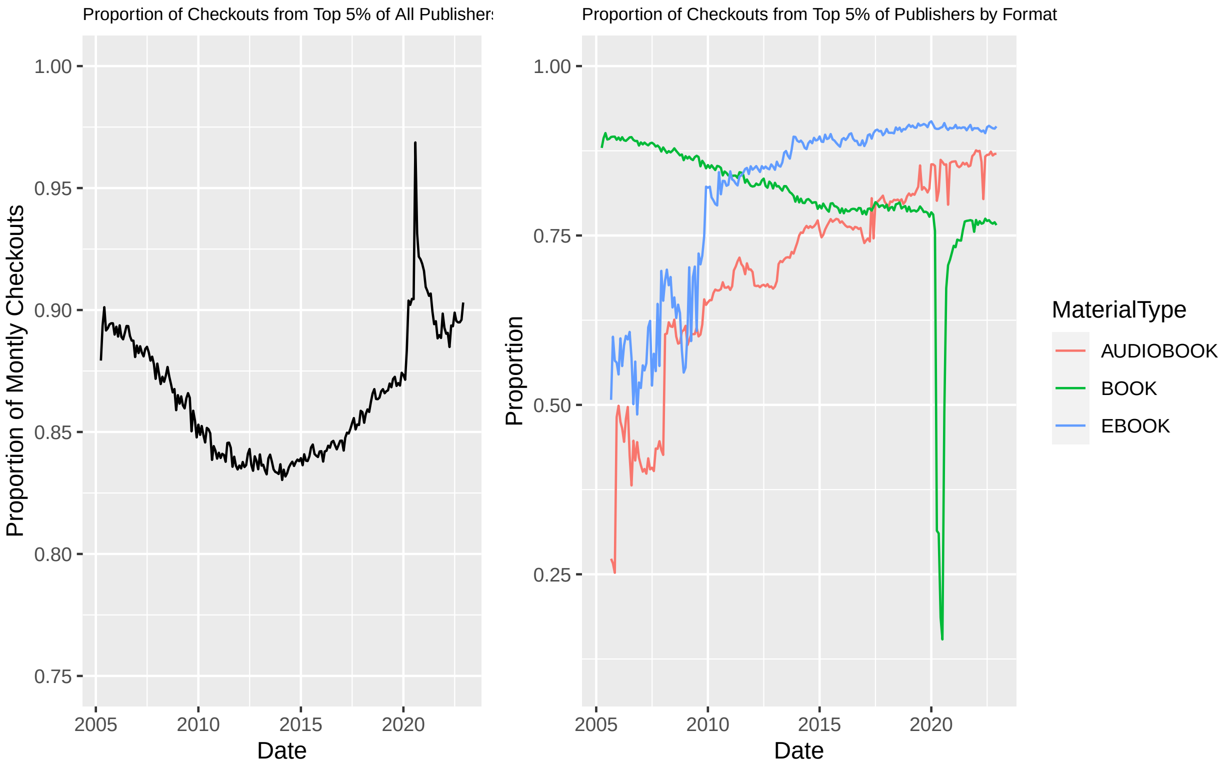

Figure 5: Measure of Publisher Conglomeration

5 Killer Plot

Our killer plot is a novel visualization that provides insight into both the popularity and sentiment of a particular novel over time. At its most fundamental level, the killer plot shows a time series of the number of monthly checkouts of a particular book at the Seattle Public Library over a given range of months. At each month within this range, we display a specific word that is most distinctive among all words in the reviews of the novel from this month. We calculate the distinctiveness of this word using a weighted log odds metric. The size of the word is relative to its distinctiveness and the color corresponds to the rating of the review (i.e. green words correspond to 5-star reviews, yellow to 3 or 4-star reviews, and red to 1 or 2-star reviews).

Below, figures 6 and 7 are two examples of our killer plot for Suzanne Collins' *Catching Fire* and Paula Hawkins' *The Girl on the Train: A Novel*, respectively.

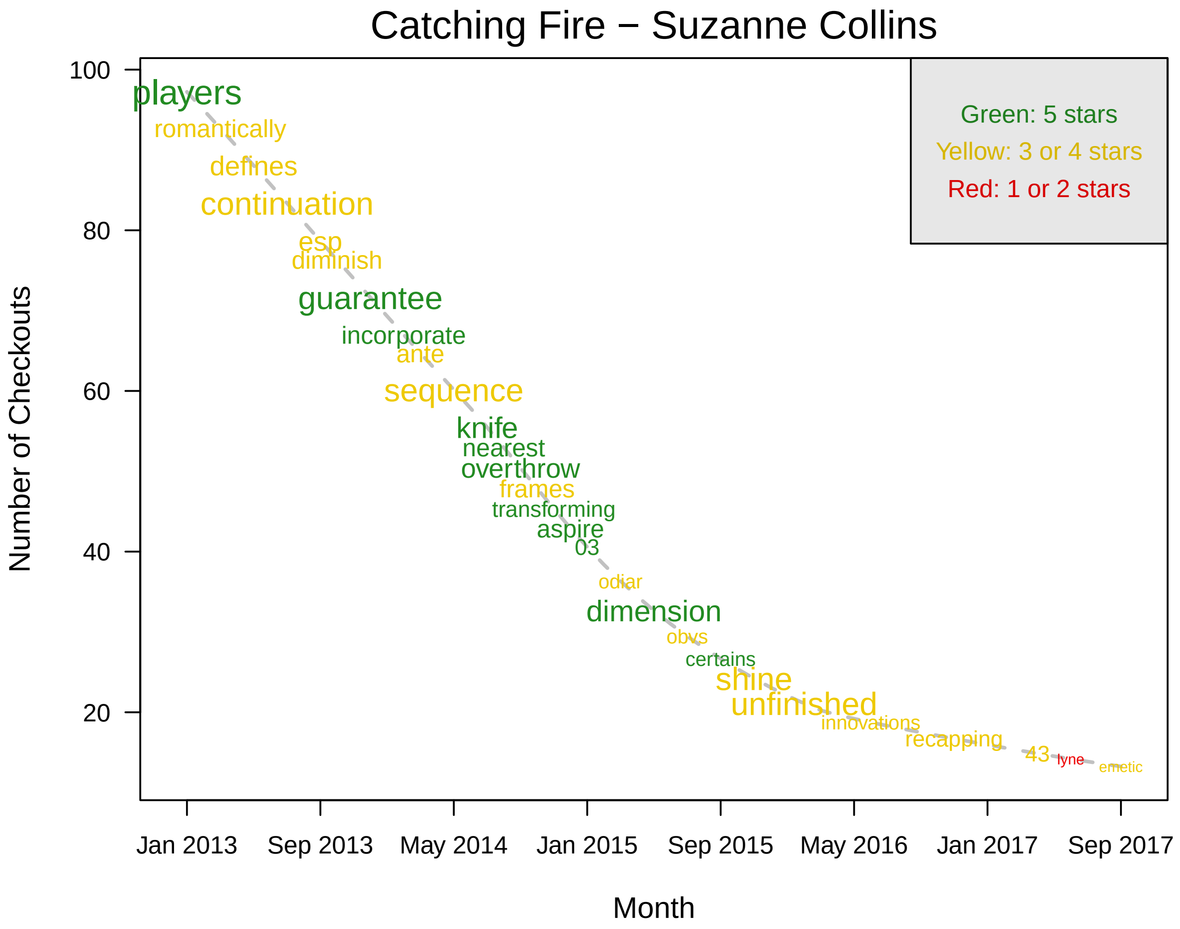

Figure 6: Killer Plot for Catching Fire by Suzanne Collins

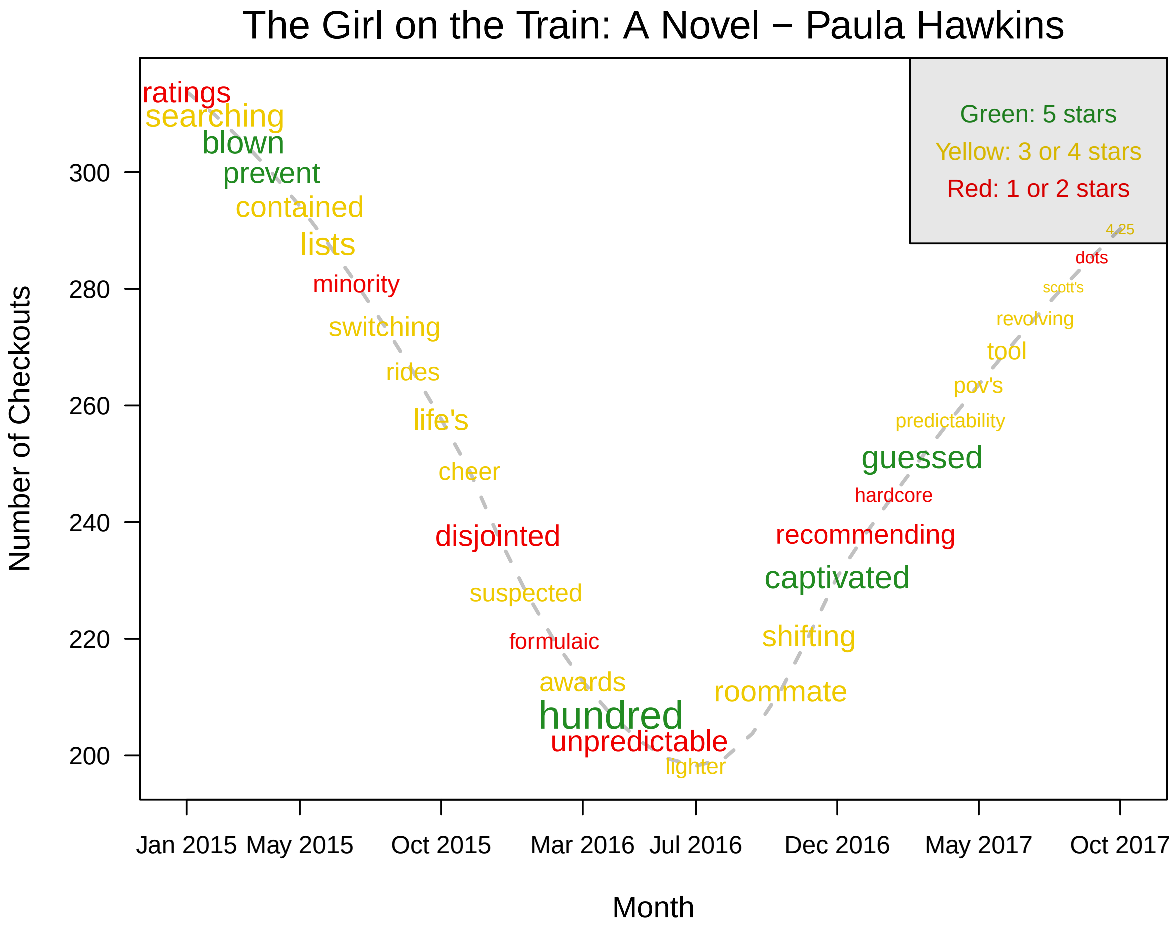

Figure 7: Killer Plot for The Girl on the Train by Paula Hawkins

With this representation, we hope to provide a unified representation of both the popularity of the novel in terms of checkouts from the library, but also the larger public sentiment towards the novel as is revealed through the review data. If, for example, a book initially achieves widespread popularity and then experiences another later resurgence, this plot allows us to visualize and interpret the sentiment surrounding its initial reception as well as its later resurgence. This can be particularly useful in understanding potential reasons behind why a particular novel may be experiencing a revival in the literary marketplace.

6 Conclusion

6.1 Discussion of Results

Do popular books explode onto the scene or accumulate readers more gradually?

How do the popular success and critical acclaim of novels differ?

How has recent conglomeration of publishers changed the literary marketplace?

6.2 Limitations

6.3 Future Improvements

7 Appendix

7.1 Fuzzy Title Matching

To make best use of our UCSD Book Graph Goodreads data, we need to match unique book IDs from Goodreads reviews to book checkouts from the Seattle Public Library. The Seattle data, however, is messy. Author information is often encoded in the title field, which is itself often full of misspellings and other typographical oddities. Dataset alignment, then, is a challenging problem in its own. As a first attempt, we employ a fuzzy string matching algorithm that computes pairwise similarity between the Goodreads and Seattle book titles and reports a match when that similarity crosses a given threshold. Our current similarity measure is Jaro-Winkler, which we selected primarily because it weighs matches at the beginning of a string more heavily than matches towards the end. Given the erroneous additions frequently appended to the end of Seattle library book titles, Jaro-Winkler ought to perform well. After parameter optimization, we report that approximately thirty percent of Seattle titles have been matched to Goodreads book IDs. With more nuanced title preprocessing and multivariable matching across author and publisher, we anticipate that the match rate could approach 50 percent.

7.2 Popularity Curve Clustering

Beyond string matching, we also report some initial research into time series dimensionality reduction and clustering. In considering book popularity over time, it is natural to ask if there are certain popularity patterns that recur across books, publishers, and genres. If these common patterns can be identified, then perhaps they can be predicted. After consulting the literature on time series clustering, we present a pipeline that we believe could identify clusters of books that exhibit similar popularity patterns. This pipeline computes monthly popularity time series curves and transforms them through z-normalization, discrete cosine transform, dynamic time warping, and k-medoids clustering.

Z-normalization makes the shape of popularity curves comparable across books with dramatically different raw checkout numbers. Discrete cosine transform reduces the dimensionality by decomposing and recomposing the time series using a smaller number of cosine waves. Dynamic time warping defines a similarity score for popularity curves with potentially non-overlapping time domains. And finally, k-medoids clustering modififes the more familiar k-medoids clustering by replacing average-computed centroids with median-computed centroids. K-medoids is a less common choice than k-means, so the selection is worth commenting upon. Compared with k-means, k-medoids is less sensitive to outliers and replaces the creation of artificial average-based centroids with median-based ones This sidesteps the need to average different popularity curves together during centroid creation – a questionable practice to be sure.

We hope that this unsupervised machine learning pipeline will, once implemented, make possible a variety of supervised and predictive modeling tasks.