ISyE 6740 – Spring 2025

Project Proposal

**Team Member Names:** Jeff Triemer

**Project Title:** The Evolving Faces of Genre Through Time

Contents

[Problem Statement 2](#_Toc191043544)

[Methodology 2](#_Toc191043545)

[Data source 3](#_Toc191043546)

[Evaluation and Final Results 3](#_Toc191043547)

[Citations 4](#_Toc191043548)

# Problem Statement

At the point of sale, frequently books are either sold or left behind based on what the cover communicated to the reader and whether or not it seemed to fit their preferences. In such a crowded and competitive marketplace, we must assume that publishers and authors make every effort to advertise their product as effectively as possible. Book covers are a crucial tool in the publishers’ arsenal to do just that. To that end, I wonder what publishers have done over time to design their book covers in this aim. My expectation is that covers have been optimized to quickly communicate to readers a given subject or genre for the book in question. As time has passed and publishers have become more competitive, has the cream separated from the crop? Have publishers refined their approach and settled on a clear and consistent approach? Do some publishers appear to have a different approach from others? Or alternatively, have books led efforts to differentiate themselves from their competition, leading to increasingly diverse bookshelves by genre?

My suspicion is that all publishers alike have converged on very similar approaches to book covers to effectively communicate their substance. In fact, this appears to have been backed up by some researchers. Some researchers have effectively used information from the covers of books to effectively predict their genre, reporting a 94% top 3 accuracy, when combining book title and cover data- 59% using just the book cover imagery alone (Kjartansson & Ashavsky, 2017). In this paper, we are proposing a technique that will utilize dimensionality reduction, image clustering, and simple linear regression to attempt to determine how book covers have become more or less similar over time and use some analysis and visualization to interpret why or how that might be.

# Methodology

We will use OpenLibrary’s API to pull covers by their release years and then apply our own initial preprocessing, namely RGB normalization and image padding (to keep the aspect ratio unaltered). Next, we will apply dimensionality reduction to the images, either t-SNE or UMAP as PCA will likely wash out some context that may be important to our use case. Then we will be ready to apply the spectral clustering.

First, we will determine the number of clusters using some exploratory analysis and the number of genres. Then we will use an example book cover sample to show how effective spectral clustering is as a predictor of the book’s genre classification. Note that the 59% accuracy referenced from the research (Kjartansson & Ashavsky, 2017) includes some logic to extract title from a book’s cover- we will likely miss this context in our simplified approach and find a worse performance, but this will show the overall fit of our clusters by genre. Instead, we will compare the clustering to some random estimation of genre. This example may exclude any temporal split so that we can first understand the imagery cluster’s relationships with book genre, and we can plot a graph of covers for visual exploratory data analysis.

Then we will fit new clusters based on some temporal data, to do this I will first do some exploratory analysis into what temporal data would be the best fit for this use case. It may be that we group the data by decade, pre-determined publishing “eras” or just year, but this will require some inspection of the data at hand to see determine the best approach.

Next, once we have our clusters over time, we can compare and contrast their results. First, while typically applied more in k-means, I’d like to find observe our Within Cluster sum of Squares (WCSS) to determine whether or publishers have become precise in their book cover designs over time. My hypothesis is that publishers have become more “conformist” or targeted in their cover designs, and we will see more compact convergence of clusters over time. That said, these results may rely heavily on the quality of our data, dimensionality reduction, and overall cluster fit to book genre.

I believe the best approach, particularly for its interpretability, may be to use simple linear regression with published years or decade as an independent variable and WCSS as the dependent variable. Once I check the assumptions of the linear regression and verify the overall fit, I can use the beta of our temporal feature and its value to interpret the magnitude, direction, and significance of the relationship between time and the WCSS of our temporal book cover clusters. Finally, as an added benefit, I plan to plot these clustering graph results over time to visually represent how the book cover clusters have changed over time with a few examples.

# Data source

I will be pulling a sample of books and their covers from OpenLibrary’s API. This will include features like title, release date (either by edition or original book release), average and number of reviews, genre, subject(s), description text, and finally of course the cover image

Our data quality will be crucial to the success of the unsupervised learning we intend to apply, and further the final simple linear regression. Unfortunately, using this source there will be some work to pull book covers by their individual release date. This source contains data for both the overall book and further each potential edition for a given book. Each book and even edition may have multiple possible covers, but it appears release dates are at the best level of granularity at the edition level. I have already done the work to set up a pull for the image of each cover for a sampled book, but some additional work will be required to find a reliable way to pull the release date for as many covers as possible.

It is also worth noting that we will have titles and book descriptions from this data. We could add to our clustering approach and add the data from the language in the title and book descriptions, which may further improve the clustering as it has in the aforementioned research. This will be a lever to pull if we find our initial example clustering lackluster or are just interested in taking the additional step. This will alter the eventual evaluation of the results, but it should all the same indicate how publishers have refined this aspect of their marketing approach over time.

# Evaluation and Final Results

As mentioned briefly in the proposed methodology, we will have a number of outputs from our process that we can use to interpret and evaluate our results. First, we have the benchmark of random genre assignment to use in comparison with our clustering results. We can use a confusion matrix here for easy interpretation. Next, we can visually interpret the clusters to explore how the clusters change over time and see if any particular cluster or genre has an interesting dynamic over time. If we add modeling based on the text data, we will also evaluate the results of the NLP approach applied.

Finally, we will use the outputs of our simple linear regression to first determine if it’s a good fit of the data, it could be that simple linear regression is not the most appropriate form of regression. Once we have a decent fit, we can interpret the results to reject the null hypothesis that there is no significant relationship between the published date and the fit of our clusters.

# Citations

Kjartansson, S., & Ashavsky, A. (2017). *Can you Judge a Book by its Cover?* https://cs231n.stanford.edu/. https://cs231n.stanford.edu/reports/2017/pdfs/814.pdf