Eric Randall
Predict 452: Section 55
Assignment 1: Automated Data Acquisition

Reissuing out of print albums can be a good way to reintroduce a musician's body of work to a new generation of music listeners. Album reissues can be a solid source of revenue for record labels. Vinyl sales made up 9% of the music industry sales in 2017 - 14.32 million units, a record-setting amount, and the top selling record of 2017 was a 50th anniversary remaster & reissue. The challenge set forth by management is identifying well regarded “underground” records as a record label publishing obscure records from decades past, especially in a world where tastemakers and influencers are no longer coming from a single point of origin such as music publications or the radio, but millions of voices on the internet. How can we harness the voices on the internet in order to make a more informed decision on what to consider for reissue?

For this project we scraped data that was freely available, but in unstructured form, on the internet. The data comes from popular user-rating website RateYourMusic.com. On the website, users rate albums on a scale 1-5 in half point increments. The ratings are then aggregated into an overall score for each album. Releases are browsable in various methods - by artist, label, in lists by users, often of which following a specific theme, by genre, and by year, and are sortable by rating among other things. Therefore a user can easily find the top albums of a genre, or by year. I found specific interest in the top albums by year, as it provides a linear view of musical history. The data was acquired using a Python script that made use of the Scrapy package, an open sourced web crawling and scraping platform. The script was run late at night in order to be less of a burden on the website’s servers. Instead of using a headless browser to navigate the website in order to scrape, the script takes advantage of the structured nature of the URLs, and therefore goes directly to each URL and scrapes directly.

The dataset contains 8 fields. First, the year of the record's release (*year*), and the ranking of the Album within that year of release (*ranking*). Ranking is some algorithm that measures user ratings versus number of ratings. A release with a higher overall rating, but with less reviews in total will rate lower than a release slightly lower in rating but with more reviews. This helps rule out the possibility of a few overzealous fans highly rating an obscure album, giving it a perfect score and launching it to the top of the charts. However, for our use, an undiscovered jewel of a record may be one that is highly rated but with relatively few real thought out reviews in general. We scraped the first 4 pages of each year's ranking since 1960. There are self-explanatory fields *artist* and *album* (Album title), *genre* which is assigned by users and the overall average rating (1-5, *rating*). The dataset also has the number of ratings (*ratings)* and the number of reviews (*reviews*) - ratings are from when a user ranked an album 1-5 whereas a review is a user submitted written review of the album that may contain a description of the album or the album's background.

The most reviews for the albums in the dataset come in 2009 - but the reviews tend to be higher as the albums get older, peaking in 1978. The charts lean heavily towards Progressive Rock, with 246 albums, followed by Film Score, Hard Bop, Singer/Songwriter, Modern Classical and Avant-Garde Jazz as the top 6. Exploring the dataset, we see that most of the lowest-rated albums in the dataset are from 2018. The highest *rated* album in the set is surprisingly a film score by Bruce Broughton, with a 4.38 average score but only 17 ratings, followed up by classical piece The Marriage of Figaro and the highly and often rated 1963 Jazz album The Black Saint and the Sinner Lady by Mingus.

We cannot blindly use this dataset to serve up best sellers but it is a good starting point for research into vinyl reissues for the future of our firm.