Stage 2 Report

Team members: Niko Escanilla, Derek Hancock, Srujith Poondla
Team name: Data Vaders

I. Web Sources

For this project we decided to gather information on music albums. We gathered this information from the following two web sources:

1. Metacritic.com

Metacritic aggregates reviews of music from leading critics. They calculate a metascore, which shows how well an item has scored from critics. They have a sizeable collection of music albums on their website that we decided to extract.

2. Wikipedia.com

For each year, Wikipedia has a page that lists music albums released during that year along with some information about each album. By extracting the information from several years of Wikipedia pages, we can gather a sizeable amount of music album data.

II. How We Extracted Data

1. Metacritic.com

To extract structured music album data from Metacritic, we observed that each page contains around 200 albums. Inspecting the HTML structure helped us to find a common method to extract all albums data in a loop. Each album details are wrapped under * *and *class="product_wrap"*. Inside this tag there are few other tags to get more details such as *<a>* for url and title, * * and *class = "metascore"* for metascore and url is used to get more details on Label and released year. We iterate over multiple pages till we get 3000 albums. Then we pass through each album url and collect the label details by analyzing the HTML structure. We used BeautifulSoup to beautify the scrapped HTML structure and pandas to create the table.

2. Wikipedia.com

To extract structured music album data from Wikipedia, we observed that each page (i.e. each year) contains 12 tables, a table for each month of the year. Inspecting the HTML tags, we noticed that the 12 tables fall under the class called *wikitable*. In each *<table>* tag with the *wikitable* class, each tuple was placed in a *<tr>* tag, and each column value was placed in a *<td>* tag. In turn, we used BeautifulSoup and pandas to extract the data.

III. Entity and Table Description

As mentioned above, we decided to extract music albums, specifically those released in the last three years.

1. Table A - Metacritic

The data from Metacritic contains the album's name, artist, label (the company that published the album), release date, and a list of genres it belongs to. It also includes a meta score which is Metacritic's rating based on all of its critics' ratings. We also added an empty producer field because that is included in the wikipedia but not in metacritic. This allows both our sets of data to have the same schema.

2. Table B - Wikipedia

The Wikipedia data also contains the album's name, artist, label, release date, producer, and a list of genres it belongs to. We also added an empty metacritic field because it is included in the metacritic data but not in wikipedia. Thus our two tables have the same schema.

IV. Open Source Tools

We used pandas and Beautiful Soup to help extract the data from the pages. Beautiful Soup makes it easier to extract data based on its tag in an html page, and pandas helps with data manipulation and saving the data once it is extracted.