CS839 –Project Stage 2 Report
Extraction of Music from Structured Web Sources

Team Members:
- Sreyas krishna Natarajan (snatarajan6@wisc.edu)
- Sugadev Chellakkannu (chellakkannu@wisc.edu)
- Vignesh Thirunavukkarasu (thirunavukka@wisc.edu)

I. Web Data Sources:

a) BestBuy:
The first data source that we selected was BestBuy (www.bestbuy.com).
BestBuy is leading provider of technology products, services and solutions. It provides a wide variety of music albums sorted by artists and genre for users to buy and listen to.

b) Metacritic:
The second data source that we went with was Metacritic (www.metacritic.com). Metacritic aggregates critic reviews and provides metascores for music based on reviews. It provides a consolidated list of music albums sorted by artists and genre which we extracted.

II. Extraction of Structured Data:

To extract data, we first analysed the structure and format of the HTML of web pages of BestBuy and Metacritic. We decided to extract Title, Artist, Genre, Release, Rating fields. We used Scrapy open source to crawl the pages. XPath and CSS selectors were used to extract the relevant fields.

BestBuy:
Codebase can be found here:
https://github.com/sugadev/CS839/blob/master/Stage%202/Code/cs839/spiders/BestBuy_spider.py

Steps we performed to extract is this:

1. We hardcoded the base URL which we need to crawl.
2. The Listing pages of BestBuy listed the Music information and with all the relevant fields we needed to extract.
3. We then hard code our scraper to find and extract the relevant fields (Title, Artist, Genre, Release and Rating).
4. Then the information is stored in a CSV file.
5. Iterate through all the listings of listing page.
6. we then used Spider to iterate through the next pages.
7. Again, perform all the steps from 2-6.

Metacritic:
Codebase can be found here:
https://github.com/sugadev/CS839/blob/master/Stage%202/Code/cs839/spiders/Metacrytic_spider.py

We performed the same type of analysis and extraction of data as BestBuy except that the listing page of the website doesn't provide us any of the relevant fields.

1. we had to perform an additional operation of taking the link of each music in each listing page and parse the pages from the link extracted to extract the relevant fields.
2. The Genre field of Metacritic consisted of comma-separated attributes which we needed to iterate through the field to get the attribute values as a list.
3. Then iterate through all the music on the listing page.

III. Entity of choice:

The entity that we chose was music albums, majorly from genres – pop and rock.

a) Table A – BestBuy (bestbuy_music.csv): The table from BestBuy contains the music albums' title, artist, genre, release and rating**.**

b) Table B – Metacritic (metacritic_music.csv): The table from Metacritic contains the music albums' title, artist, genre, release and rating.

The sites apparently did not have the same schema and we had choose to select the above 5 columns: Title, artist, genre, release and rating; to get the same schema for both.

Tuples:
BestBuy – 4360
Metacritic – 4330
Schema:

Title	**:** name of the music album
Artist	**:** artist of the album
Genre	: conventional category that identifies some pieces of music
Release	**:** date of release of the album
Rating	**:** rating the album has received on that particular site.

IV. Open Source Tools:

We used the open source crawling tool named **Scrapy** on top of the **Python** ecosystem to crawl structured data from the two selected web sources.

Scrapy is an application framework for crawling web sites and extracting structured data which can be used for a wide range of useful applications, like data mining, information processing or historical archival. Even though Scrapy was originally designed for web scraping, it can also be used to extract data using APIs or as a general-purpose web crawler