CRAWLING AND EXTRACTING STRUCTURED DATA FROM WEB PAGES

DATA SCIENCE (CS838) PROJECT STAGE II

Rohit Kumar Sharma
rsharma@cs.wisc.edu

Arjun Balasubramanian
balarjun@cs.wisc.edu

Aditya Rungta
aditaker@cs.wisc.edu

April 13, 2019

Introduction

In this report, we describe how to perform Wrapper based Information extraction to extract structured data from websites. Specifically, we demonstrate how a wrapper can be used to parse semi-structured data such as HTML pages and transform it into a tabular format. For this, we have chosen to extract movie information from popular websites such as Internet Movie Database (IMDB) [1] and Metacritic [2].

Methodology

Web Sources

The entity type which we chose to analyse is *Movies*. We selected two sources with overlapping real-world entities as the information extracted will be used to study in the subsequent stages of the project.

Internet Movie Database or IMDB[1] is an online database of information related to films, television programs, home videos and video games, and internet streams, including cast, production crew and personnel biographies, plot summaries, trivia, and fan reviews and ratings.

Metacritic [2] is a website that aggregates reviews of media products: films, TV shows, music albums, video games, and formerly, books. For each product, the scores from each review are averaged.

Crawling and Extraction

We chose to extract best movies across different genres from both the websites. This will also ensure that we have some overlapping entities as required.

The open source tool that we used to extract data from HTML pages is Beautiful Soup [3]. Beautiful Soup is a Python library for pulling data out of HTML and XML files. It provides idiomatic ways to navigate, search, and modify the parse tree.

Results

Statistics of information extracted

Web Source	File name	Number of Records
IMDB	imdb.csv	3000
Metacritic	metacritic.csv	3000

Schema of the Tables

Attribute	Datatype	Description
ID	Integer	Primary Key
Name	String	The name of the movie
Year	Integer	The year in which the movie was released
Runtime	String	The duration of the movie in min
User Rating	Decimal	The average of user rating on respective website
Metascore	Integer	The score provided by Metacritic.com on a scale of 1 - 100
Director	String	The director of the movie
Actors	String	The star cast of the movie
Genre	String	The genre of the movie

Figure 1 shows the attributes highlighted that were extracted from IMDB [1]. Figures 2 and 3 show the same for Metacritic [2].

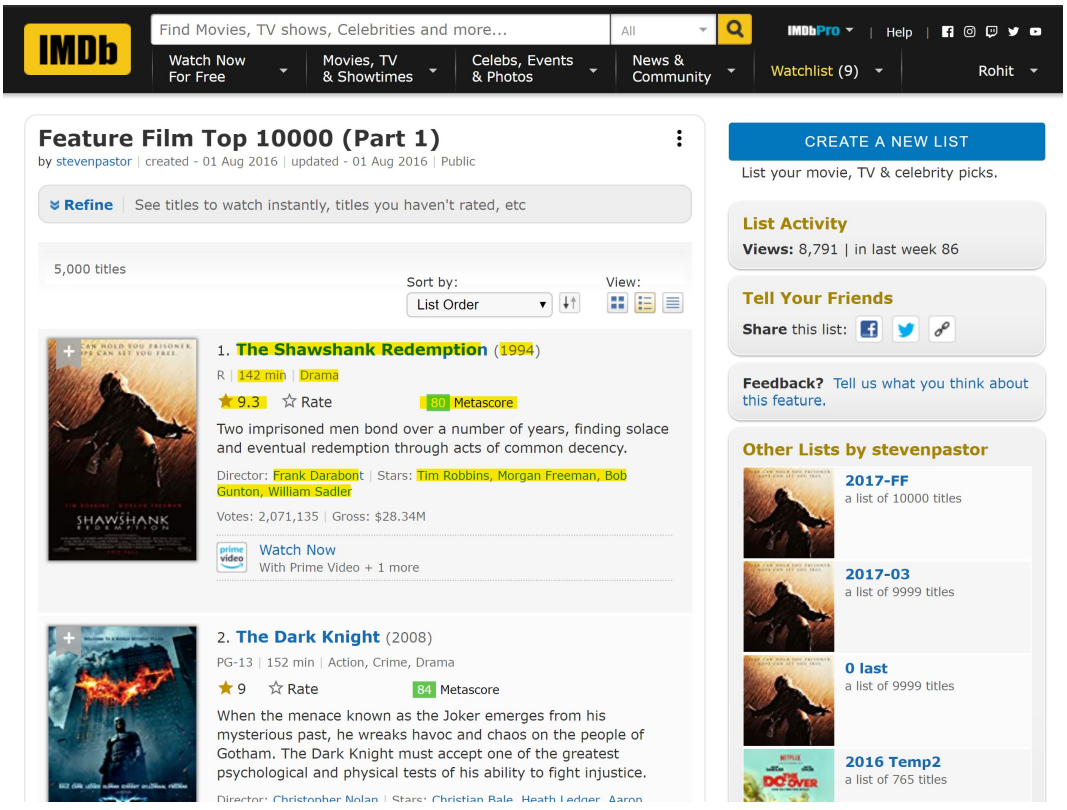

Figure 1: Attributes in IMDB

As visible in Figure 1, it is easy to extract all of the required information listed in the schema table. We inspected the source of the HTML page and identified common patterns of tags required to extract each of the attributes. We then proceeded to extract information using the Beautiful Soup library. One tricky case in extracting information from IMDB was to handle missing metacritic ratings. However, this became easy to handle since we had a thorough understanding of the wrapper structure for IMDB and Beautiful returned *null* when an attribute value was not present.

As visible in the below figures, to extract the movie information from metacritic.com, we had to navigate to the movie page as the information we needed was present on that page. We parsed the html of the movie page(Part 2) using beautifulSoup library in python to extract all the required attributes.

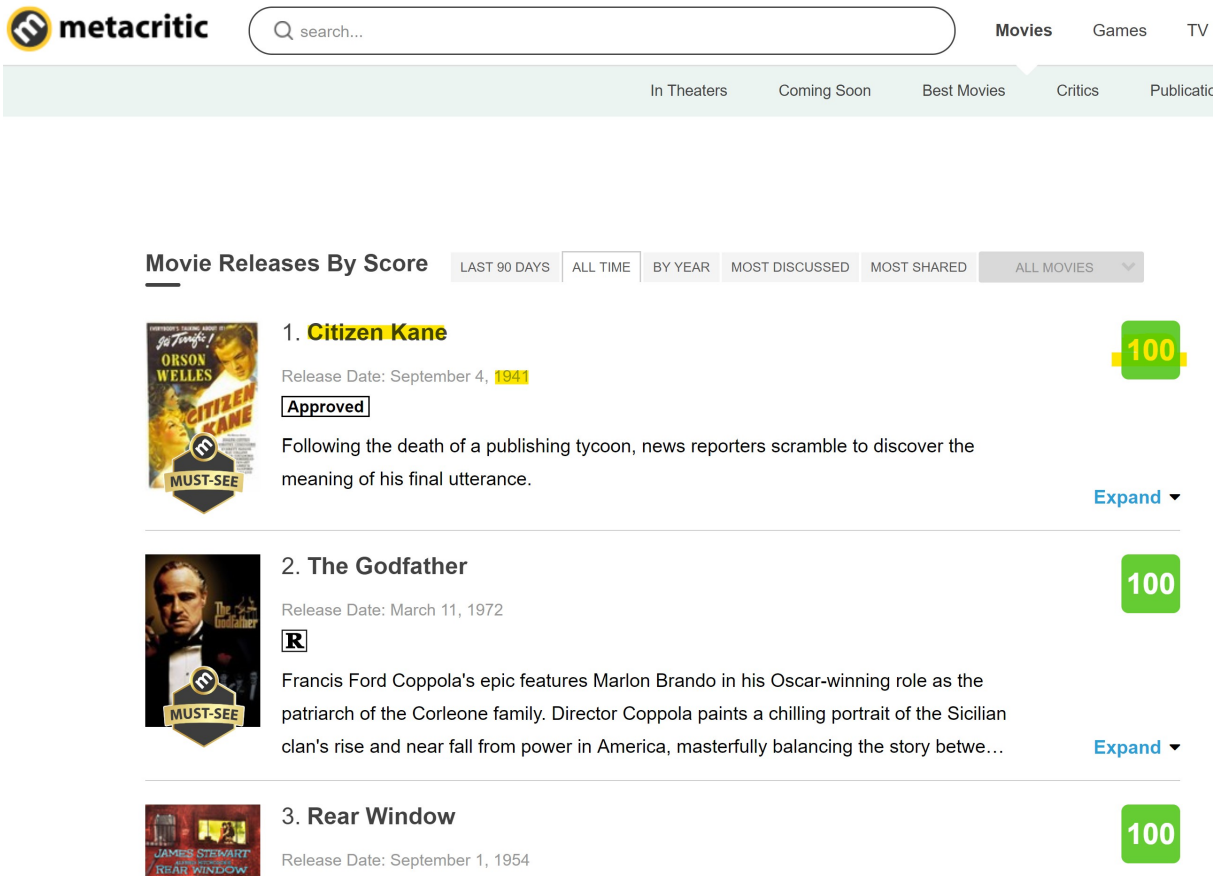

Figure 2: Attributes in Metacritic - Part 1

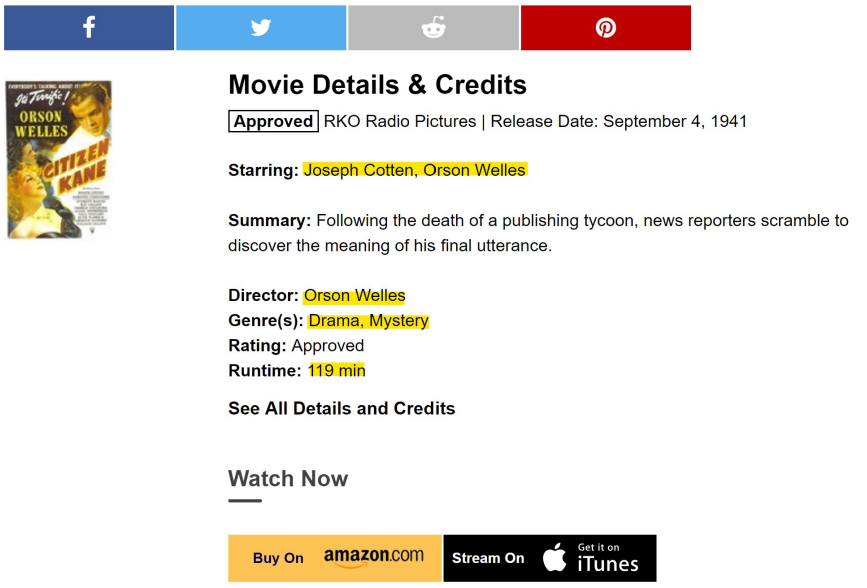

Figure 3: Attributes in Metacritic - Part 2

References

[1] Internet Movie Database. https://www.imdb.com/

[2] Metacritic. https://www.metacritic.com/

[3] Beautiful Soup. https://www.crummy.com/software/BeautifulSoup/