Project Stage 2
CS 839, Spring 2019

Group 26

Akshaya Kalyanaraman (Email: kalyanarama3@wisc.edu)
Elena Milkai (Email: milkai@wisc.edu)

1 Web sources did we use to extract the structured data

- The first Web source is the IMDB which is an online database for movies television and video games. This source created the tableA.csv. We extracted movies that were created between the years 2000-2017.

  https://www.imdb.com/search/title?release_date=2000-01-01,2000-12-31&sort=num_votes,desc&ref_=adv_prv

- The second source is the THE MOVIE DB which is also an online database for movies and TV shows. This source created the tableB.csv. We extracted the "Top rated movies".

  https://www.themoviedb.org/movie/top-rated?language=en-US

2 Method used to extract the data

We extracted the data by making requests and getting responses from the Web sources. We requested the page source of each webpage and we get it as an answer. Then we had to understand the HTML structure of each Web source. IMDB and THE MOVIE DATABASE have totally different structures in their HTMLs. So, we used a different xpath to extract the attributes from the two Web sources. Both Web sources uses title, date of release, rating and description for the movies they store. We extracted those 4 attributes with a different xpath based on the source. We store our data in data-frames and at the end in .csv files.

3 Type of entity extracted

We extracted movies.

The attributes that we extracted from both the sources:

1. Title: The title of the movie (text data).
2. Date_of_Release: The date that the movie was released. In tableA (IMDB) is just the year of release (ex: 2000) while in tableB (THE MOVIE DATABASE) is the whole date (ex: October 20, 1995).
3. Rating: The rating of the movie (people votes). In tableA (IMDB) is in the form X out of 10 (ex: 8.5) while in tableB (THE MOVIE DATABASE) is X out of 100 (ex: 90)
4. Description: A brief description of the movie (text data).

These attributes were actually common in both the Web sources and the most informative for a movie description. Moreover, they can be used in order to compare two movies in different datasets.

The first table (in tableA.csv) is the table created by IMDB and it has 3296 entries.
The second table (in tableB.csv) is the table created by THE MOVIE DB and it has 7100 entries.
There were about 800 common movies from both tables A and B.

4 Open Source Tools

- To parse our HTML document and extract the containers, we used a Python module named BeautifulSoup, the most common web scraping module for Python. This tool can easily extract the DOM tree from a website and also has some built-in functions that allows the user to locate the html tags of a website.

- We used Pandas to store and manipulate our data during the extraction. Pandas is a software library written for the Python for data manipulation and analysis. It offers data structures and operations helpful when working with tables.