CS 839 Data Science

Project Stage 2

< Group 14 >

1. The names of all team members.
   - Wen-Fu Lee (wlee256@wisc.edu)
   - Yuan-Ting Hsieh (yhsieh28@wisc.edu)
   - Yahn-Chung Chen (chen666@wisc.edu)
2. A description of the two Web data sources that you have selected.
   - We decide to extract movies and related data. The first web source is “IMDB”. It is now the world’s most popular and authoritative source for movie, TV and celebrity content. They offer a searchable database of more than 185 million data items including more than 3.5 million movies, TV and entertainment programs and 7 million cast and crew members.
   - The second web source is rotten tomatoes. It is the most trusted recommendation sources for quality entertainment. As the leading online aggregator of movie and TV show reviews from professional critics, Rotten Tomatoes offers the most comprehensive guide to what's Fresh. The world famous Tomatometer score represents the percentage of positive professional reviews for films and TV shows and is used by millions of fans to help with their viewing decisions. Rotten Tomatoes designates the best-reviewed movies and TV shows as Certified Fresh™. That accolade is awarded with a Tomatometer score of 75% and higher and a required minimum number of reviews.
3. A description of how you have extracted structured data from the two data sources.
   - First we observe the template/html tags of the source website. Then we use BeautifulSoup to parse the website and add rules to extract the attributes we want.
4. Describe the type of entity you extract, briefly describe the two tables, list the number of tuples per table.
   - The type of entity: movie
   - The two tables
     - One is for IMDB movies, and the other one is for Rotten Tomatoes movies
     - In these two tables, both include the following attributes
       - movie_no: movie number
       - movie_name: movie name
       - movie_year: movie year
       - movie_certificate: movie certificate
       - movie_runtime: movie runtime
       - movie_genre: movie genre
       - movie_score: imdb score
       - movie_gross: movie gross
       - movie_director: movie directors
       - movie_star: movie stars

- movie_writer: movie writer
- tomatoter: tomatoter review socre
- audience: audience review score

- The number of tuples
    - IMDB: 3000
    - Rotten Tomatoes: 3012

5. The names of open-source tools you have used in this project stage and a brief description of what they do.
    - BeautifulSoup: It is the most commonly used open-source library used for building website crawler. It can easily extract the DOM tree from a website and then allow user to locate html tags of a website with some built-in functions.
    - Requests: It is an elegant and simple HTTP library for Python, built for human beings.
    - Pandas: It is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. It is used a lot in data science/machine learning problems to deal with data cleaning/exploration/transformation/integration tasks.