CS 839 – Project

Stage 2: Crawling and extracting structured data from Web pages

Group Members: Pradeep Kumar, Pulkit Kapoor, Sonu Agarwal

Description of the two Web data sources

We decided to extract movie data from two different data sources.
Source1: http://www.imdb.com/
Source2: https://www.allmovie.com/

We converged on these sources after manually browsing through several websites which provide movie information. We looked for a pair of websites which provided a large number of common attributes and all these attributes were present in some fixed format. We then decided the set of attributes which we wanted to extract from the two sources.

Description of how structured data was extracted

- We used a python package (request) to get html source pages which contain intended movie data.
- We then visually inspected the html page to identify the structure (class, division, span, tag) corresponding to each attribute which we wanted to extract. We inspected this structure manually created DOM tree of the html page.
- Once we decided where each attribute existed in the hierarchy, we then used another python package (BeautifulSoup) to parse the html page. This gave us the DOM tree of the html page.
- Now we used the learnt structure to automatize the extraction process. We iteratively searched for the corresponding structure (division, class, span, tag) to extract each attribute, which were decided in the beginning.
- Once the values of attributes were found, some preprocessing was done (stripping the string to eliminate additional source, decoding and encoding consistently) and values were stored in the data list and finally all data were stored in csv file.

Type of entity extracted and details:

Movie data have been extracted from two sources. Data corresponding to each source has been stored in separate table (in separate files).

Data Source 1: https://www.allmovie.com/
Number of tuples: 3650

Attributes: Title, Certificate, Genre, Rating, Running Time, Directors, Writers (NULL), Stars Cast, Country, Language (NULL), Budget (NULL), Gross (NULL), Release Date, Production Company

Data Source 2: http://www.imdb.com/
Number of tuples: 3700
Attributes: Title, Certificate, Genre, Rating, Running Time, Directors, Writers, Stars Cast, Country, Language, Budget, Gross, Release Date, Production Company

Note:

- Attribute_Name (NULL)" represents that all values of this feature have been set to NULL as the corresponding source didn't provide values for this attribute.
- All missing values have been replaced by NULL.
- Multiple values for the same attribute have been joined using pipe symbol (“|”).

Names of open-source tools used:

- requests: It provides the method to fetch html source page given a link.
- BeautifulSoup: It provides the methods to parse the html to create a DOM tree and methods to iterative search for values with given structure (class, division, tag etc).