University of Wisconsin-Madison

Computer Science 839

Data Science: Principles, Algorithms, and Applications

Web Crawling and Data Extracting

Team Member:	Contact Email:
Jiatong Li	jli872@wisc.edu
Siyu Wang	swang739@wisc.edu
Xinyu Zhang	xzhang959@wisc.edu

April 12, 2019

Jiatong Li
Siyu Wang
Xinyu Zhang

1 Web Sources

(1) Internet Movie Database: Internet Movie Database (IMDb) is the world's most popular and authoritative source for movie, TV and celebrity content. Find ratings and reviews for the newest movie and TV shows.

(2) The Movie Database: The Movie Database (TMDb) is a popular, user editable database for movies and TV shows.

2 Method - How we extracted

We extract data from two different movie sources based on following steps:

(1) IMDb: **a.** We used Scrapy recursive web crawling framework to extract data from multiple pages. **b.**We obtained structured data from each page with the help of XPath and CSS selectors.

(2) TMDb: **a.** We first extracted the links of each movie from the main movie list. **b.** We input the extracted web pages to Scrapy. **c.** We extracted interested information from each page.

3 Type of Entity

Movie: We selected movies as our type of entity.

4 Information of Two Tables

Number of tuples: Since the requirement says, we should have at least 3k tuples and no more than 10k tuples. We kept 5k tuples in each of our tables.

Web Sources	Number of tuples
IMDb	5000
TMDb	5000

Attributes: In both of these two sources, we extracted the following attributes: movie, directors, stars, year, runtime and genres. The type of these attributes are all strings.

Schema:

Attributes	Discription
movie	This attribute describes the name of the movie.
directors	This attribute includes names of all the directors.
stars	This attribute includes names of all the stars.
year	This attribute describes the release year of the movie.
runtime	This attribute describes the length of the movie.
genres	This attribute decribes the type of the movie.

5 Open-source Tool

Scrapy: Scrapy is a web crawling framework, written in Python. Originally designed for web scraping, it can also be used to extract data using APIs or as a general-purpose web crawler. It is currently maintained by Scrapinghub Ltd., a web-scraping development and services company. Using Scrapy, we define spiders for each web source. And Scrapy provides the framework for processing the response for each request fired. Also, support for crawling the multiple links in the HTML is provided by Scrapy.