INF 558/CSCI 563: Building Knowledge Graphs

Homework 1: Crawling

Released: Jan 16th, 2020
Due: Jan 23rd, 2020 @ 23:59

Ground Rules

This homework must be done individually. You can ask others for help with the tools, however, the submitted homework has to be your own work.

Summary

In this homework, you will utilize web crawlers to collect webpages and extract data from the Internet Movie Database (IMDb) website (https://www.imdb.com/).
A web crawler is a program/bot that systematically browses the World Wide Web (WWW), typically for the purpose of web indexing (web spidering). It starts with a list of seed URLs to visit, and as it visits each webpage, it finds the links in that web page, and then visits those links and repeats the entire process. There are two web crawler technologies you are required to use in this homework:

- Scrapy (https://scrapy.org): A python library crawler
- import.io (https://import.io/): An interactive crawling platform

Task 1: Scrapy (7 points)

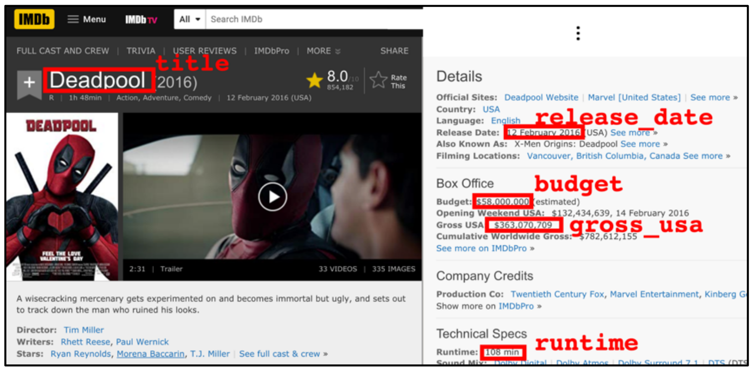
Task 1.1 (3 points)

Crawl at least **5000** webpages of **Sci-Fi movies/shows** in IMDb using Scrapy. Extract and generate the following attributes for each webpage:

`id`	unique id for the webpage
`url`	url of the webpage
`timestamp_crawl`	timestamp of the crawling event

`title`
`release_date`
`budget`
`gross_usa`
`runtime`
	see Figure 1
	if attribute doesn't exist, set as empty string

Store your crawled data into a JSON-Lines (`.jl`) file. In this file format, each line is a valid JSON object (dictionary) that holds the attributes listed above for a single crawled webpage. You can check the attached file `sample.jl` to understand the format. While crawling, please make sure you obey the website's politeness rules (i.e. sleep time between requests) in order to avoid getting banned.



Deadpool (2016)
8.0/10 854,182
R | 1h 48min | Action, Adventure, Comedy | 12 February 2016 (USA)

A wisecracking mercenary gets experimented on and becomes immortal but ugly, and sets out to track down the man who ruined his looks.

Director: Tim Miller
Writers: Rhett Reese, Paul Wernick
Stars: Ryan Reynolds, Morena Baccarin, T.J. Miller | See full cast & crew »

Details
Official Sites: Deadpool Website | Marvel [United States] | See more »
Country: USA
Language: English
Release Date: 12 February 2016 (USA) See more »
Also Known As: X-Men Origins: Deadpool See more »
Filming Locations: Vancouver, British Columbia, Canada See more »

Box Office
Budget: $58,000,000 (estimated)
Opening Weekend USA: $132,434,639, 14 February 2016
Gross USA: $363,070,709
Cumulative Worldwide Gross: $782,612,155
See more on IMDbPro »

Company Credits
Production Co: Twentieth Century Fox, Marvel Entertainment, Kinberg G
Show more on IMDbPro »

Technical Specs
Runtime: 108 min
Sound Mix: Dolby Digital | Dolby Atmos | Dolby Surround 7.1 | DTS (DTS

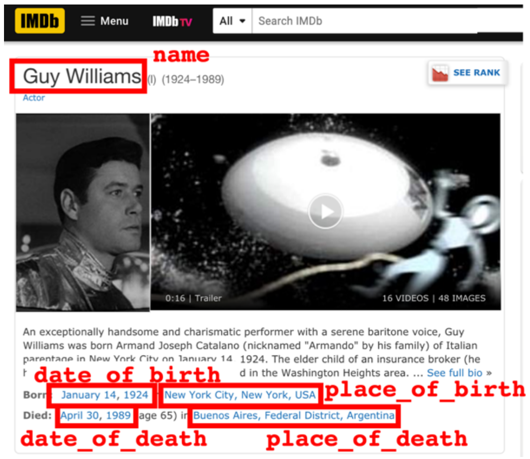
Figure 1: An example movie/show webpage with the required attributes

Task 1.2 (3 points)

Similar to the previous task, crawl at least **5000** webpages of cast (actors and actresses) in IMDb using Scrapy. Extract and generate the following attributes for each cast webpage:

`id`	unique id for the webpage
`url`	url of the webpage
`timestamp_crawl`	timestamp of the crawling event
`name`	see Figure 2 if attribute doesn't exist, set as empty string
`date_of_birth`	
`place_of_birth`	
`date_of_death`	
`place_of_death`	

Similarly, store your crawled data into a JSON-Lines (`.jl`) file.



Guy Williams (I) (1924–1989)

An exceptionally handsome and charismatic performer with a serene baritone voice, Guy Williams was born Armand Joseph Catalano (nicknamed "Armando" by his family) of Italian parentage in New York City on January 14, 1924. The elder child of an insurance broker (he h d in the Washington Heights area. ... See full bio »

Born: January 14, 1924 New York City, New York, USA

Died: April 30, 1989 age 65) in Buenos Aires, Federal District, Argentina

Figure 2: An example cast webpage with the required attributes

Task 1.3 (1 points)

Answer the following questions (no more than 2 sentences for each question):

1.3.1 What is the seed URL(s) you used for each task?

1.3.2 How did you manage to only collect movie/show or cast pages?

1.3.3 Did you need to discard irrelevant pages? If so, how?

1.3.4 Did you collect the required number of pages? If you were not able to do so, please describe and explain your issues.

Task 2: import.io (3 points)

Use the interactive platform of **import.io** to crawl and extract the same attributes as in Task 1.1 from a total of **400 Romance movies/shows** in IMDb. Store your output in a comma-separated values (`.csv`) file. You are expected to create a chain of at least 2 chained extractors, in which at least 1 extractor's output is used to generate the seed URLs of another extractor (e.g. the movie/show webpage extractor). The general expected chain pipeline is shown in Figure 3. Describe the chain you constructed. For each extractor in the chain: describe its inputs, outputs and role (what is it doing) and attach a screenshot of the run including the top output rows (by clicking "Preview Data" in the "Run History" screen), as seen in Figure 4.

- The free trial version of the service is limited to 1000 URL queries, which should be sufficient for the chain of extractors you are expected to construct for this task.
- You may use the "URL Generator" to create the list of seed URL(s) of your first extractor in the chain. This is not mandatory but may make the task easier for you.

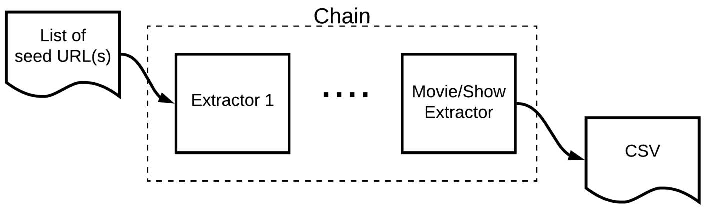
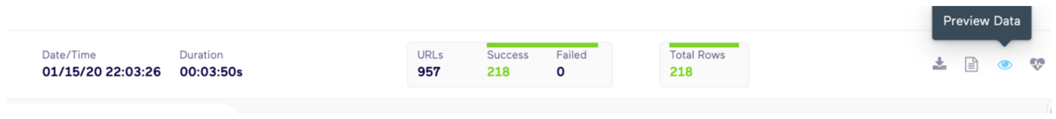


Figure 3: The general extraction pipeline you are expected to build



Date/Time 01/15/20 22:03:26
Duration 00:03:50s
URLs 957
Success 218
Failed 0
Total Rows 218

Figure 4: A screenshot example of an extractor's run

Submission Instructions

You must submit (via Blackboard) the following files/folders in a single `.zip` archive named `Firstname_Lastname_hw01.zip`:

- `Firstname_Lastname_hw01_report.pdf`: pdf file with your answers to Task 1.3 and 2
- JSON-Lines files containing the data you crawled using Scrapy for Task 1.1 and 1.2:
  - `Firstname_Lastname_hw01_scrapy_title.jl`: Generated data from Task 1.1
  - `Firstname_Lastname_hw01_scrapy_cast.jl`: Generated data from Task 1.2
- `Firstname_Lastname_hw01_importio_title.csv`: A Comma-Separated values file of the generated data results from Task 2
- `source`: This folder includes all the code you wrote to accomplish Task 1 (i.e. your Scrapy crawler, seed files, your script/program to eliminate unwanted pages and store webpages into JSON-Lines format, etc…)

Additional Notes

1. We also provide you with a script called `post_processing.py`, you can use this script to validate the structure of your JSON-Lines files in Task 1.1:
   a. Prerequisites: python 3 and the `ujson` package (`pip install <package>`)
   b. Usage: The script takes one argument `<path>` which is the path of your `jl` file, processes the file and prints a message to let the user know if it is valid.
   `python post_processing.py /home/users/sample.jl` will print:
   ```
   Process: (0/2)...
   Process: (1/2)...
   Finished processing, looks good, found 2 entries.
   ```
2. It is your responsibility to validate the structure of the outputs in tasks 1.2 and 2 prior to submitting.