

DSCI 558: Building Knowledge Graphs

Homework 1: Crawling

Released: Jan 25th, 2021

Due: Jan 31st, 2021 @ 23:59

Ground Rules

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

Summary

In this homework, you will utilize web crawlers to collect webpages and extract data from four movie review websites:

- Metacritic
- IMDB
- Rotten Tomatoes
- Rogerebert

Each student will receive two random websites to crawl, one for Task 1 and one for Task 2. Please check the **Website_Assignment.pdf** document for your website assignment.

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. You are required to use **Scrapy** (<https://scrapy.org>) in this homework.

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 3
- JSON-Lines files containing the data you crawled using Scrapy for Task 1 and 2:
 - **Firstname_Lastname_hw01_movies.jl**: Generated data from Task 1
 - **Firstname_Lastname_hw01_cast.jl**: Generated data from Task 2
- **source**: This folder includes all the code you wrote to accomplish Task 1 and 2 (i.e. your Scrapy crawler, seed files, your script/program to eliminate unwanted pages and store webpages into JSON-Lines format, etc....)

Task 1: Movies (5 points)

Crawl at least 5000 webpages of Comedy/Drama movies from IMDB, Metacritic or RottenTomatoes using Scrapy.

Extract and generate the following attributes (*all attribute names are lowercase*) for each webpage. **If an attribute doesn't exist, set as empty string/list depending on the field type.**

Attribute	Type	Note
Id	str	Unique id for the webpage (self-generated)
url	str	Url of the webpage
timestamp_crawl	str	Timestamp of the crawling event
title	str	see Figure 1, 2, 3, 4 • Extract genres if you are crawling IMDB , Metacritic and RottenTomatoes . Otherwise, extract running_time .
director	str	
cast	list of str	
genres/running_time	list of str if genres	
	str if running_time	
release_date	str	



OUR FRIEND

Title

Critics Consensus

Our Friend's occasionally frustrating approach to dramatizing its fact-based story is often offset by a trio of starring performances led by a never-better Jason Segel.



83%

TOMATOMETER
Total Count: 82



90%

AUDIENCE SCORE
Verified Ratings: 10

[SEE SCORE DETAILS](#)

MOVIE INFO

OUR FRIEND tells the inspiring and extraordinary true story of the Teague family—journalist Matt (Casey Affleck), his vibrant wife Nicole (Dakota Johnson) and their two young daughters—and how their lives are upended by Nicole's heartbreaking diagnosis of terminal cancer. As Matt's responsibilities as caretaker and parent become increasingly overwhelming, the couple's best friend Dane Fauchaux (Jason Segel) offers to come and help out. As Dane puts his life on hold to stay with his friends, the impact of this life altering decision proves greater and more profound than anyone could have imagined.

Genre: Drama, Comedy

Genre

Original Language: English

Director: Gabriela Cowperthwaite

Director

Producer: Michael A. Pruss, Teddy Schwarzman, Ryan Stowell, Kevin J. Walsh

Writer: Brad Ingelsby

Release Date (Theaters): Jan 22, 2021 Limited

Release Date

Release Date (Streaming): Jan 22, 2021

Runtime: 2h 4m

Production Co: Scott Free Productions, STX International, Black Bear Pictures

Aspect Ratio: Flat (1.85:1)

CAST & CREW



Jason Segel
Dane Fauchaux



Dakota Johnson
Nicole Teague



Casey Affleck
Matt Teague



Gwendoline Christie
Teresa



Marielle Scott
Kat



Cherry Jones
Faith Pruett

Cast (Dont need to "Show All")

[Show all Cast & Crew](#)

Figure 1. Movie Information from Rotten Tomatoes

+

Our Friend

(2019)

Title

★ 7.3



649

★ Rate This

The Friend *(original title)*

R | 2h 4min | Drama | 22 January 2021 (USA)

Release Date

2:31 | Trailer

5 VIDEOS | 30 IMAGES

After receiving life-altering news, a couple finds unexpected support from their best friend, who puts his own life on hold and moves into their family home, bringing an impact much greater and more profound than anyone could have imagined

Director: [Gabriela Cowperthwaite](#) **Director**

Writers: [Brad Ingelsby](#) (screenplay by), [Matthew Teague](#) (based on the article entitled 'The Friend' by)

Stars: [Dakota Johnson](#), [Jason Segel](#), [Isabella Kai](#) [See full cast & crew »](#)

Cast (Dont need to "See full cast")

Storyline

Edit

After receiving life-altering news, a couple finds unexpected support from their best friend, who puts his own life on hold and moves into their family home, bringing an impact much greater and more profound than anyone could have imagined

[Plot Summary](#) | [Add Synopsis](#)

Plot Keywords: [based on real events](#) | [singing in a car](#) | [See All \(2\) »](#)

Genres: [Drama](#) **Genre**

Motion Picture Rating (MPAA)
 Rated R for language | [See all certifications »](#)

Parents Guide: [Add content advisory for parents »](#)

Figure 2. Movie Information from IMDB

Our Friend

2021

Title

METAScore

Mixed or average reviews based on 19 Critic Reviews

See All

56

USER SCORE

No score yet

tbd

VOTE NOW

0 1 2 3 4 5 6 7 8 9 10

Play Sound

Now Playing: The Friend: Jason Segel-Standup



Movie Details & Credits

Gravitas Ventures | Release Date: January 22, 2021 | Not Rated

Release Date

Starring: Ahna O'Reilly, Azita Ghanizada, Casey Affleck, Cherry Jones, Dakota Johnson, Denée Benton, Gwendoline Christie, Isabella Kai Rice, Jake Owen, Jason Bayle, Jason Segel, Jeronimo Spinx, John McConnell, Marielle Scott, Michael Papajohn, Reed Diamond, Violet McGraw

Cast

Summary: [Formerly known as "The Friend."] After learning that his terminally ill wife has six months to live, a man welcomes the support of his best friend who moves into their home to help out. His impact on the whole family is much greater than anyone could have imagined.

Director: Gabriela Cowperthwaite

Genre(s): Drama

Rating: Not Rated

Runtime: 124 min

Director
Genre

Figure 3. Movie Information from Metacritic

Film Credits



Cast

- Casey Affleck as Matthew Teague
- Dakota Johnson as Nicole Teague
- Jason Segel as Dane Fauchaux
- Gwendoline Christie as Teresa
- Cherry Jones as Faith
- Ahna O'Reilly as Gale
- Jake Owen as Aaron

Director

Gabriela Cowperthwaite

Writer

Brad Ingelsby

Cinematographer

Joe Anderson

Editor

Colin Patton

Composer

Rob Simonsen

Title

Our Friend

Release date

2021

★★★

Rated R for language.

124 minutes

running time

Figure 4. Movie Information from Rogerebert

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.

Task 2: Actor/Actress (4 points)

Similar to the previous task, crawl at least 5000 webpages of cast (actors and actresses) from IMDb, Metacritic or RottenTomatoes using Scrapy. Extract and generate the following attributes for each cast webpage. **If an attribute doesn't exist, set as empty string/list if the attribute type is string/list and set as -1 if the attribute type is integer.**

Attribute	Type	Note
Id	str	Unique id for the webpage (self-generated)
url	str	Url of the webpage
timestamp_crawl	str	Timestamp of the crawling event
name	str	Name of actor/actress
movies <ul style="list-style-type: none">• title• year• role	list of json object with 3 fields <ul style="list-style-type: none">• title: str• year: integer• role: str	<p>A list of up to 5 most recent movies that the actor/actress participates in.</p> <p>Each movie should have name, release year and role of actor/actress. (See Figure 5, 6, 7, 8).</p> <p>It is fine if an actor/actress has less than 5 movies.</p>

Task 3: Report (1 points)

Answer the following questions (no more than 3 sentences for each question):

- What is the seed URL(s) you used for each task?
- How did you manage to only collect movie or cast pages?
- Did you need to discard irrelevant pages? If so, how?
- Did you collect the required number of pages? If you were not able to do so, please describe and explain your issues.



Chris Evans Name

Highest Rated: 🍅 100% [Superpower Dogs \(2019\)](#)

Lowest Rated: 🌿 14% [Playing It Cool \(2014\)](#)

Birthday: Jun 13, 1981

Birthplace: Boston, Massachusetts, USA

Actor Chris Evans began his acting career in typical fashion, but it was his rapid rise to stardom which was unusual. Bitten by the acting bug in the first grade, Evans started out appearing in school plays and theater camp; from there it was a quick jump to local community theater, and later, an internship for a casting office. Once Evans made friends with a few agents on the job, ... [Read more](#)

FILMOGRAPHY

5 most recent movies

Movies		Title	Role	BOX OFFICE	Year
TOMATOMETER®	AUDIENCE SCORE	TITLE	CREDIT		YEAR ↕
🍅 97%	🍿 92%	Knives Out	Ransom Drysdale (Character)	\$165.4M	2019
🍅 94%	🍿 90%	Avengers: Endgame	Steve Rogers/ Captain America (Character)	\$858.4M	2019
🍅 100%	🍿 88%	Superpower Dogs	Narrator	-	2019
No Score Yet	🍿 57%	The Red Sea Diving Resort	Ari Levinson (Character)	-	2019
🍅 85%	🍿 91%	Avengers: Infinity War	Steve Rogers (Character)	\$678.8M	2018
🌿 29%	🍿 78%	The Red Sea Diving Resort	Unknown (Character)	-	2018
🍅 73%	🍿 85%	Gifted	Frank (Character)	\$24.8M	2017
🍅 90%	🍿 89%	Captain America: Civil War	Steve Rogers/ Captain America (Character)	\$408.1M	2016
🍅 76%	🍿 83%	Avengers: Age of Ultron	Steve Rogers/ Captain America (Character)	\$459M	2015
🌿 14%	🍿 33%	Playing It Cool	Me (Character), Executive Producer	-	2014

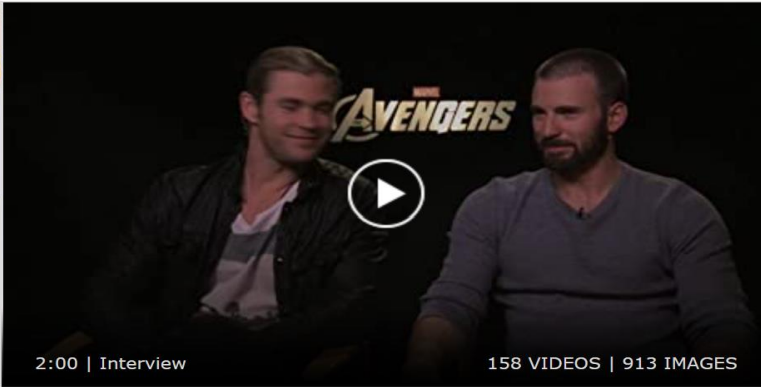
Figure 5. Actor Information from Rotten Tomatoes

Chris Evans (V)

Name

Top 500

Actor | Producer | Soundtrack



Christopher Robert Evans began his acting career in typical fashion: performing in school productions and community theatre. He was born in Boston, Massachusetts, the son of Lisa (Capuano), who worked at the Concord Youth Theatre, and G. Robert Evans III, a dentist. His uncle is congressman [Mike Capuano](#). Chris's father is of half German and half ... [See full bio](#) »

Filmography

Show all | Show by... | Edit

Jump to: [Actor](#) | [Producer](#) | [Soundtrack](#) | [Director](#) | [Thanks](#) | [Self](#) | [Archive footage](#)

Actor (57 credits)

Hide

Title	Year	Role
Lightyear (announced)	2022	Buzz Lightyear (voice)
Bermuda (pre-production)		Dr. Fisk (rumored)
Little Shop of Horrors (pre-production)		Orin Scrivello (rumored)
The Gray Man (pre-production)		Lloyd Hansen
Don't Look Up (filming)	2021	Peter Isherwell
Scott Pilgrim vs. the World Water Crisis (Video)	2020	Lucas Lee
Defending Jacob (TV Mini-Series)	2020	Andy Barber - After (2020) ... Andy Barber - Job (2020) ... Andy Barber - Wishful Thinking (2020) ... Andy Barber - Visitors (2020) ... Andy Barber - Damage Control (2020) ... Andy Barber Show all 8 episodes
Knives Out	2019	Ransom Drysdale
The Red Sea Diving Resort	2019	Ari Levinson
Avengers: Endgame	2019	Steve Rogers / Captain America
Captain Marvel	2019	Steve Rogers (uncredited)

Figure 6. Actor Information from IMDB

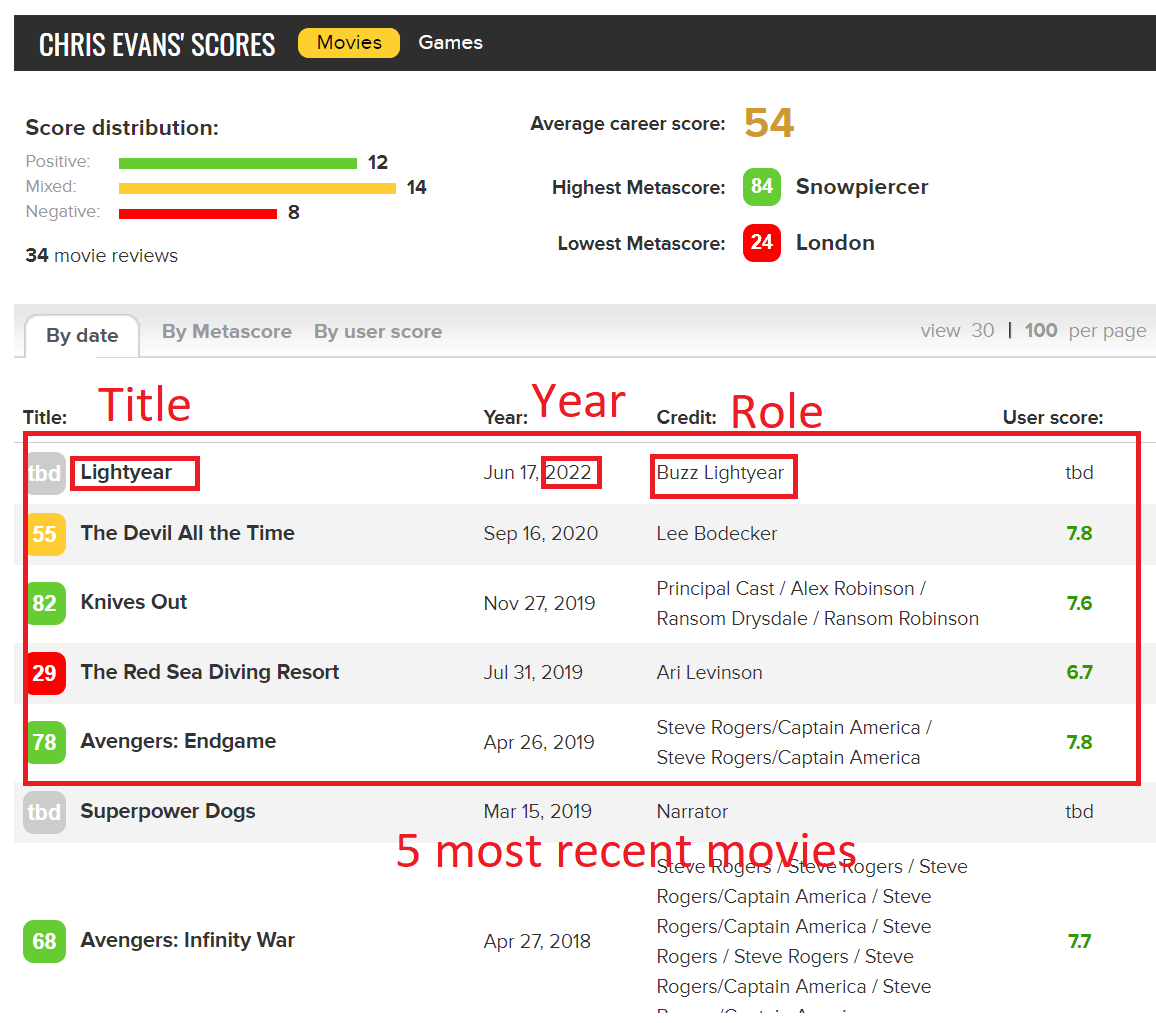


Figure 7. Actor Information from Metacritic

Chris Evans

Name

[Find on IMDB](#)
[Find on Wikipedia](#)

5 most recent movies

Reviews


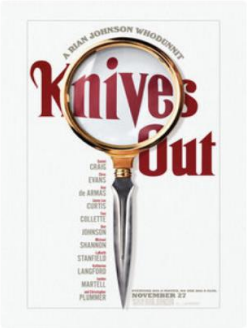
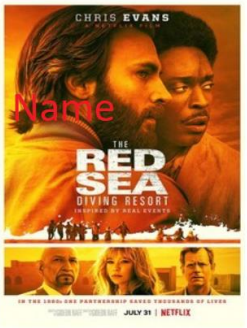





 <div><div>title</div><div>year</div><div>Defending Jacob</div><div>2020</div><div>Brian Tallerico</div><div>★★</div><div>Andy Barber</div><div>Role</div></div>	 <div><div>Knives Out</div><div>(2019)</div><div>Brian Tallerico</div><div>★★★★</div><div>Hugh Robinson</div></div>	 <div><div>Name</div><div>The Red Sea Diving Resort</div><div>(2019)</div><div>Brian Tallerico</div><div>★</div><div>Ari Levinson</div></div>	 <div><div>Running with Beto</div><div>(2019)</div><div>Christy Lemire</div><div>★★★</div><div>Himself</div></div>
 <div><div>Avengers: Endgame</div><div>(2019)</div><div>Brian Tallerico</div><div>★★★</div><div>Steve Rogers / Captain America</div></div>	 <div><div>Avengers: Infinity War</div><div>(2018)</div><div>Matt Zoller Seitz</div><div>★★★</div><div>Steve Rogers / Nomad</div></div>	 <div><div>gifted</div><div>(2017)</div><div>Glenn Kenny</div><div>★★★★</div><div>Frank Adler</div></div>	 <div><div>Captain America: Civil War</div><div>(2016)</div><div>Matt Zoller Seitz</div><div>★★★★</div><div>Steve Rogers / Captain America</div></div>

Figure 8. Actor Information from Rogerebert