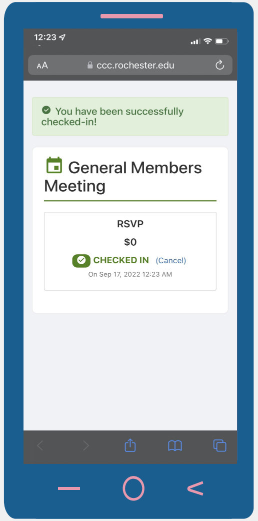
Webscrape with UDSC

Data Science Essentials

Scan for Attendance :)



You have been successfully checked-in!

General Members Meeting

RSVP

$0

CHECKED IN (Cancel)

On Sep 17, 2022 12:23 AM

01

What is Webscraping?

Webscraping /web ‘skrāpiNG/

noun (gerund)

The automated extraction of data from a website.
This data is then collected and exported into a cleaner format.

Examples:

1. Don’t you know the coolest people are into webscraping?
2. Webscraping saved me so much time on clicking links and importing data!
3. My boss likes me the most because I know webscraping XD

Why Should I Webscrape?

- Data is gold
- Previous GMMs, we talked about data life cycle, and spend some time talking about data collection
    - Refresher: Data collection *defines* your analysis!
- Pre-existing, pre-made:
    - Kaggle, Google Public Datasets
- The webscraping process is designed entirely be you, and you can see the data through a completely more personal lens
- In other words, you may find something entirely new and interesting

02

To Get Started

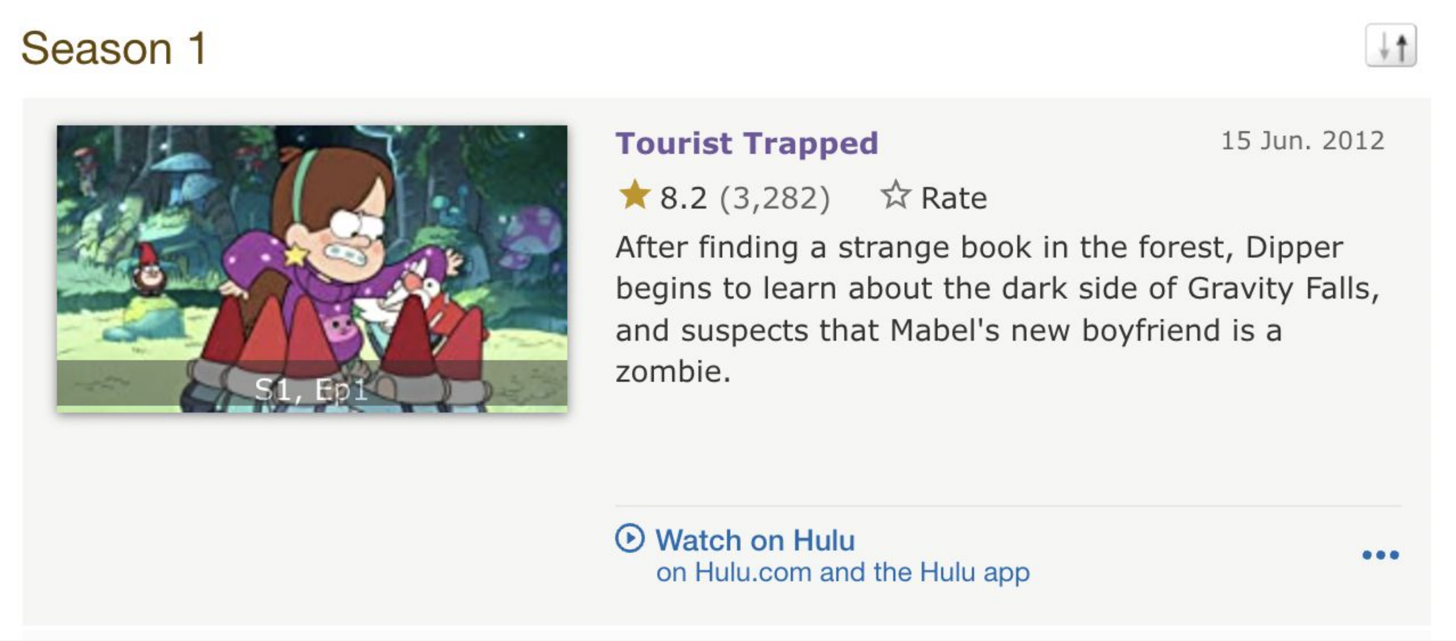
The Big Idea

- We are going to attempt to webscrape information about Gravity Falls Season 1, specifically episode titles, air dates, and their synopses
  - IMDB - ethics?
- Websites are built using HTML
  - Inspect
- Tools
  1. Requests: Get HTML
  2. BeautifulSoup: Read through HTML
  3. Selenium: Interact with dynamic website

Setting Up

- Install requests
- Install BeautifulSoup
- If time allows: Install Selenium
    - Download Chrome if you don't already have it
    - Download Chromedriver
- Have your python text-editor ready, code will be sent to you shortly
- Have open the wikipedia page titled "List of Marvel Cinematic Universe films"

Demo Part 1.1.1 (sans Selenium)



Season 1

Tourist Trapped

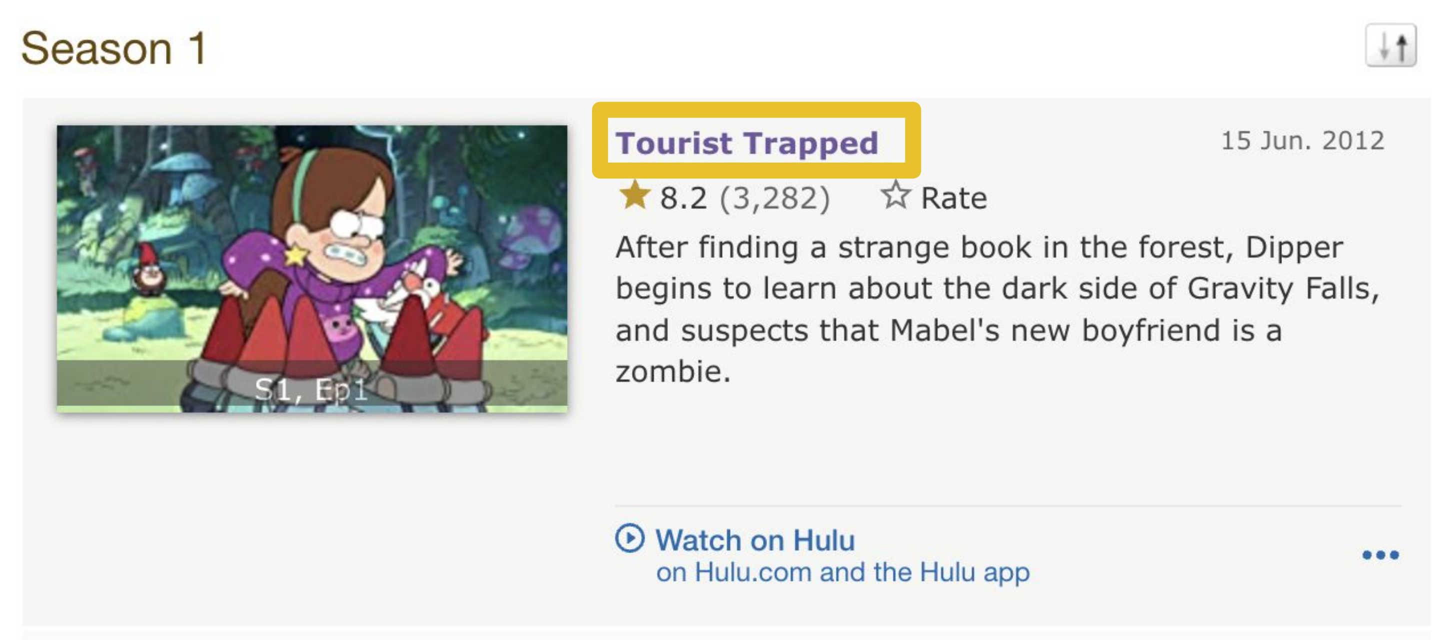
15 Jun. 2012

★ 8.2 (3,282) ☆ Rate

After finding a strange book in the forest, Dipper begins to learn about the dark side of Gravity Falls, and suspects that Mabel's new boyfriend is a zombie.

Watch on Hulu
on Hulu.com and the Hulu app

Demo Part 1.1.1 (sans Selenium)



Season 1

S1, Ep1

Tourist Trapped

15 Jun. 2012

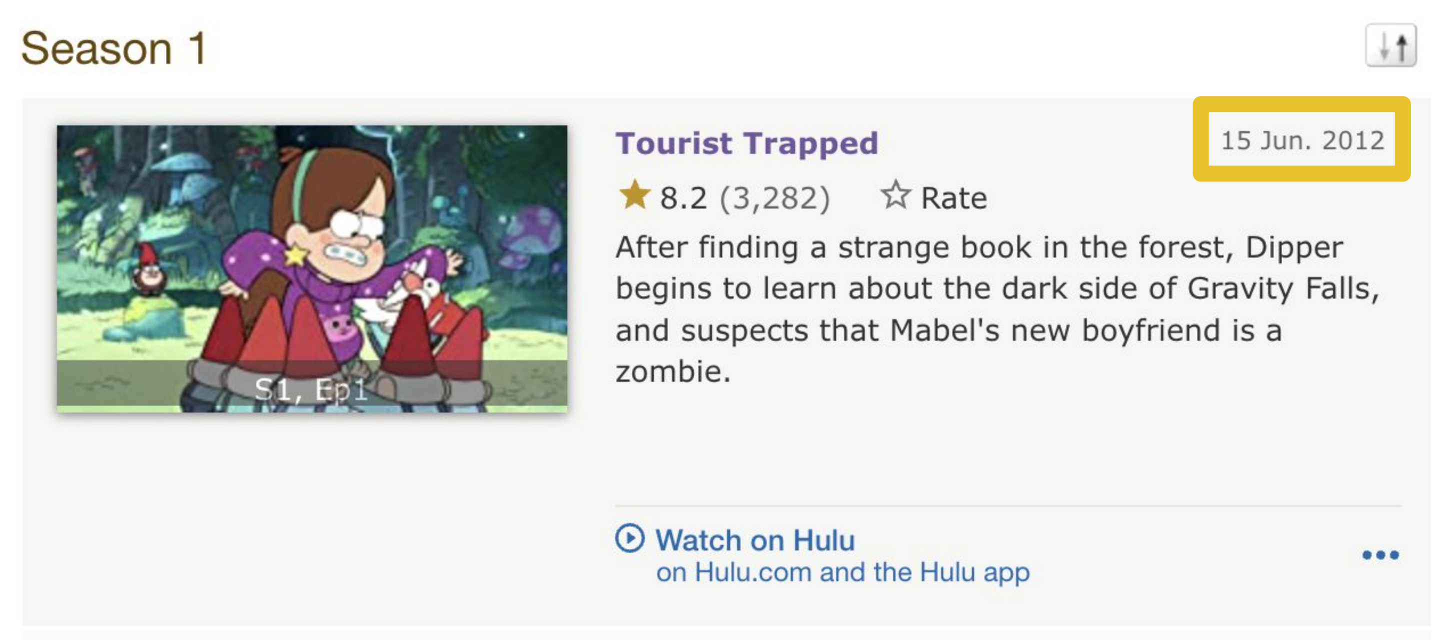
★ 8.2 (3,282) ☆ Rate

After finding a strange book in the forest, Dipper begins to learn about the dark side of Gravity Falls, and suspects that Mabel's new boyfriend is a zombie.

Watch on Hulu
on Hulu.com and the Hulu app

•••

Demo Part 1.1.2 (sans Selenium)



Season 1

S1, Ep1

Tourist Trapped

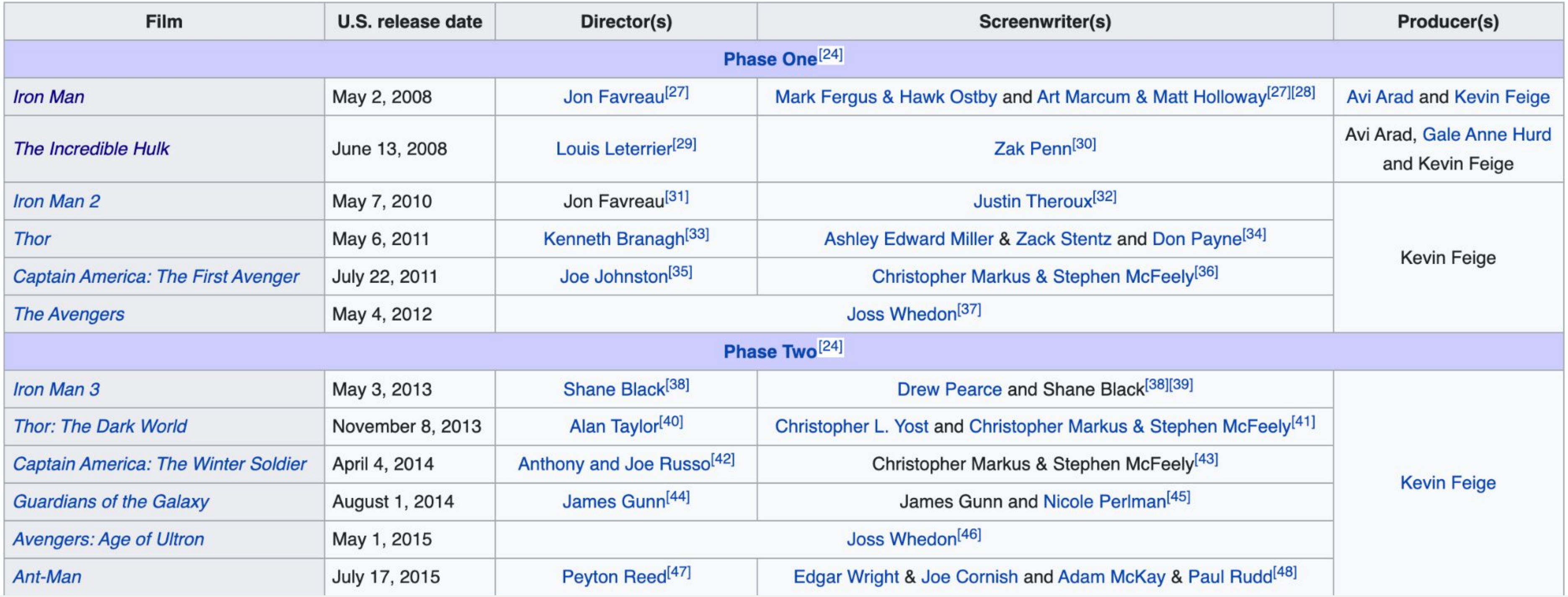
15 Jun. 2012

★ 8.2 (3,282) ☆ Rate

After finding a strange book in the forest, Dipper begins to learn about the dark side of Gravity Falls, and suspects that Mabel's new boyfriend is a zombie.

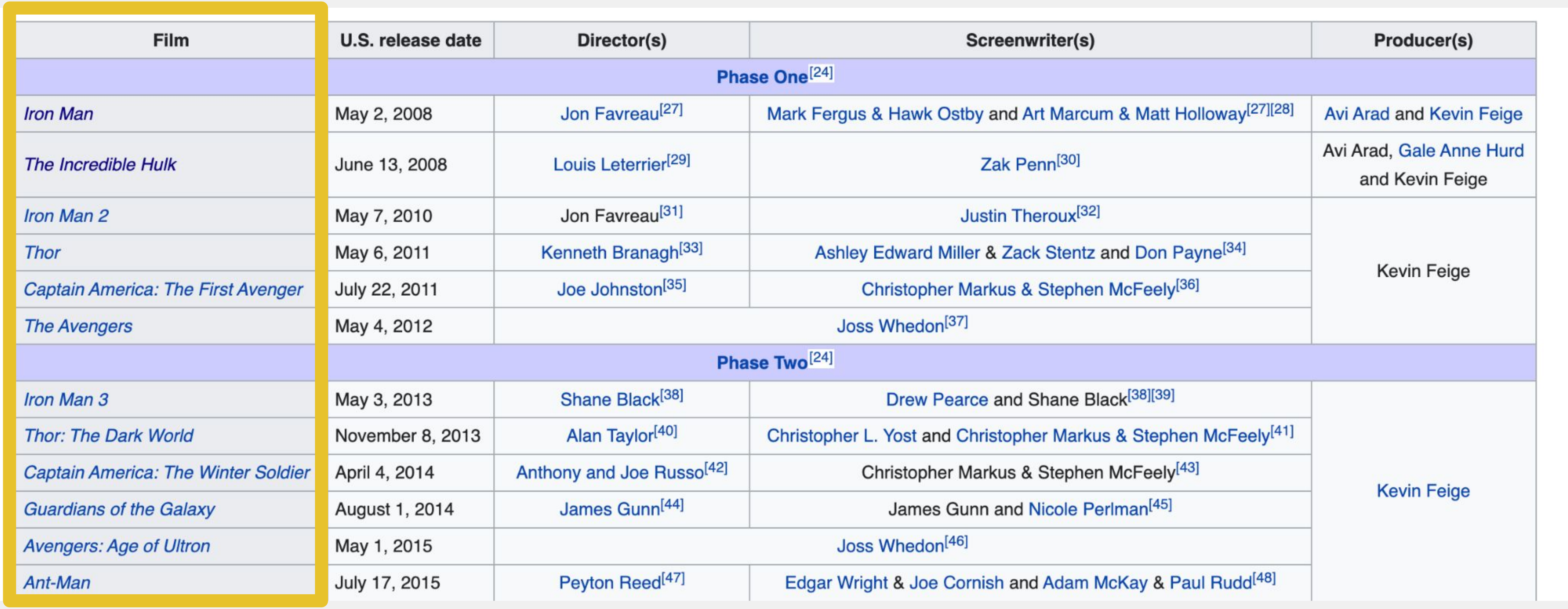
Watch on Hulu
on Hulu.com and the Hulu app

Demo Part 1.2.1 (sans Selenium)



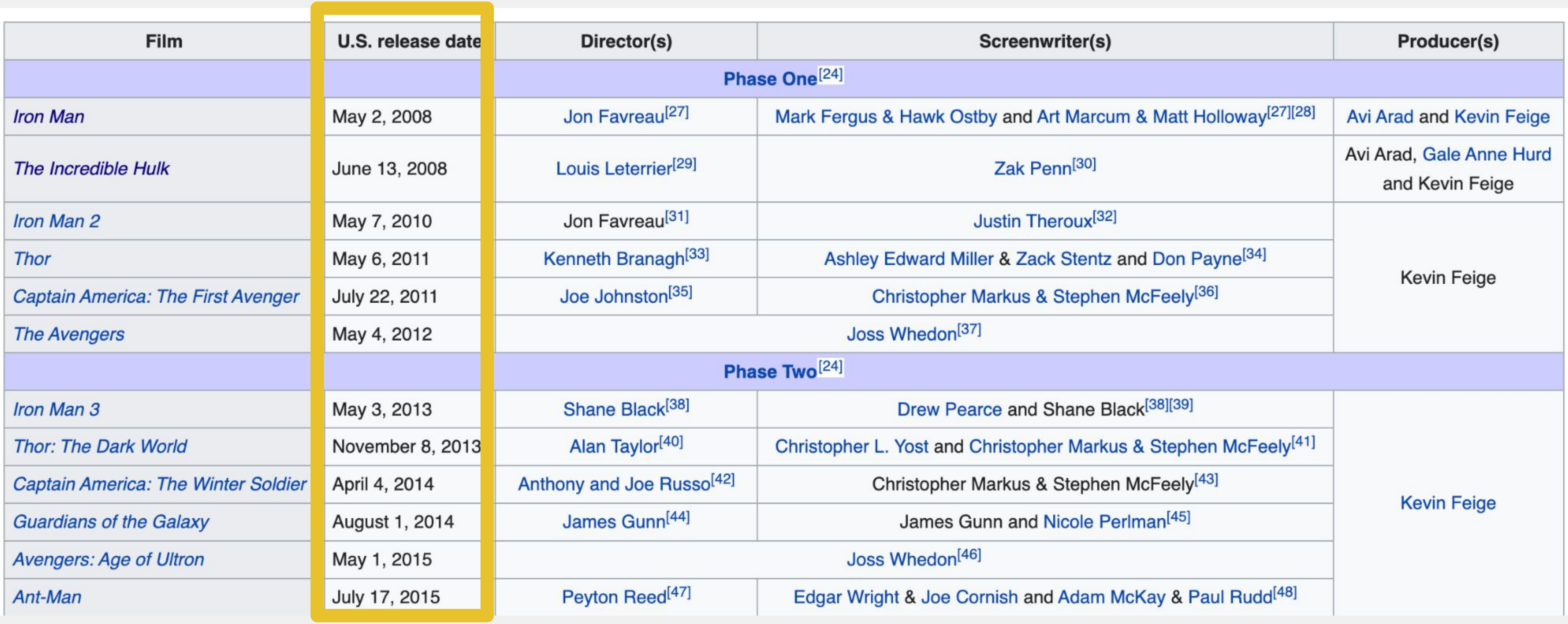
Film	U.S. release date	Director(s)	Screenwriter(s)	Producer(s)
Phase One[24]				
Iron Man	May 2, 2008	Jon Favreau[27]	Mark Fergus & Hawk Ostby and Art Marcum & Matt Holloway[27][28]	Avi Arad and Kevin Feige
The Incredible Hulk	June 13, 2008	Louis Leterrier[29]	Zak Penn[30]	Avi Arad, Gale Anne Hurd and Kevin Feige
Iron Man 2	May 7, 2010	Jon Favreau[31]	Justin Theroux[32]	Kevin Feige
Thor	May 6, 2011	Kenneth Branagh[33]	Ashley Edward Miller & Zack Stentz and Don Payne[34]	
Captain America: The First Avenger	July 22, 2011	Joe Johnston[35]	Christopher Markus & Stephen McFeely[36]	
The Avengers	May 4, 2012	Joss Whedon[37]		
Phase Two[24]				
Iron Man 3	May 3, 2013	Shane Black[38]	Drew Pearce and Shane Black[38][39]	Kevin Feige
Thor: The Dark World	November 8, 2013	Alan Taylor[40]	Christopher L. Yost and Christopher Markus & Stephen McFeely[41]	
Captain America: The Winter Soldier	April 4, 2014	Anthony and Joe Russo[42]	Christopher Markus & Stephen McFeely[43]	
Guardians of the Galaxy	August 1, 2014	James Gunn[44]	James Gunn and Nicole Perlman[45]	
Avengers: Age of Ultron	May 1, 2015	Joss Whedon[46]		
Ant-Man	July 17, 2015	Peyton Reed[47]	Edgar Wright & Joe Cornish and Adam McKay & Paul Rudd[48]	

Demo Part 1.2.2 (sans Selenium)



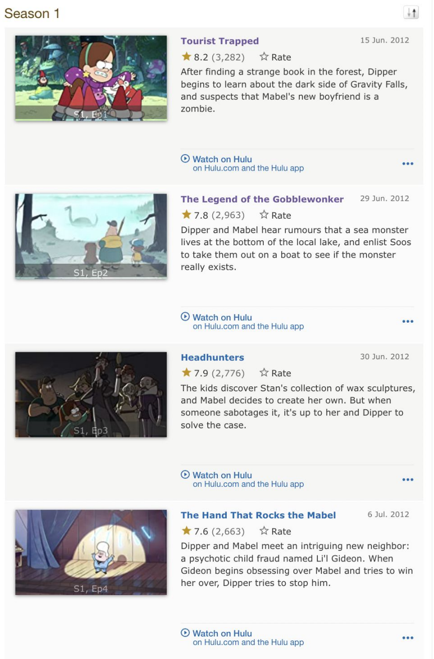
Film	U.S. release date	Director(s)	Screenwriter(s)	Producer(s)
Phase One[24]				
Iron Man	May 2, 2008	Jon Favreau[27]	Mark Fergus & Hawk Ostby and Art Marcum & Matt Holloway[27][28]	Avi Arad and Kevin Feige
The Incredible Hulk	June 13, 2008	Louis Leterrier[29]	Zak Penn[30]	Avi Arad, Gale Anne Hurd and Kevin Feige
Iron Man 2	May 7, 2010	Jon Favreau[31]	Justin Theroux[32]	Kevin Feige
Thor	May 6, 2011	Kenneth Branagh[33]	Ashley Edward Miller & Zack Stentz and Don Payne[34]	
Captain America: The First Avenger	July 22, 2011	Joe Johnston[35]	Christopher Markus & Stephen McFeely[36]	
The Avengers	May 4, 2012	Joss Whedon[37]		
Phase Two[24]				
Iron Man 3	May 3, 2013	Shane Black[38]	Drew Pearce and Shane Black[38][39]	Kevin Feige
Thor: The Dark World	November 8, 2013	Alan Taylor[40]	Christopher L. Yost and Christopher Markus & Stephen McFeely[41]	
Captain America: The Winter Soldier	April 4, 2014	Anthony and Joe Russo[42]	Christopher Markus & Stephen McFeely[43]	
Guardians of the Galaxy	August 1, 2014	James Gunn[44]	James Gunn and Nicole Perlman[45]	
Avengers: Age of Ultron	May 1, 2015	Joss Whedon[46]		
Ant-Man	July 17, 2015	Peyton Reed[47]	Edgar Wright & Joe Cornish and Adam McKay & Paul Rudd[48]	

Demo Part 1.2.2 (sans Selenium)



Film	U.S. release date	Director(s)	Screenwriter(s)	Producer(s)
Phase One[24]				
Iron Man	May 2, 2008	Jon Favreau[27]	Mark Fergus & Hawk Ostby and Art Marcum & Matt Holloway[27][28]	Avi Arad and Kevin Feige
The Incredible Hulk	June 13, 2008	Louis Leterrier[29]	Zak Penn[30]	Avi Arad, Gale Anne Hurd and Kevin Feige
Iron Man 2	May 7, 2010	Jon Favreau[31]	Justin Theroux[32]	Kevin Feige
Thor	May 6, 2011	Kenneth Branagh[33]	Ashley Edward Miller & Zack Stentz and Don Payne[34]	
Captain America: The First Avenger	July 22, 2011	Joe Johnston[35]	Christopher Markus & Stephen McFeely[36]	
The Avengers	May 4, 2012	Joss Whedon[37]		
Phase Two[24]				
Iron Man 3	May 3, 2013	Shane Black[38]	Drew Pearce and Shane Black[38][39]	Kevin Feige
Thor: The Dark World	November 8, 2013	Alan Taylor[40]	Christopher L. Yost and Christopher Markus & Stephen McFeely[41]	
Captain America: The Winter Soldier	April 4, 2014	Anthony and Joe Russo[42]	Christopher Markus & Stephen McFeely[43]	
Guardians of the Galaxy	August 1, 2014	James Gunn[44]	James Gunn and Nicole Perlman[45]	
Avengers: Age of Ultron	May 1, 2015	Joss Whedon[46]		
Ant-Man	July 17, 2015	Peyton Reed[47]	Edgar Wright & Joe Cornish and Adam McKay & Paul Rudd[48]	

Demo Part 2 (with Selenium)



Season 1

Tourist Trapped 15 Jun. 2012

★ 8.2 (3,282) ☆ Rate

After finding a strange book in the forest, Dipper begins to learn about the dark side of Gravity Falls, and suspects that Mabel's new boyfriend is a zombie.

The Legend of the Gobblewonker 29 Jun. 2012

★ 7.8 (2,963) ☆ Rate

Dipper and Mabel hear rumours that a sea monster lives at the bottom of the local lake, and enlist Soos to take them out on a boat to see if the monster really exists.

Headhunters 30 Jun. 2012

★ 7.9 (2,776) ☆ Rate

The kids discover Stan's collection of wax sculptures, and Mabel decides to create her own. But when someone sabotages it, it's up to her and Dipper to solve the case.

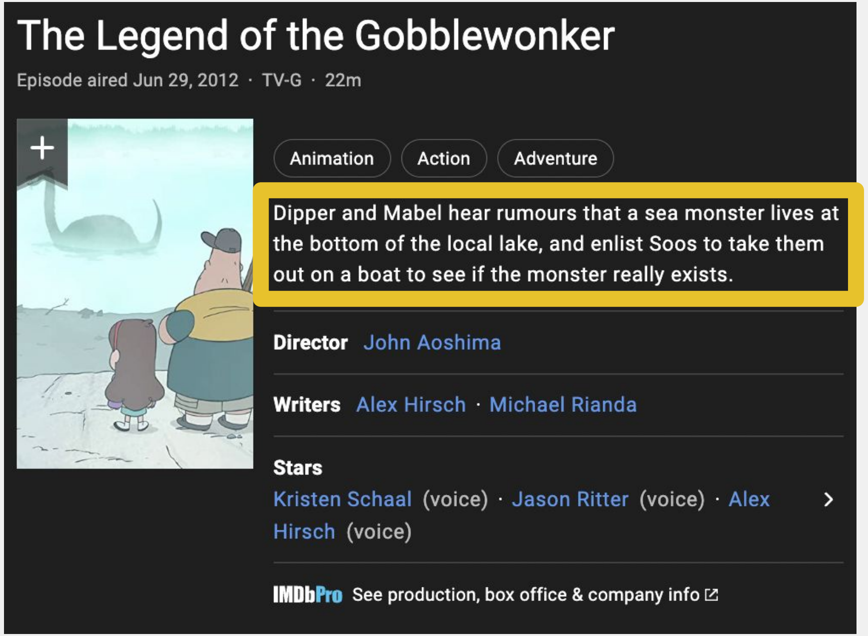
The Hand That Rocks the Mabel 6 Jul. 2012

★ 7.6 (2,663) ☆ Rate

Dipper and Mabel meet an intriguing new neighbor: a psychotic child fraud named Li'l Gideon. When Gideon begins obsessing over Mabel and tries to win her over, Dipper tries to stop him.

I want my program to click the hyperlinks!

Then, scrape the synopsis



The Legend of the Gobblewonker

Episode aired Jun 29, 2012 · TV-G · 22m

Dipper and Mabel hear rumours that a sea monster lives at the bottom of the local lake, and enlist Soos to take them out on a boat to see if the monster really exists.

Director John Aoshima

Writers Alex Hirsch · Michael Rianda

Stars
Kristen Schaal (voice) · Jason Ritter (voice) · Alex Hirsch (voice)

IMDbPro See production, box office & company info

Thanks for Coming!

Contact Us:

Ethan

eleung8628

eleung6@u.rochester.edu

Joyce

joyceeeemeng

jmeng10@u.rochester.edu

Veronica

vronyc_13

vchistay@u.rochester.edu

Irene

iirene.y

iyoo@u.rochester.edu

Amelia

lv.ame

alui8@u.rochester.edu

Instagram

CCC

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**

Graphics by Storyset

Appendix

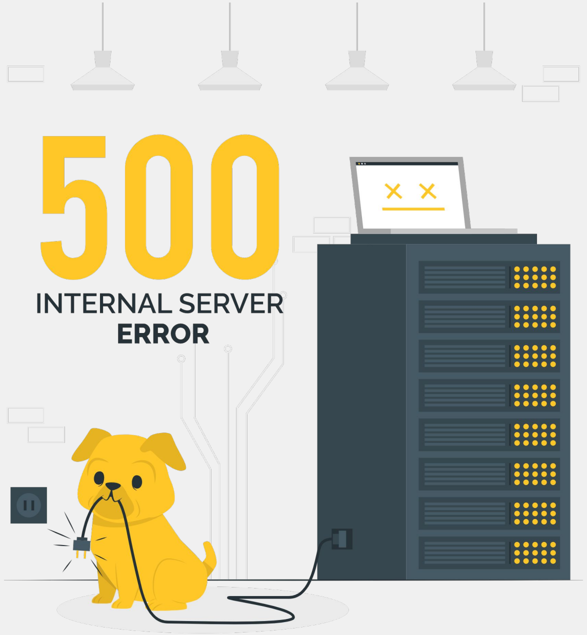
Ethical Webscraping

- The Golden Rule:
  - If the website has an API, use the API
- robots.txt
  - Access by typing in the url of the website, then add “/robots.txt”
  - Guidelines for webscraping that specific website
  - Eg: https://en.wikipedia.org/robots.txt
- User Agent String
- Provide your contact information to the website, make your intentions clear
- Request data at a reasonable rate

HTML Crash Course

- Webscraping-friendly websites have a structured HTML (you'll know what I mean with more practice)
    - Eg: Wikipedia infobox is highly unfriendly
- Tags vs Elements vs Attributes
    - Tags: starting and ending parts of HTML element. Tags hold the Element
        - Everything between < and > are tags
        - Eg: <a></a>
    - Elements: general content within the tags
    - Attributes: describe the elements
        - Found within starting tag
        - Eg: <p align="center"></p>
        - .select(), rather than .find()

You’ve reached the end



(thanks for looking at the appendix if you got so far :))