

# Project 5: Amazon web crawling in Python + REGEX

In this project we use BS- BeautifulSoup and REGEX to find some products by crawling an Amazon.com page

## WE'LL COVER THE FOLLOWING ^

- Solution

A **Web crawler**, sometimes called a spider, is an Internet bot that systematically browses the World Wide Web, typically for the purpose of Web indexing (web spidering). In this project, we crawl the [amazon.com](https://www.amazon.com) website > Movies & TV > 'startrek' (see the image below). Then, we find the list of movies with 'bonus' content.

amazon **Try Prime** Movies & TV startrek prime student 50% off Prime

Departments Browsing History Your Amazon.com Today's Deals Gift Cards & Registry Sell Help

Movies & TV New Releases Best Sellers Deals Blu-ray TV Shows Kids & Family Anime All Genres Amazon Video Prime Video Your Video Library Trade-In

1-16 of 1,261 results for Movies & TV: "star trek" Sort by Relevance

Show results for

Any Department Movies & TV Movies TV

Refine by

International Shipping Ship to Australia

Amazon Prime prime

Eligible for Free Shipping Free Shipping by Amazon

Format Blu-ray DVD Amazon Video Blu-ray 3D See more

Genre Science Fiction Action & Adventure Mystery & Thrillers Drama Kids & Family Comedy Military & War Fantasy Romance Documentary

Actor Leonard Nimoy William Shatner DeForest Kelley

Showing results for star trek. Search instead for startrek. Showing most relevant results. See all results for star trek.

Format: Blu-ray | DVD | Amazon Video | Blu-ray 3D | See more

**Star Trek** 2009 PG-13 CC Amazon Video \$3.99 - \$13.99 Rent or Buy 4.5 stars 94 Starring: John Cho, Ben Cross, Bruce Greenwood, et al. Directed by: J. J. Abrams Runtime: 2 hrs 6 mins

**Star Trek Beyond** 2016 PG-13 CC Amazon Video \$0.00 Watch with a Prime membership \$14.99 - \$19.99 Buy 4.5 stars 2,734 Starring: John Cho, Simon Pegg, Chris Pine, et al. Directed by: Justin Lin Runtime: 2 hrs 2 mins

**Star Trek Into Darkness** 2013 PG-13 CC Amazon Video \$3.99 - \$13.99 Rent or Buy 4.5 stars 14,474 Starring: John Cho, Peter Weller, Simon Pegg, et al. Directed by: J. J. Abrams Runtime: 2 hrs 12 mins

Sponsored

Jason Bourne [Blu-ray] \$12.32 \$26.98 prime 4.5 stars 1,950

Quantum Leap: The Complete Series \$32.98 \$479.98 prime 4.5 stars 246

Zumba Fitness Exhilarate Body Sha... \$32.40 \$39.98 prime 4.5 stars 1,267

Amazon web scraping for Startrek DVD movies with 'bonus' content

## Solution #

The problem solution uses [BeautifulSoup](https://www.crummy.com/software/BeautifulSoup/BeautifulSoup/). A detailed explanation of the code is

out-of-scope for this course (hint: read the BS docs). In this code, we first

extract HTML data and format/convert into BS's table using BS's

`BeautifulSoup()` function, then find and extract the movies from the HTML code.

```
import re
from pprint import pprint
import csv
import requests

import requests
from bs4 import BeautifulSoup
def crawl_amazon_web(page,WebUrl):
    if(page>0):
        url = WebUrl
        code = requests.get(url)
        plain = code.text
        s = BeautifulSoup(plain, "html.parser")

        for link in s.findAll('a', {'class':'s-access-detail-page'}):
            movie_title = link.get('title')
            m = re.search( r'Bonus',movie_title)
            if m:
                print(movie_title)
                html_link = link.get('href')
                print(html_link)

crawl_amazon_web(1,'https://www.amazon.com/s/ref=nb_sb_noss_2?url=search-alias%3Dmovies-tv&fi
```

Expected output (Startrek movies with bonus content) :

```
Star Wars: The Force Awakens (Plus Bonus Features)
https://www.amazon.com/Star-Wars-Force-Awakens-Features/dp/B019EG1TC8
Rogue One: A Star Wars Story (With Bonus Content)
https://www.amazon.com/Rogue-One-Story-Bonus-Content/dp/B01N7FYJ7H
```

Easy!

This solution has been **adopted and extended** from the [Dev.to post](#) written by Pranay Das.

