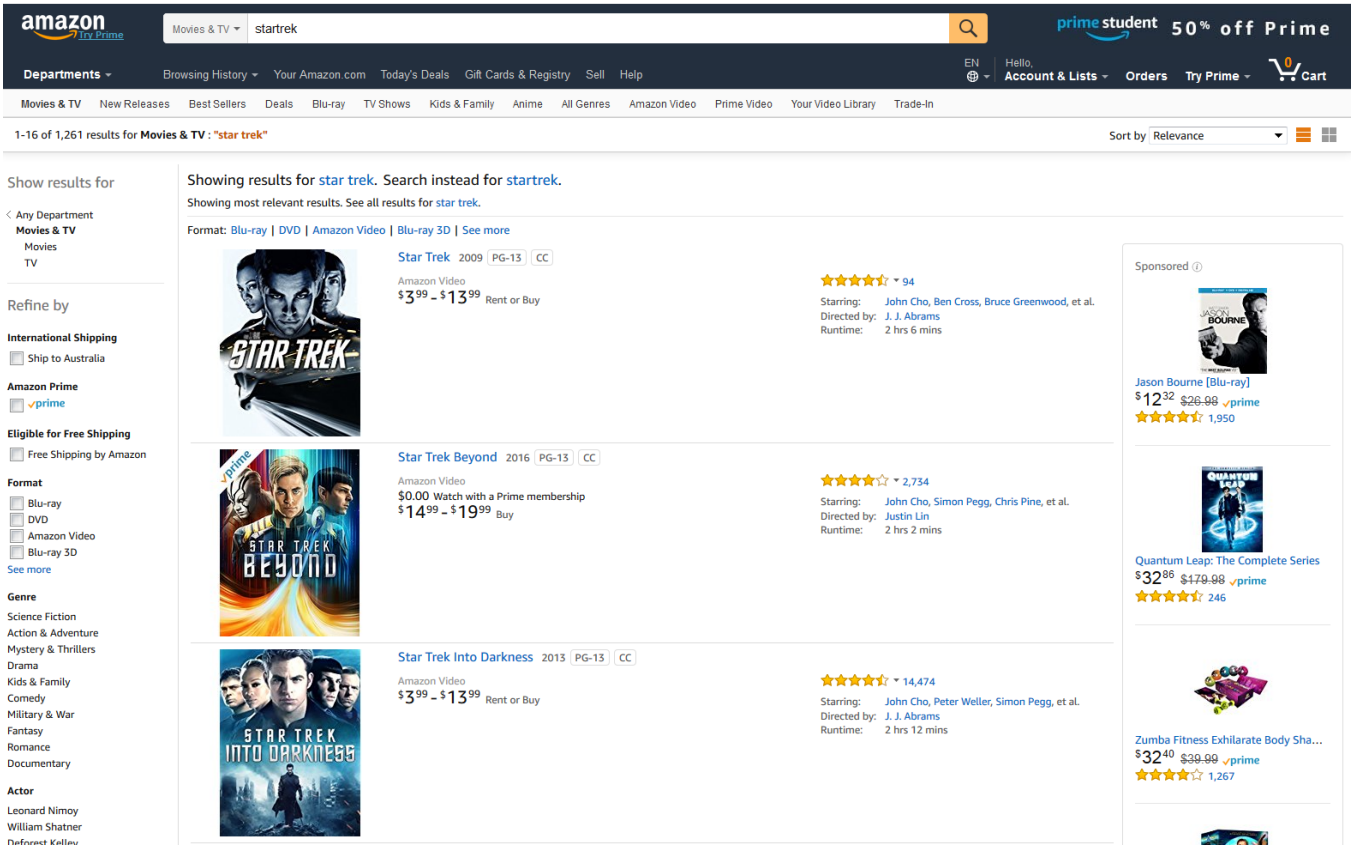
Project 5: Amazon web crawling in Python + REGEX

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 website > Movies & TV > ‘startrek’ (see the image below). Then, we find the list of movies with ‘bonus’ content.



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

Solution

The problem solution uses 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.