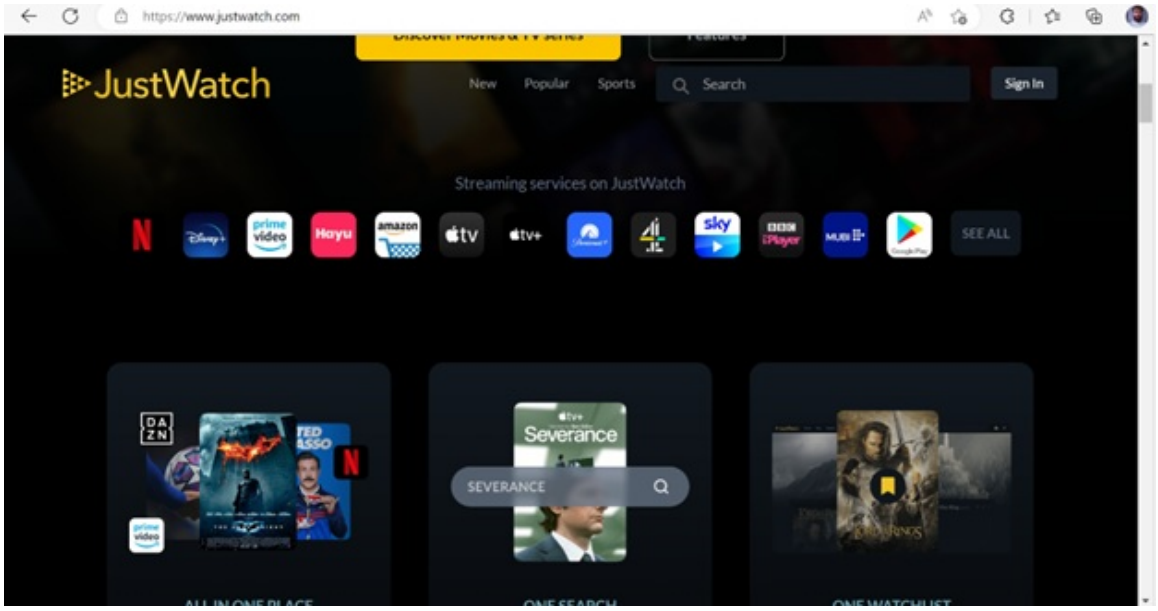
[Justwatch is an online aggregator of all major streaming platforms] (https://www.justwatch.com/)

Instead of scraping all streaming platforms individually which will require more resource/time (e.g., different login credentials) we can simply scrape this website easily and get required parameters



Justwatch Main Page

In [18]:

```
# Justwatch url

justwatch_url = 'https://www.justwatch.com/uk'
```

In [216]:

```
# Necessary imports
import requests
import pandas as pd
from bs4 import BeautifulSoup as bs
import time
from urllib.parse import urljoin
import warnings
from selenium import webdriver
from webdriver_manager.chrome import ChromeDriverManager
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.webdriver.support.ui import Select
warnings.filterwarnings("ignore")
chrome_options = webdriver.ChromeOptions()
# chrome_options.add_argument('headless')
chrome_options.add_experimental_option('excludeSwitches', ['enable-logging'])
chrome_options.add_argument("start-maximized")
chrome_options.add_argument("--incognito")
from selenium.webdriver.common.keys import Keys
```

In [135]:

```
# Different provider url

netflix = '/provider/netflix'
disney_plus = '/provider/hotstar'
iplayer = '/provider/bbc-iplayer'
apple_tv = '/provider/apple-tv-plus'
```

```
amazon = '/provider/amazon-prime-video'
```

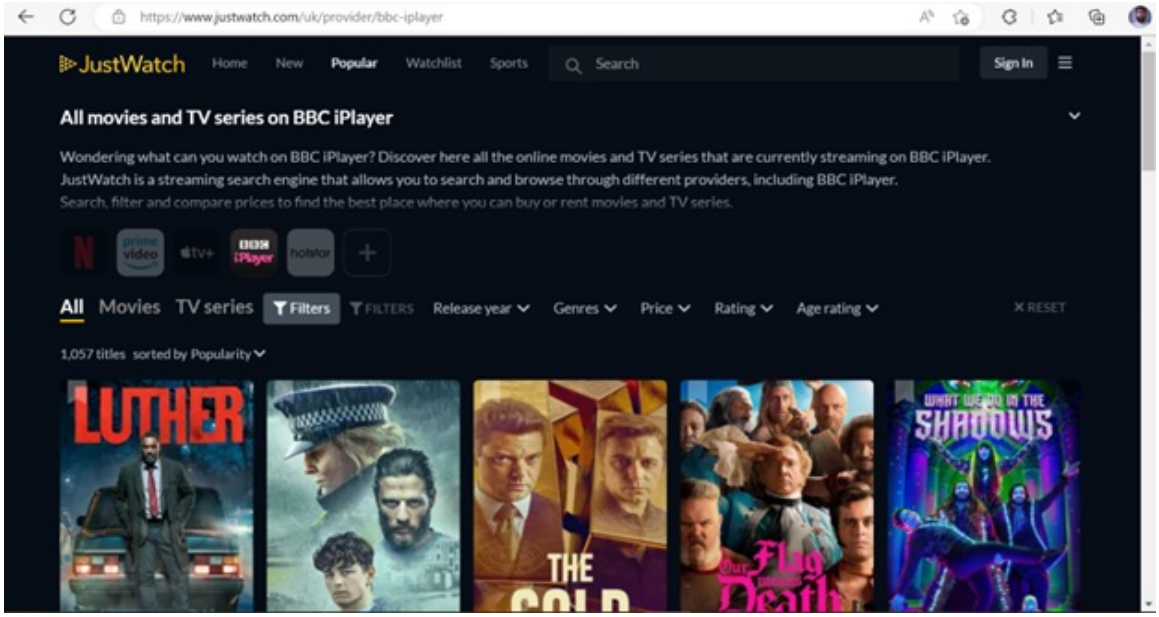
In [55]:

```
# Let us say we want to extract information of the bbc_iplayer source

iplayer_source_url = justwatch_url + iplayer
iplayer_source_url
```

Out[55]:

```
'https://www.justwatch.com/uk/provider/bbc-iplayer'
```



BBC Iplayer Page

In [152]:

```
# Install Driver

driver = webdriver.Chrome(ChromeDriverManager().install(), chrome_options=chrome_options)
```

In [142]:

```
# Open url in driver
driver.get(iplayer_source_url)
```

In [138]:

```
# Scroll to bottom to load all movies/shows

last_height = driver.execute_script("return document.body.scrollHeight")
while True:
    # Scroll down to bottom
    driver.execute_script("window.scrollTo(0, document.body.scrollHeight);")
    # Wait to load page
    time.sleep(2)
    # Calculate new scroll height and compare with last scroll height
    new_height = driver.execute_script("return document.body.scrollHeight")
    if new_height == last_height:
        break
    last_height = new_height
```

In [139]:

```
# Get tags of all movies/shows

soup = bs(driver.page_source, 'lxml')
all_movies = soup.findAll('div', class_='title-list-grid__item')
print(len(all_movies))
```

1049

In [141]:

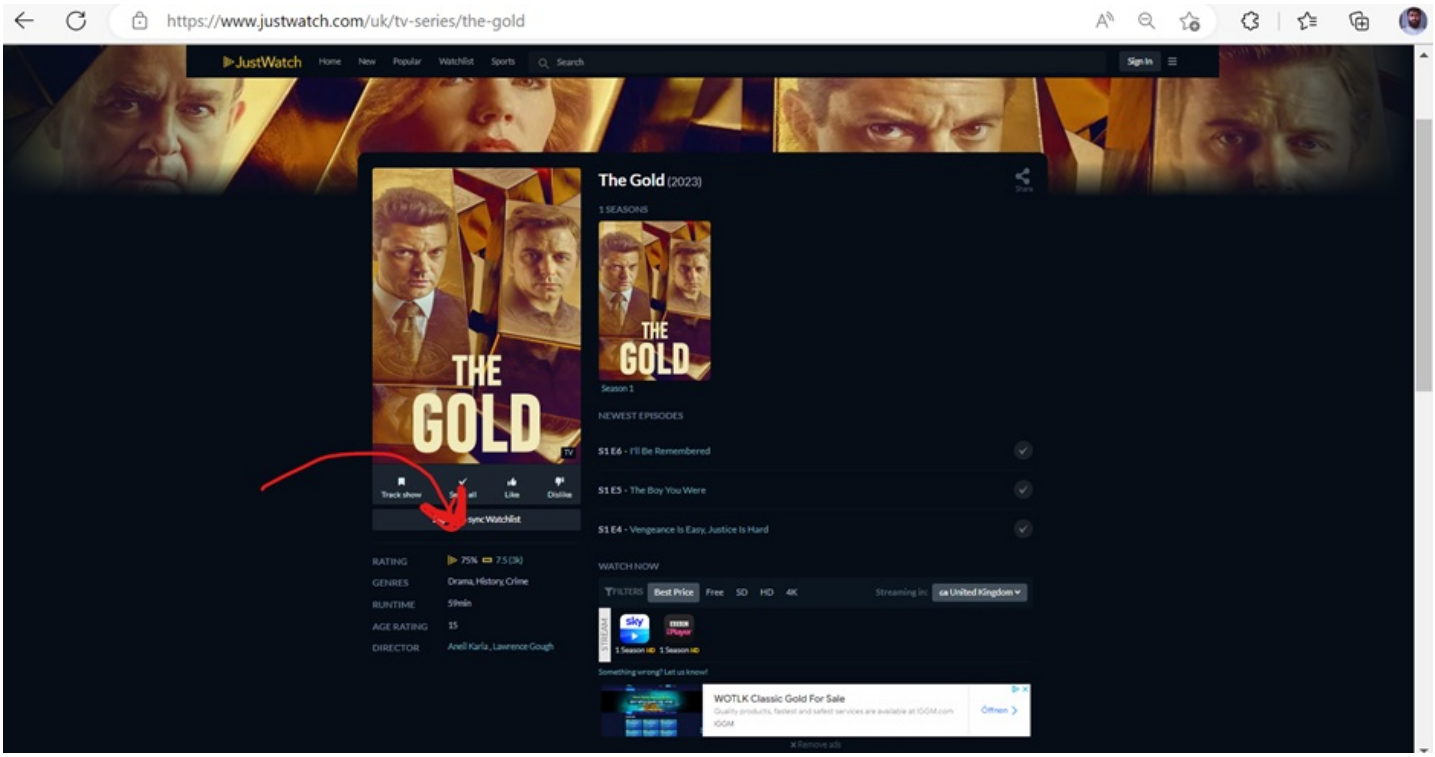
```
# Collect all movies links

all_movies_links = []
for movie in all_movies:
    all_movies_links.append('https://www.justwatch.com' + movie.find('a')['href'])

all_movies_links[:5]
```

Out[141]:

```
['https://www.justwatch.com/uk/tv-series/luther',
 'https://www.justwatch.com/uk/tv-series/happy-valley',
 'https://www.justwatch.com/uk/tv-series/the-gold',
 'https://www.justwatch.com/uk/tv-series/our-flag-means-death',
 'https://www.justwatch.com/uk/tv-series/what-we-do-in-the-shadows']
```



The Gold on BBC Iplayer (Taking IMDB URL from Justwatch)

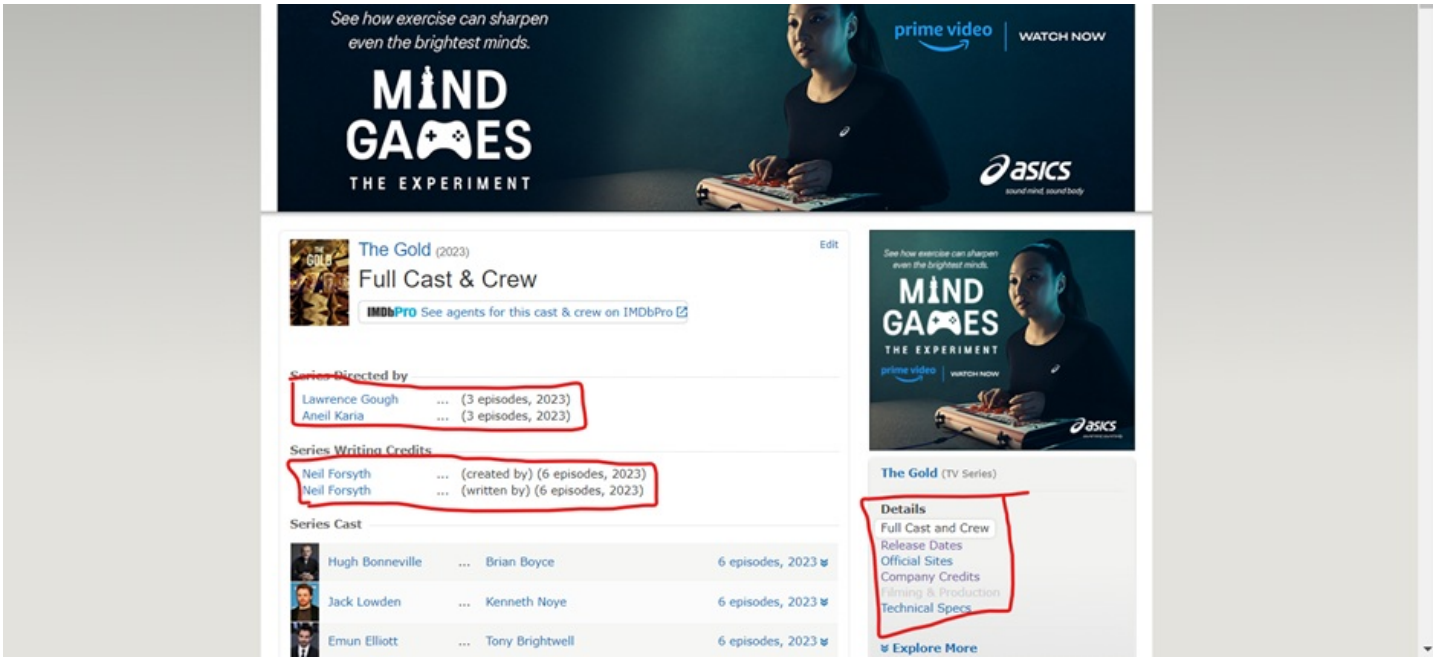
Since justwatch has limited information about movies/shows we are extracting imdb url from justwatch and extracting full information from justwatch

In [129]:

```
for movie_link in all_movies_links[:5]:
    res = requests.get(movie_link)
    soup = bs(res.content, 'lxml')
    imdb_link = soup.find('div', {'v-uib-tooltip': 'IMDB'}).find('a')['href']
    print(imdb_link)
```

```
https://www.imdb.com/title/tt1474684/?ref_=justwatch
https://www.imdb.com/title/tt3428912/?ref_=justwatch
https://www.imdb.com/title/tt14063678/?ref_=justwatch
https://www.imdb.com/title/tt11000902/?ref_=justwatch
https://www.imdb.com/title/tt7908628/?ref_=justwatch
```

Extracting Information from IMDB URL for *"The Gold"*



In [229]:

```
the_gold_justwatch_url = 'https://www.justwatch.com/uk/tv-series/the-gold'
res = requests.get(the_gold_justwatch_url)
soup = bs(res.content, 'lxml')

# Create data dict

data = {
    'Source': 'BBC Iplayer',
    'Movie/Show': '',
    'Parameter Name': '',
    'Parameter Value': ''

}
director = soup.find('div', text="Director").findNextSibling().text.strip()
age_rating = soup.find('div', text="Age rating").findNextSibling().text.strip()
all_streaming_platforms = ', '.join([stream['title'] for stream in soup.find('div', clas
s_='price-comparison__grid__row__holder').findAll('img')])
the_gold_imdb_url = 'https://www.imdb.com/title/tt14063678/?ref_=justwatch'
full_credits_url = f"{the_gold_imdb_url.split('?')[0]}fullcredits?ref_=tt_ov_st_sm"
print(full_credits_url)
```

```
https://www.imdb.com/title/tt14063678/fullcredits?ref_=tt_ov_st_sm
```

In [204]:

```
# Data Extraction

driver.get(the_gold_imdb_url)
time.sleep(2)
soup = bs(driver.page_source, 'lxml')
# print(soup.prettify())
```

In [214]:

```
title = soup.find('h1', class_='sc-b73cd867-0 cEmnhL').text.strip()
imdb_rating = soup.find('span', class_='sc-e457ee34-1 gvYTvP').text.strip()
episodes  = soup.find('div', {'data-testid': 'episodes-header'}).find('span', class_='ip
c-title__subtext').text.strip()
top_stars = ', '.join([star.text for star in soup.find('a', text='Stars').findNextSiblin
g().findAll('a')])
top_casts = ', '.join([' - '.join([a.text for a in cast.findAll('a')]) for cast in soup.
findAll('div', class_='sc-bfec09a1-7 dpBDvu')])
creator = soup.find('span', text='Creator').findNextSibling().text.strip()
genres = ', '.join([genre.text for genre in soup.find('span', text='Genres').findNextSib
ling().findAll('a')])
release_date = soup.find('a', text='Release date').findNextSibling().text.split('(')[0].
strip()
```

```
release_country = soup.find('span', text='Country of origin').findNextSibling().text.strip()
language = soup.find('span', text='Language').findNextSibling().text.strip()
production_company = soup.find('a', text='Production company').findNextSibling().text.strip()
runtime = soup.find('span', text='Runtime').findNextSibling().text.strip()
plot_summary = soup.find('div', {'data-testid': 'storyline-plot-summary'}).text.strip()
similar_shows = ', '.join([show.text for show in soup.find('section', {'data-testid': 'MoreLikeThis'}).findAll('span', {'data-testid': 'title'})])
```

In [247]:

```
data['Movie/Show'] = title
data['Parameter Name'] = [
    'Director', 'Age Rating', 'All Streaming Platforms', 'IMDB Rating', 'Episodes',
    'Top Stars', 'Top Casts', 'Creator', 'Genres', 'Release Date', 'Release Country',
    'Language', 'Production Company', 'Runtime', 'Plot Summary', 'Similar Shows']
data['Parameter Value'] = [director, age_rating, all_streaming_platforms,
                           imdb_rating, episodes, top_stars, top_casts, creator,
                           genres, release_date, release_country, language, production_company,
                           runtime, plot_summary, similar_shows]
data
```

Out[247]:


```
{'Source': 'BBC Iplayer',
 'Movie/Show': 'The Gold',
 'Parameter Name': ['Director',
  'Age Rating',
  'All Streaming Platforms',
  'IMDB Rating',
  'Episodes',
  'Top Stars',
  'Top Casts',
  'Creator',
  'Genres',
  'Release Date',
  'Release Country',
  'Language',
  'Production Company',
  'Runtime',
  'Plot Summary',
  'Similar Shows'],
 'Parameter Value': ['Aneil Karia ,  Lawrence Gough',
  '15',
  'Sky Go, BBC iPlayer',
  '7.5',
  '6',
  'Hugh Bonneville, Jack Lowden, Emun Elliott',
  'Hugh Bonneville - Brian Boyce, Jack Lowden - Kenneth Noye, Emun Elliott - Tony Brightwell, Charlotte Spencer - Nicki Jennings, Tom Cullen - John Palmer, Stefanie Martini - Marnie Palmer, Sean Harris - Gordon Parry, Dominic Cooper - Edwyn Cooper, Amanda Drew - CS Cath McLean, Sean Gilder - DI Neville Carter, Daniel Ings - Archie Osborne, Nichola Burley - Brenda Noye, Silas Carson - Harry Bowman, Peter Davison - Assistant Commissioner Gordon Stewart, Paul Thornley - Max Goodman, Ellora Torchia - Sienna Rose, Adam Nagaitis - Micky McAvoy, Dorothy Atkinson - Jeannie Savage',
  'Neil Forsyth',
  'Crime, Drama, History',
  'February 12, 2023',
  'United Kingdom',
  'English',
  'ViacomCBS International Studios',
  '58 minutes',
  "Drama series inspired by true events surrounding the 1983 Brink's-Mat robbery, and the remarkable story that followed. The Gold takes a pulsating journey into a 1980's world awash with cheap money and loosened morals to tell an extraordinary and epic tale.",
  'Better, Happy Valley, Funny Woman, Bank of Dave, Unforgotten, Stonehouse, Nolly, Rogue Heroes, A Spy Among Friends, Desperate Measures, C.B. Strike, Luther: The Fallen Sun']}
```


In [248]:

```
df = pd.DataFrame(data)
df
```

Out[248]:

	Source	Movie/Show	Parameter Name	Parameter Value
0	BBC Iplayer	The Gold	Director	Aneil Karia , Lawrence Gough
1	BBC Iplayer	The Gold	Age Rating	15
2	BBC Iplayer	The Gold	All Streaming Platforms	Sky Go, BBC iPlayer
3	BBC Iplayer	The Gold	IMDB Rating	7.5
4	BBC Iplayer	The Gold	Episodes	6
5	BBC Iplayer	The Gold	Top Stars	Hugh Bonneville, Jack Lowden, Emun Elliott
6	BBC Iplayer	The Gold	Top Casts	Hugh Bonneville - Brian Boyce, Jack Lowden - K...
7	BBC Iplayer	The Gold	Creator	Neil Forsyth
8	BBC Iplayer	The Gold	Genres	Crime, Drama, History
9	BBC Iplayer	The Gold	Release Date	February 12, 2023
10	BBC Iplayer	The Gold	Release Country	United Kingdom
11	BBC Iplayer	The Gold	Language	English
12	BBC Iplayer	The Gold	Production Company	ViacomCBS International Studios
13	BBC Iplayer	The Gold	Runtime	58 minutes
14	BBC Iplayer	The Gold	Plot Summary	Drama series inspired by true events surroundi...
15	BBC Iplayer	The Gold	Similar Shows	Better, Happy Valley, Funny Woman, Bank of Dav...

In [277]:

```
df = df.set_index(['Source', 'Movie/Show', 'Parameter Name'])
df
```

Out[277]:

			Parameter Value
Source	Movie/Show	Parameter Name	
BBC Iplayer	The Gold	Director	Aneil Karia , Lawrence Gough
		Age Rating	15
		All Streaming Platforms	Sky Go, BBC iPlayer
		IMDB Rating	7.5
		Episodes	6
		Top Stars	Hugh Bonneville, Jack Lowden, Emun Elliott
		Top Casts	Hugh Bonneville - Brian Boyce, Jack Lowden - K...
		Creator	Neil Forsyth
		Genres	Crime, Drama, History
		Release Date	February 12, 2023
		Release Country	United Kingdom
		Language	English
		Production Company	ViacomCBS International Studios
		Runtime	58 minutes
		Plot Summary	Drama series inspired by true events surroundi...
		Similar Shows	Better, Happy Valley, Funny Woman, Bank of Dav...