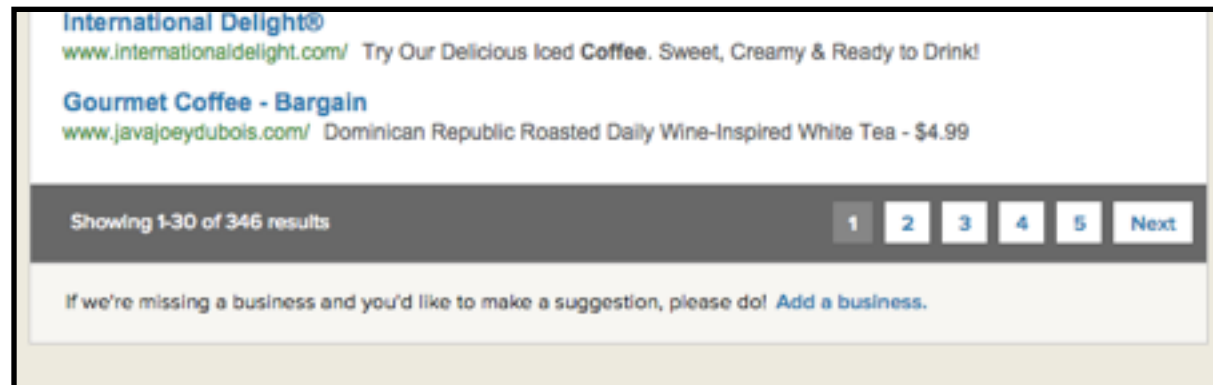
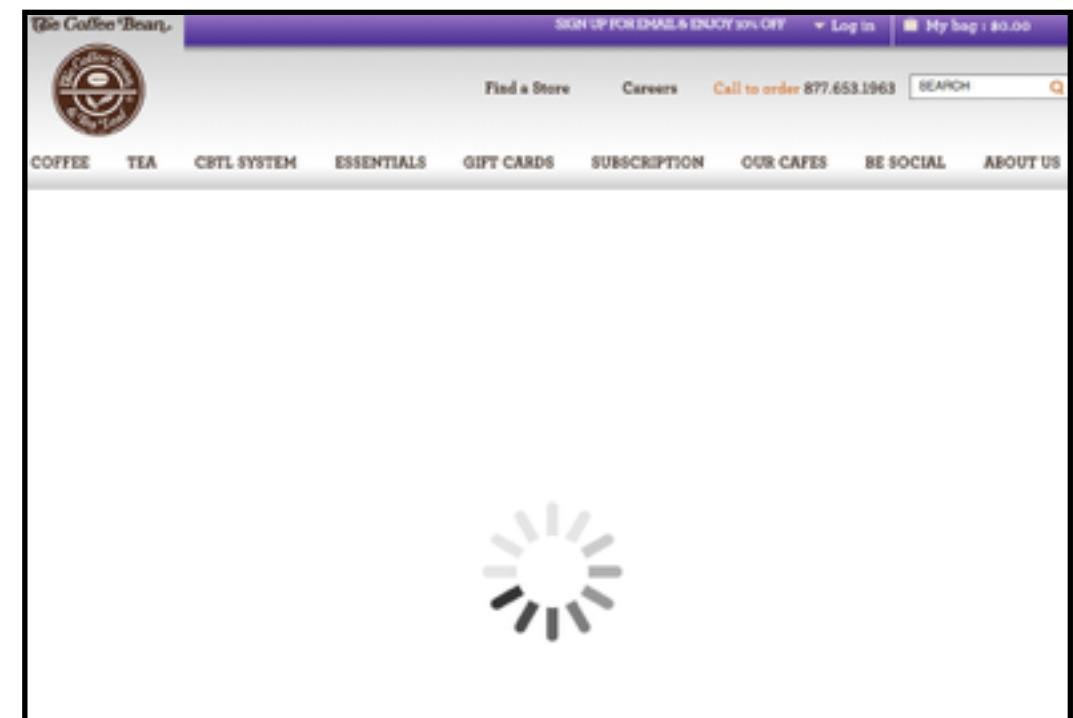


Some Obstacles



Results on Multiple Pages



Dynamic Web Page

Results on Multiple Pages

Today's **GOAL**

Collect the **cast list** of **Top 250 comedy movies** sorted based on US Box Office Income

(http://www.imdb.com/search/title?genres=comedy&sort=boxoffice_gross_us)

Pagination

Sometimes data we are interested in is on multiple pages. For instance, IMDB list of top comedy movies.

49.



22 Jump Street (2014)

Add to Watchlist

\$192M

★★★★★☆☆☆☆ 7.2/10

After making their way through high school (twice), big changes are in store for officers Schmidt and Jenko when they go deep undercover at a local college.
Dir: [Phil Lord](#), [Christopher Miller](#) With: [Channing Tatum](#), [Jonah Hill](#), [Ice Cube](#)
[Action](#) | [Comedy](#) | [Crime](#)

112 mins. 

50.



Cars 2 (2011)

Add to Watchlist

\$191M

★★★★★☆☆☆☆ 6.4/10

Star race car Lightning McQueen and his pal Mater head overseas to compete in the World Grand Prix race. But the road to the championship becomes rocky as Mater gets caught up in an intriguing adventure of his own: international espionage.
Dir: [John Lasseter](#), [Brad Lewis](#) With: [Owen Wilson](#), [Larry the Cable Guy](#), [Michael Caine](#)
[Animation](#) | [Adventure](#) | [Comedy](#) | [Family](#)

106 mins. 

1-50 of 705,050 titles.

Next »

http://www.imdb.com/search/title?genres=comedy&sort=boxoffice_gross_us

http://www.imdb.com/search/title?genres=comedy&sort=boxoffice_gross_us&start=51

http://www.imdb.com/search/title?genres=comedy&sort=boxoffice_gross_us&start=101

Structure of URL

http://www.imdb.com/search/title?genres=comedy&sort=boxoffice_gross_us&start=51

List of comedy movies

Sorted by US box office gross income

Offset = 51

Change the offset and you'll have different pages of the list.

```
url_base = 'http://www.imdb.com/search/title?genres=comedy&sort=boxoffice_gross_us'
for i in range(1,250,50):
    # Concatenate two parts of the URL string
    url = url_base + '&start=' + str(i)
    response = requests.get(url)
    ...
```

[python]

Next

Today's **GOAL**

Collect **comments** of a news article on **The Guardian**
(<http://www.theguardian.com/technology/2015/jan/28/artificial-intelligence-will-not-end-human-race>)

Extract Comments

Artificial intelligence: how clever do we want our machines to be?

29 Nov 2014 338

Elon Musk: artificial intelligence is our biggest existential threat

27 Oct 2014 673

Google buys two more UK artificial intelligence startups


23 Oct 2014 27

comments (133)

This discussion is closed for comments.

Order by Oldest Threads Collapsed


1 2 3

ubiktd Jon 3d ago

i'm not sure i trust a company that collaborated with the government to put back doors in it's systems so that they could spy on everyone.

2 ↑


Report

popcornmaster 3d ago

This debate between 'will destroy' and 'won't destroy' will be used as the opening montage to a feel good historical robot film, when Skynet inevitably rises up. Starting the robot equivalent of Tom Cruise.

2 ↑


Report

popcornmaster → popcornmaster 3d ago

Staring* god damn it

1 ↑


Report

AlexAnder1 → popcornmaster 3d ago

Starring*

1 ↑

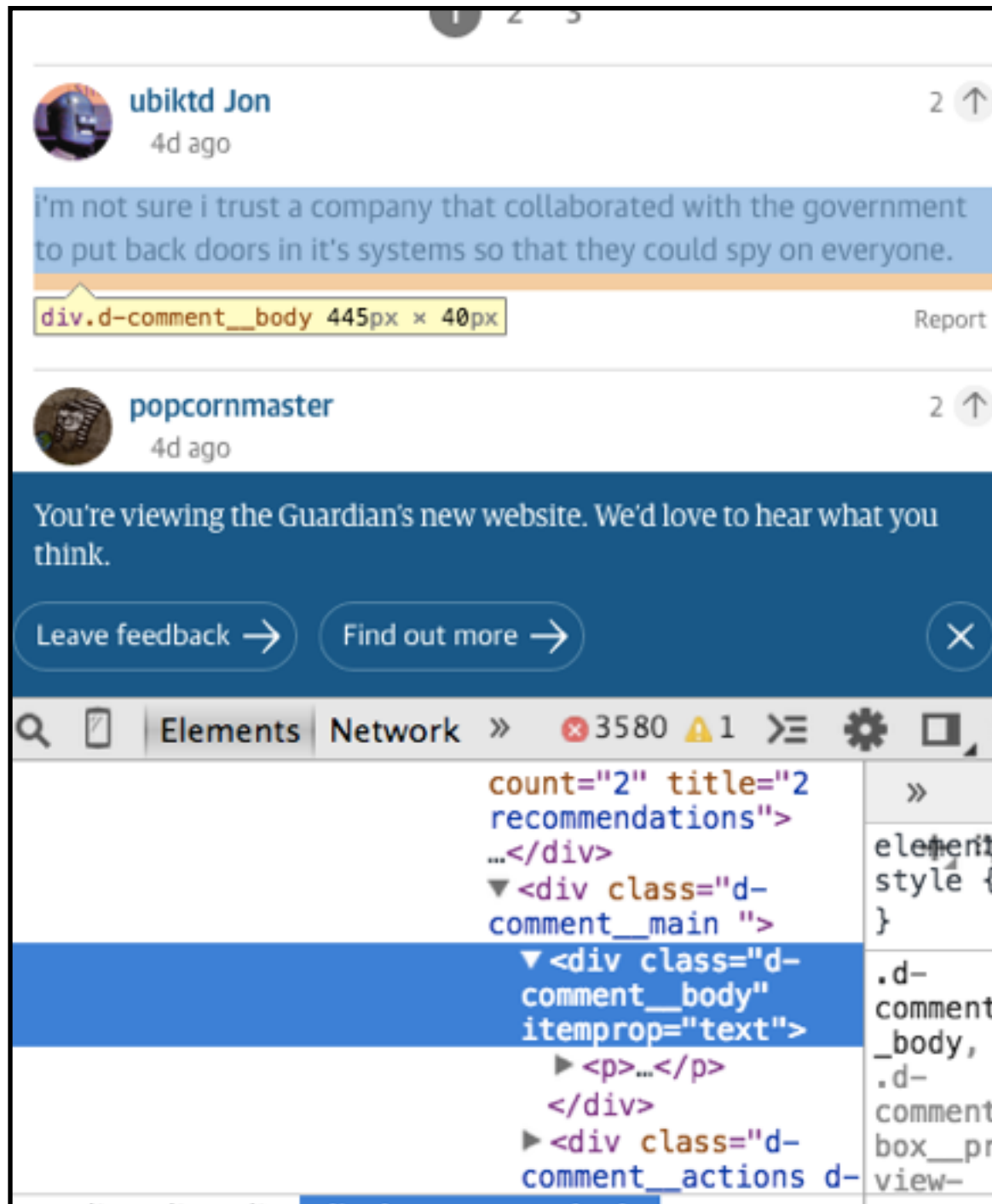
Report

You're viewing the Guardian's new website. We'd love to hear what you think.

Find out more →

Leave feedback →

Our Old Ways



```
url_base = 'http://www.theguardian.com/technology/
2015/jan/28/artificial-intelligence-will-not-end-human-
race'
```

```
response = requests.get(url)
```

```
soup = BeautifulSoup(response.content)
```

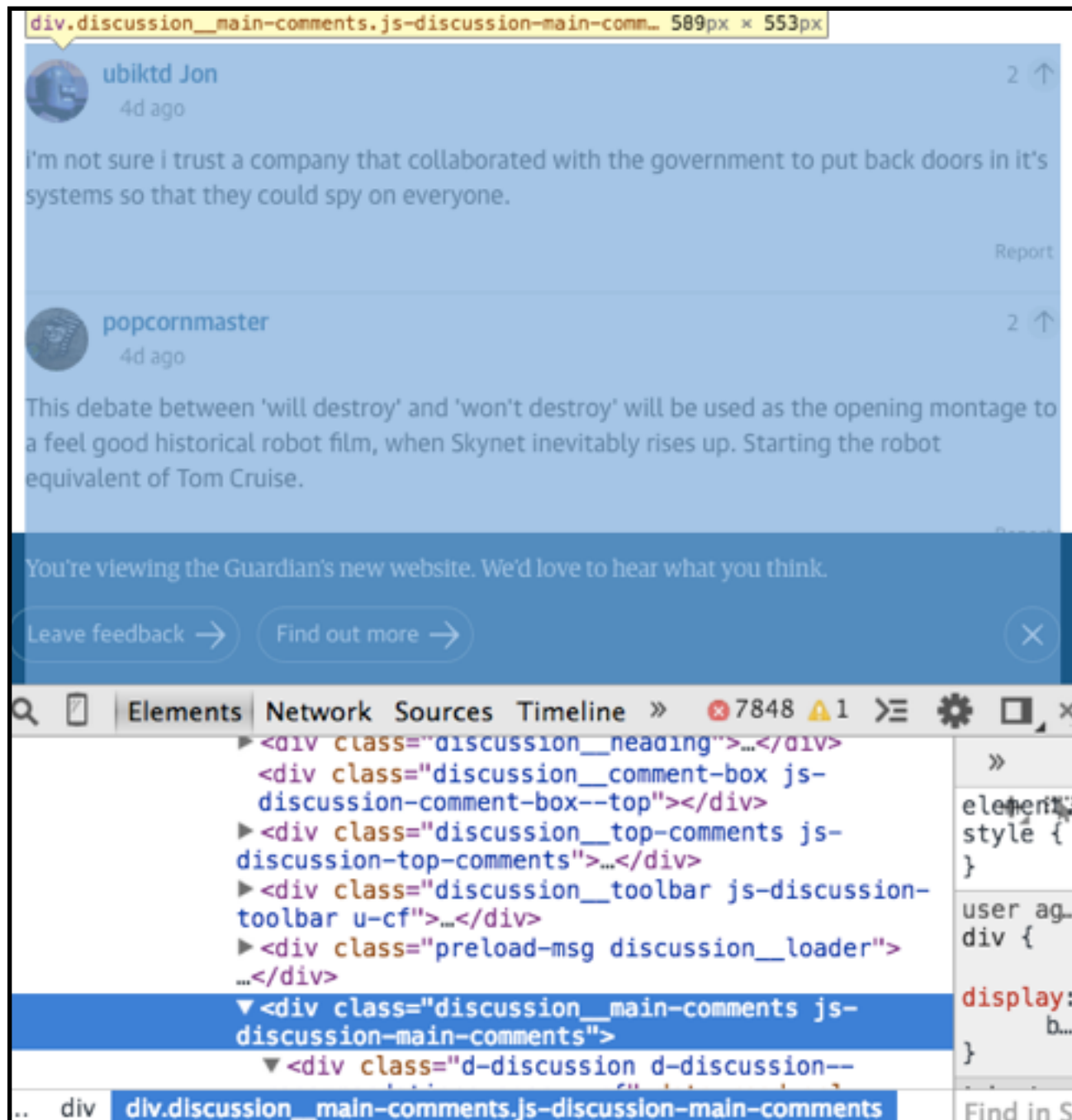
```
mydivs = soup.findAll('div', class_ = 'd-
comment__body', limit = 1)
```

[python]

mydivs will be empty!

Compare

Inspect Element



Page Source

```
pagination"></div>
</div>
<div class="preload-msg
discussion_loader">Loading comments... <a
href="/discussion/p/45a9a class="accessible-
link">Trouble loading?</a><div class="is-
updating"></div></div>
<div class="discussion_main-comments js-
discussion-main-comments"></div>
<div class="discussion__comment-box
discussion__comment-box--bottom js-discussion-
comment-box--bottom"></div>
<button class="discussion__show-button button-
show-more button button--large button--
primary js-discussion-show-button" data-link-
name="more-comments">
<i class="i i-plus-white"></i>
View more comments
</button>
```

The element containing comments is empty in the source code!

Dynamic Content

Inspect Element

Shows the page content after scripts are run on the server side.

Page Source

Shows the initial content of the page before scripts are run.

Requests gets the initial content of the page, same as what we see from Page Source.

Solution:

Use Selenium

Selenium

Python library to mechanize the browser.

```
from selenium import webdriver  
browser = webdriver.Firefox()
```

```
url = 'http://www.imdb.com'  
browser.get(url)
```

[python]



Opens a FireFox window,
loads IMDB website.

Scrape Dynamic Web Pages

When using **Selenium**, we can wait for a certain element to load, or wait for a certain amount of time, and then look for the data we want to extract in the Page.

```
url_base = 'http://www.theguardian.com/technology/2015/jan/28/artificial-intelligence-will-not-end-human-race'
```

```
browser.get(url)
```

```
time.sleep(5)
```

```
page_content = browser.page_source
```

[python]

`page_content` now contains the comments.

Also Consider API's

API (Application Program Interface):

Provides programmatic access to a website's data.

- + **Structured Data**
- + **Easier than web scraping**
- **You have to register for a key and use it with every request, so not anonymous**
- **Limited rate and access**