Web Scraping

Presented by Trixia Belleza

what is web scraping?

Fetching data from websites

why web scrape?

Big Data Analytics

amazon

★★★★★ **FITS PERFECTLY ON ON MY SIDEWALK!**

By Amazon Customer on January 7, 2017

This speaker fits perfectly on my sidewalk, because I sold my house to buy these.

★★★★★ **Whoa Baby**

By Patrick J. Hawkins on December 18, 2015

I bought these to put in our newborn's nursery. We've been playing classical music because it is suppose to stimulate brain functionality. It seems to work stupendously. The child is only 8 weeks old and has told me that I'm a freaking moron for buying them.

★★☆☆☆ **Misleading product name...**

By Obi Wan TOP 100 REVIEWER on February 16, 2016

I was quite despondent to see them shipped in simple wooden and cardboard boxes as if these were being sent to peasants. They should have been packed in hand stained mahogany shipping containers and wrapped tightly in baby panda fur for softness.

”

The uses of web scraping for business and personal requirements are endless.

“

https://www.webharvy.com/articles/web-scraper-use-cases.html

How to web scrape?

In this seminar, we will use Python Scrapy

SETTING UP

Python Scrapy

PYTHON

A high-level programming language

PIP

PIP

PIP is a package manager for Python packages.

SCRAPY

Scrapy is a free and open-source web-scraping framework written in Python

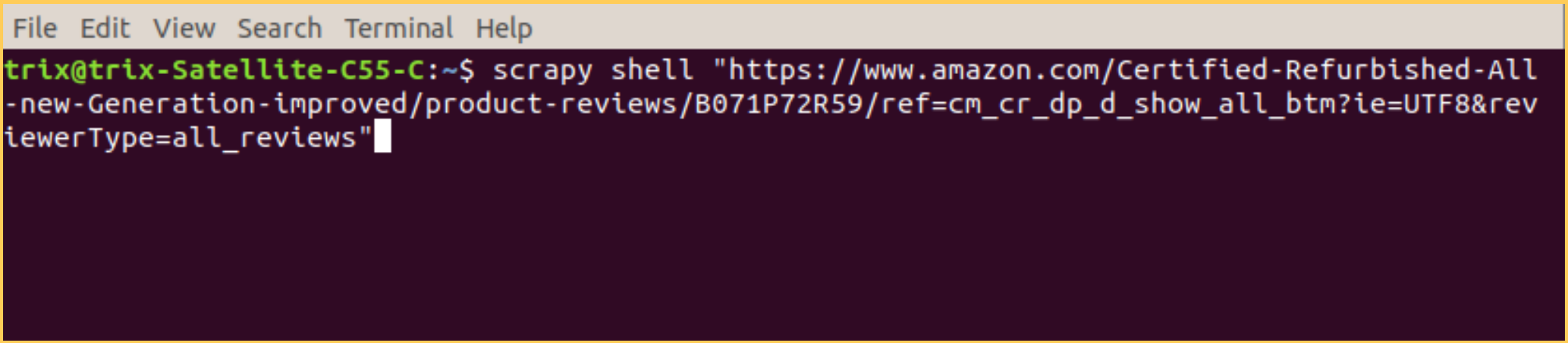
Amazon Product Reviews

Return a list of text reviews in a page.

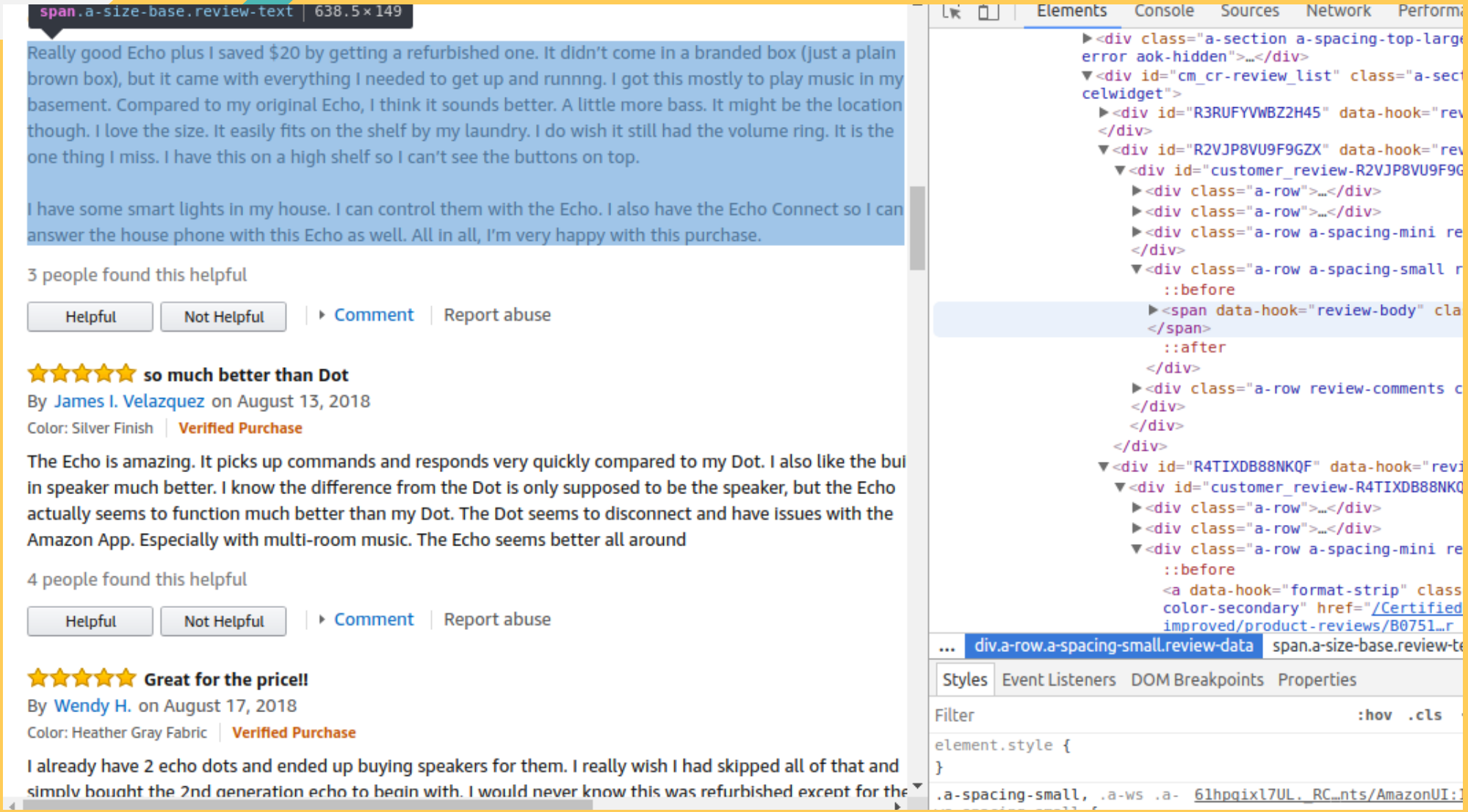
Store the text reviews of a page in a list.

In the terminal, write:

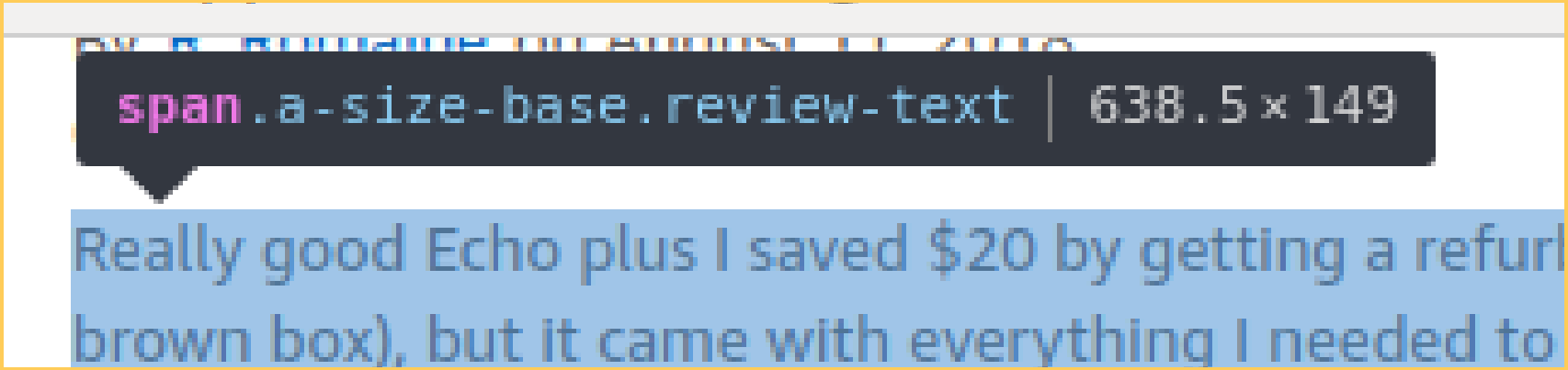
scrapy shell "<url of the page>"



Find the html element and class that defines a specific text review



Find the html element and class that defines a specific text review



Element span

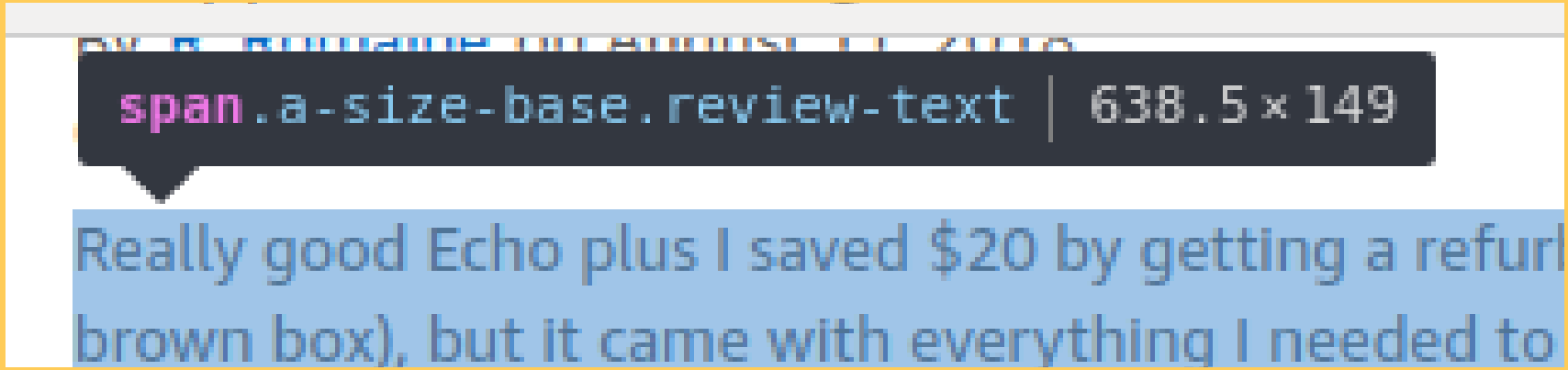
Class a-size-base review-text

In the scrapy shell, type:

response.xpath('//<element>[contains(@class, "<name of class>")]')

XPATH

XPATH IS A LANGUAGE FOR SELECTING NODES IN XML OR HTML DOCUMENTS



Element span

Class a-size-base review-text

```
>>>
>>> response.xpath('//span[contains(@class,"a-size-base review-text")]')
```

Select nodes from all span elements that contains the class "a-size-base review-text"

Returns a list of selector objects

```
>>>
>>> response.xpath('//span[contains(@class,"a-size-base review-text")]')
[<Selector xpath='//span[contains(@class,"a-size-base review-text")]' data=u'<sp
an data-hook="review-body" class="a-s'>, <Selector xpath='//span[contains(@class
,"a-size-base review-text")]' data=u'<span data-hook="review-body" class="a-s'>,
 <Selector xpath='//span[contains(@class,"a-size-base review-text")]' data=u'<sp
an data-hook="review-body" class="a-s'>, <Selector xpath='//span[contains(@class
,"a-size-base review-text")]' data=u'<span data-hook="review-body" class="a-s'>,
 <Selector xpath='//span[contains(@class,"a-size-base review-text")]' data=u'<sp
an data-hook="review-body" class="a-s'>, <Selector xpath='//span[contains(@class
,"a-size-base review-text")]' data=u'<span data-hook="review-body" class="a-s'>,
 <Selector xpath='//span[contains(@class,"a-size-base review-text")]' data=u'<sp
an data-hook="review-body" class="a-s'>, <Selector xpath='//span[contains(@class
,"a-size-base review-text")]' data=u'<span data-hook="review-body" class="a-s'>,
 <Selector xpath='//span[contains(@class,"a-size-base review-text")]' data=u'<sp
an data-hook="review-body" class="a-s'>, <Selector xpath='//span[contains(@class
,"a-size-base review-text")]' data=u'<span data-hook="review-body" class="a-s'>]
```

Getting the text data from these selectors

In the scrapy shell, type:
response.xpath('//<element>[contains(@class, "<name of class>")]/text()')

Getting the text data from these selectors

```
>>>
>>> response.xpath('//span[contains(@class, "a-size-base
[<Selector xpath='//span[contains(@class, "a-size-base r
th='//span[contains(@class, "a-size-base review-text")]/
tains(@class, "a-size-base review-text")]/text()' data=u
"a-size-base review-text")]/text()' data=u'This arrived
eview-text")]/text()' data=u'It is difficult to get the
text()' data=u'I thought I was buying an original Vizio'
'Responds better than the original remote'>, <Selector x
ht out of the box, so I hav'>, <Selector xpath='//span[c
my 3rd remote'>, <Selector xpath='//span[contains(@class
>, <Selector xpath='//span[contains(@class, "a-size-base
>>>
```

Getting the text data from these selectors

Notice that the list actually contains data reviews

Finally, extracting the data

In the scrapy shell, type:
response.xpath('//<element>[contains(@class, "<name of class>")]/text()').extract()

Finally, extracting the data

```
>>>
>>> response.xpath('//span[contains(@class, "a-size-base review-text")]/text()'
[u"I have an older Vizio Tv.  I had thrown the  original  remote out a few year
red of the prices of cable and satellite and switched to Amazon Prime and Sling
y inexpensive  and just what I needed to turn the TV on and off and control vol
egardless  of age.  And connecting - easy, just a few questions on the screen a
e, Hate the remote that came with it.  You have to push the buttons like 5 time
 least once a month.  I've really avoided using it at all costs, but have recen
 switch the input.  So, thank goodness for this replacement!", u'*edited to add
 with it, I do have a complaint.  The buttons are so freaking loud when you pus
ck, click, click, click".  The loud button clicks have not eased as it is "brok
my teenager click click clicking through YouTube, to the point that I will like
d with my set straight out of the box and came with batteries. What more could
tual set anymore, I\u2019m just glad mine can be used again!', u'It is difficul
how to do so. The battery compartment is in the bottom front of the remote. It
 an original Vizio remote control (it was not).  The material seems inferior to
the original.  Out the box, it works like the original.  I did not have to set
 now.", u'Responds better than the original remote ever did. My only complaint
s may fade over time. Good replacement for Vizio remote.', u"It worked right ou
 the remote.. BUT it feels very cheap, light, and the button press is cringe wo
f my wife is sleeping at night, there is not a chance I can use this remote in
emote after 7pm lol.. so I am Definitely buying a new one, and keeping this one
remote in 4 years. They just don't last long enough but what remote does when y
n I need to replace something that would normally be tough to find.  Our remote
g it around a bit (hahaha).  I went on Amazon and did a search and \u2018wala\u
```

Assign it to a list

In the scrapy shell, type:

```
reviews = response.xpath('//<element>[contains(@class, "<name of class>")]/text()').extract()
```

Then access each review by index:

```
reviews[i]
```

reviews[1]

```
>>> reviews[1]
u"Yay!  It works with my TV.  Hate, Hate, Hate the remote that came with it.
nd I've had to replace the batteries at least once a month.  I've really avo
estick on this TV and have to use it to switch the input.  So, thank goodnes
>>>
>>>
>>>
```

”

You can have data without information, but you cannot have information without data.

“

Daniel Keys Moran Quotes. (n.d.). BrainyQuote.com. Retrieved September 14, 2018, from BrainyQuote.com Web site: https://www.brainyquote.com/quotes/daniel_keys_moran_230935

Presented by:
Trixia Belleza
CMSC 199 - B2