

Data Ingestion and Big Data

Build a dataset from zero to solid

Mauro Pelucchi

mauro.pelucchi@gmail.com

<https://github.com/mauropelucchi>

<https://www.linkedin.com/in/mauropelucchi/>



Description

Web scraping, crawling and API are the first step to retrieve information to use for analysis and to start a new business.

In this tutorial I'll show you how to use python to set up scraping and crawling processes, how to simulate users navigation and browser behavior with a ghost browser and how to hook up and use data APIs. I will also try to explain the technical and ethical aspects that we have to consider when we approach these kinds of challenges..

You are provided a git repository (<https://github.com/mauropelucchi/pyconsk2022>) with:

- Jupyter notebooks
- Slides

During this training we will use **Google Colab**

Mauro Pelucchi

I am a senior data scientist and big data engineer responsible for the design of the "**Real-Time Labour Market Information System on Skill Requirements**" for



I currently works for



My main tasks are related to machine learning modelling, labour market analyses, and the design of big data pipelines to process large datasets of online job vacancies.

In collaboration with the **University of Milano-Bicocca**, I took part in many research projects related to the labour market intelligence systems.

I collaborate with the **University of Milano-Bicocca** as a lecturer at the Master Business Intelligence and Big Data Analytics and with the **University of Bergamo** as a lecturer in Computer Engineering.



Agenda

- Data ingestion patterns
- Scraping vs crawling
- How to build a dataset from web data
- Introduction to phantom browsers
- CSS Selector
- Scraping hands-on with Selenium
- Ethics

Data ingestion pattern

Big Data was initially created to serve as storage for all raw data, with access to any kind of historical information always available for corporations.

Over time, **Big Data distributions became more and more effective**, not only at ingesting and scaling the data, but also providing **data processing capabilities for stored and real-time data**.

Scraping vs Crawling

A **web crawler** (spider) is an Internet bot that **systematically** browses the World Wide Web, typically for the **purpose of Web indexing** (web spidering). Web search engines and some other sites use Web crawling or spidering software to update their web content or indices of others sites' web content.

The crawler navigate through a given website and typically **store all data**.

A **web scraping** tool is a technology solution to **extract data** from web sites, in a **quick, efficient** and **automated** manner, offering data in a **more structured** and easier to use format.

Web scrapers simulate human exploration of the World Wide Web by either implementing low-level hypertext transfer protocol or embedding suitable Web browsers.

Scraping vs Crawling

Web Scraping, main tasks

- Internet: most data is unstructured
 - Identifying pages
 - Understanding their structure
 - Navigating (e.g. browsing a list of results)
 - Structuring the data collection
 - Associating a structured field with a portion of the page
 - Building a structured dataset

Why?

- For research (academic, marketing, ...)
 - Data are not available in structured format, no database exists
 - Multiple sources
- For business/eCommerce
 - Market Analysis
 - Product & Price Comparison
 - Online reputation
- For Marketing/CRM
 - What do they say about me (public opinion)?
 - Where is the market moving?
 - What do the media think (e.g. scraping newspaper articles)?
 - Better target ads for customers
 - Detect Fraudulent Reviews
- ...

Tools

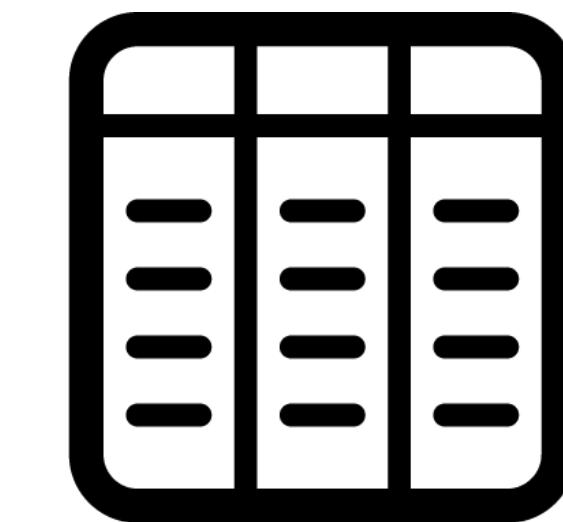
- Visual crawling and scraping tools
 - Import.io www.import.io
 - Mozenda <https://www.mozenda.com>
 - Octoparse www.octoparse.com
- Framework
 - Nutch and Solr <http://nutch.apache.org/>
 - HTTrack www.httrack.com
 - Stormcrawler
- Custom code
 - JSOUP www.jsoup.org (Java), SCRAPY <https://scrapy.org> (Python)
 - **BeautifulSoup <https://pypi.org/project/beautifulsoup4/> (Python)**
 - **Selenium [https://www.selenium.dev/](https://www.selenium.dev)**

<https://github.com/mauropelucchi/pyconsk2022>

INDIEGOGO



+ colab
python

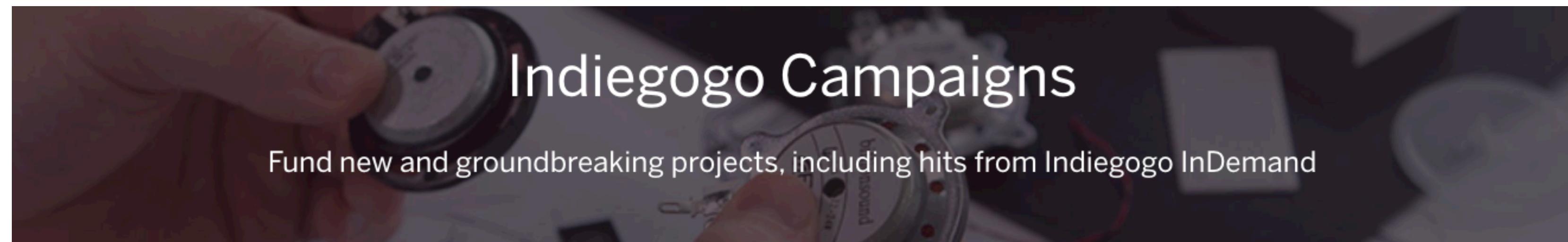
The Python logo, consisting of two interlocking snakes in blue and yellow, followed by the word "colab" in orange and "python" in gray lowercase letters.

The browser

The browser = The best way to **surf** the Internet

Open a web page

→ [https://www.indiegogo.com/explore/all?
project_type=campaign&project_timing=all&sort=trending](https://www.indiegogo.com/explore/all?project_type=campaign&project_timing=all&sort=trending)



Filter results

Search for campaigns

X

CATEGORY

All Categories

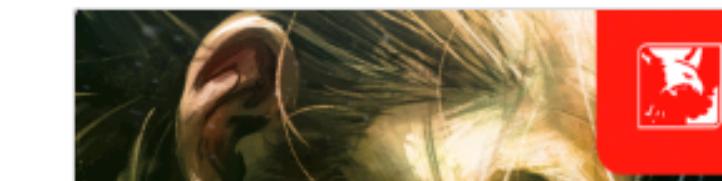
Tech & Innovation ▾

Creative Works ▾

Community Projects ▾

Sort by

Trending ▾



Web browsers use HTML (**HyperText Markup Language**) to display web pages.

HTML

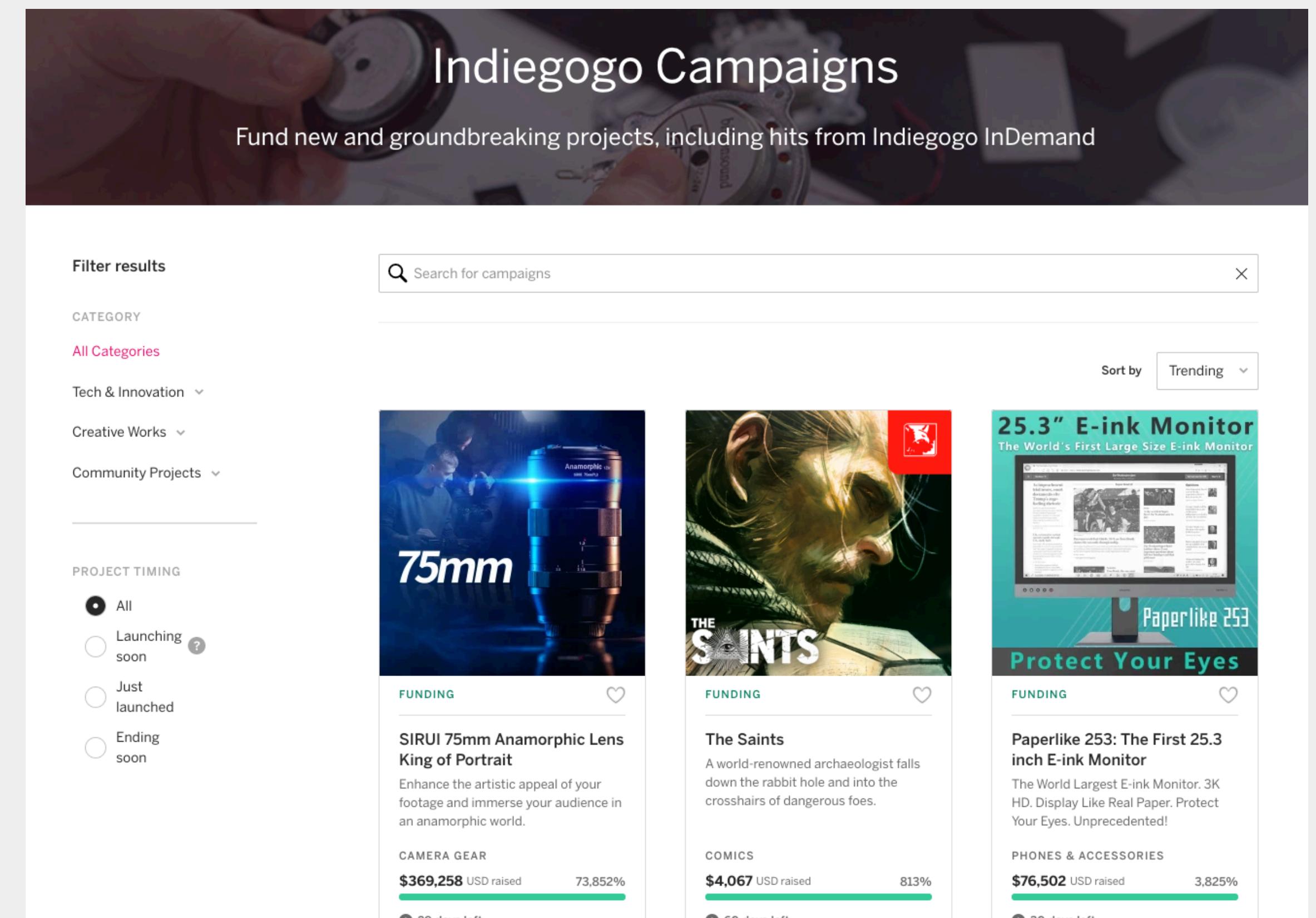
- Is composed of elements: **TAG**
- Each element has a start tag **<element>** and a closing tag **</element>**
- Each element can have an **ID**, which should be unique per page

<element id="ads"> </element>

- **Classes:** used to categorise elements and apply styles. There may be many elements with the class “my_class”.

<element class = “my_class”>

</ element>

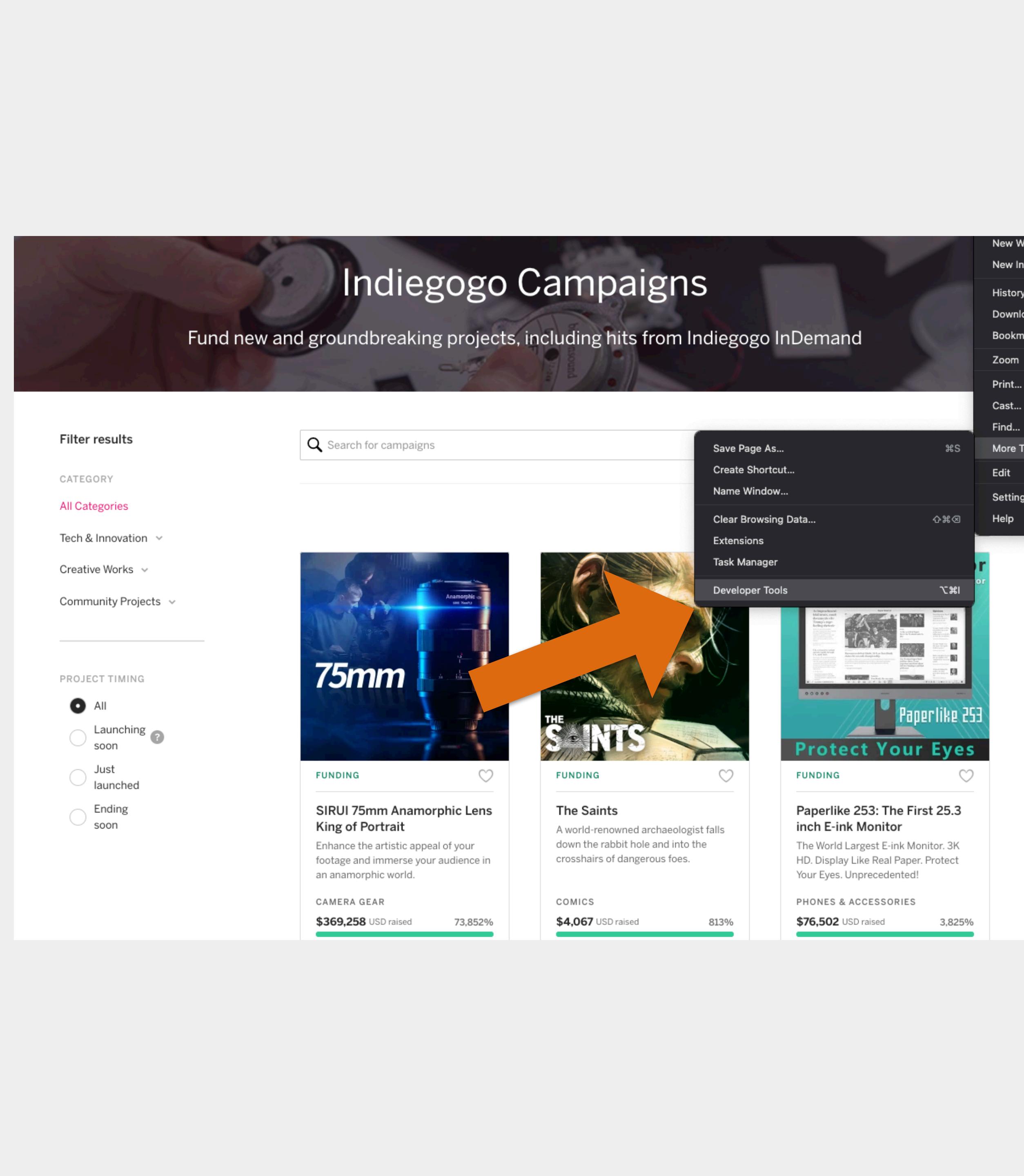


How can I view the structure
of the web page?

Developers Tools

Google Chrome

Menu> More tools> Developers tools



Start

End

The diagram illustrates the structure of an HTML tag. It features a central black text block containing the code: Google. Four orange arrows point from labels to specific parts of the code: an upward arrow on the left points to the opening tag <a>, an upward arrow on the right points to the closing tag , a downward arrow below the href attribute points to the word "Attribute", and a downward arrow below the word "Google" points to the word "Text".

```
<a href="www.google.it">Google</a>
```

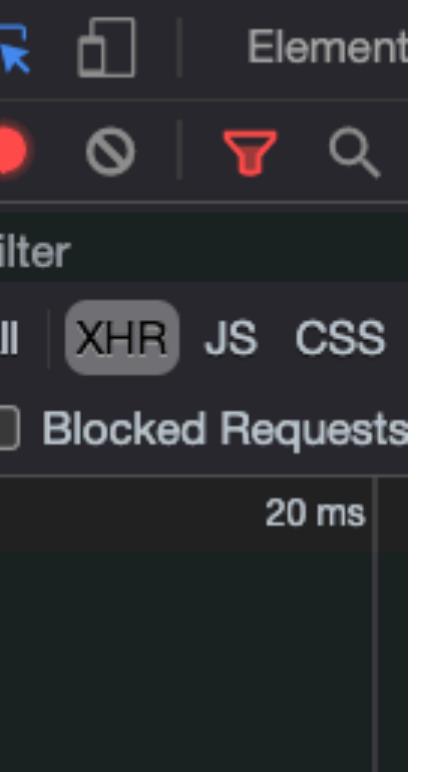
Attribute

Text

The screenshot shows the Indiegogo campaign exploration page. At the top, there's a banner with the text "Indiegogo Campaigns" and "Fund new and groundbreaking projects, including hits from Indiegogo InDemand". Below the banner, there's a search bar labeled "Search for campaigns" and a "Filter results" section. The filter section includes dropdown menus for "CATEGORY" (All Categories, Tech & Innovation, Creative Works, Community Projects) and "PROJECT TIMING" (All, Launching soon, Just launched, Ending soon). On the right side, there's a "Sort by" dropdown set to "Trending". Three campaign cards are displayed:

- SIRUI 75mm Anamorphic Lens King of Portrait**: A camera lens with the text "Anamorphic 75mm" and "75mm" prominently displayed.
- The Saints**: A campaign featuring a man with a mustache and the text "THE SAINTS".
- Paperlike 253: The First 25.3 inch E-ink Monitor**: A monitor displaying a grid of images with the text "Paperlike 253" and "Protect Your Eyes".

Two orange arrows point to specific elements: one arrow points to the "Sort by" dropdown, and another arrow points to the "Funding" section of the first campaign card.



er results
EGORY
Categories

Innovation ▾

Creative Works ▾

Community Projects ▾

JECT TIMING

All

Launching soon

Just launched

Ending soon



FUNDING

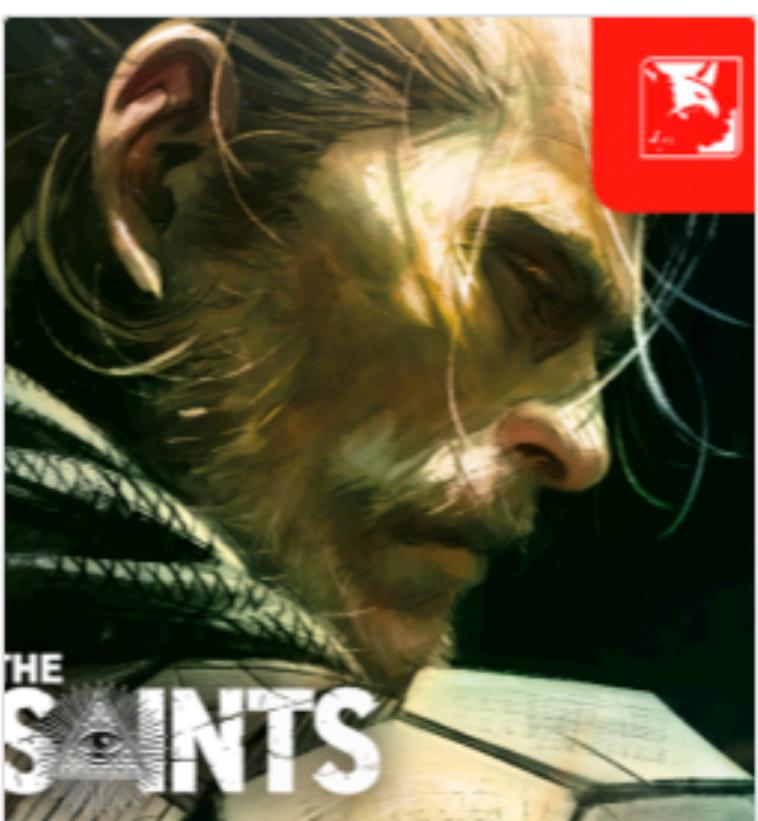
SIRUI 75mm Anamorphic Lens King of Portrait

Enhance the artistic appeal of your footage and immerse your audience in an anamorphic world.

CAMERA GEAR

\$369,258 USD raised 73,852%

29 days left



FUNDING

The Saints

A world-renowned archaeologist falls down the rabbit hole and into the crosshairs of dangerous foes.

COMICS

\$4,067 USD raised 813%

60 days left



FUNDING

Paperlike 253: The First 25.3 inch E-ink Monitor

The World Largest E-ink Monitor. 3K HD. Display Like Real Paper. Protect Your Eyes. Unprecedented!

PHONES & ACCESSORIES

\$76,502 USD raised 3,825%

30 days left

Ctrl+F

```

<div class="discoverableCard-image lazyloaded" ng-azyload="lazyLoadImage" id="discoverableCard-image" bgset="https://c1.indiegogo-media-prod-load/c_fill,f_auto,h_273,w_273/gp5aurjxy2arn8nbflwu" style="background-image: url("https://c1.indiegogo.com/media-prod-cld/image/upload/c_fill,f_auto,h_273,w_273/gp5aurjxy2arn8nbflwu.jpg");">...</div>
<div class="discoverableCard-body">
  <!-- ngIf: viewModel.isCampaign() -->
  <div ng-if="viewModel.isCampaign()" class="discoverableCard-type--crowdfunding ng-sco flex">
    <!-- end ngIf: viewModel.isCampaign() -->
    <div class="discoverableCard-title ng-binding discoverableCard-lineClamp2" ng-class="::viewModel.titleClass" gogo-test="title" ng-bind="::viewModel.discoverable.title">SIRUI 75mm Anamorphic Lens King of Portrait</div>
    <div class="discoverableCard-description ng-binding discoverableCard-lineClamp3" ng-class="::viewModel.descriptionClass" ng-bind="::viewModel.discoverable.tagline"><div class="discoverableCard-category ng-binding" test="category" ng-click="clickCategory($event)" ::viewModel.discoverable.category">Camera Gear</div>
    <!-- ngIf: viewModel.showCrowdFundingProgress() -->
    <div ng-if="viewModel.showCrowdFundingProgress()" class="discoverableCard-crowdfundingProgress ng-scope">...</div>
    <!-- end ngIf: viewModel.showCrowdFundingProgress() -->
    <!-- ngIf: viewModel.isStandardCampaign() -->
    <div ng-if="viewModel.isStandardCampaign()" class="discoverableCard-standardProgress">...</div>
  </div>
</div>
<div class="discoverableCard-body">
  <!-- ngIf: viewModel.isCampaign() -->
  <div ng-if="viewModel.isCampaign()" class="discoverableCard-type--crowdfunding ng-sco flex">
    <!-- end ngIf: viewModel.isCampaign() -->
    <div class="discoverableCard-title ng-binding discoverableCard-lineClamp2" ng-class="::viewModel.titleClass" gogo-test="title" ng-bind="::viewModel.discoverable.title">The Saints</div>
    <div class="discoverableCard-description ng-binding discoverableCard-lineClamp3" ng-class="::viewModel.descriptionClass" ng-bind="::viewModel.discoverable.tagline"><div class="discoverableCard-category ng-binding" test="category" ng-click="clickCategory($event)" ::viewModel.discoverable.category">Comics</div>
    <!-- ngIf: viewModel.showCrowdFundingProgress() -->
    <div ng-if="viewModel.showCrowdFundingProgress()" class="discoverableCard-crowdfundingProgress ng-scope">...</div>
    <!-- end ngIf: viewModel.showCrowdFundingProgress() -->
    <!-- ngIf: viewModel.isStandardCampaign() -->
    <div ng-if="viewModel.isStandardCampaign()" class="discoverableCard-standardProgress">...</div>
  </div>
</div>
<div class="discoverableCard-body">
  <!-- ngIf: viewModel.isCampaign() -->
  <div ng-if="viewModel.isCampaign()" class="discoverableCard-type--crowdfunding ng-sco flex">
    <!-- end ngIf: viewModel.isCampaign() -->
    <div class="discoverableCard-title ng-binding discoverableCard-lineClamp2" ng-class="::viewModel.titleClass" gogo-test="title" ng-bind="::viewModel.discoverable.title">Paperlike 253: The First 25.3 inch E-ink Monitor</div>
    <div class="discoverableCard-description ng-binding discoverableCard-lineClamp3" ng-class="::viewModel.descriptionClass" ng-bind="::viewModel.discoverable.tagline"><div class="discoverableCard-category ng-binding" test="category" ng-click="clickCategory($event)" ::viewModel.discoverable.category">Phones & Accessories</div>
    <!-- ngIf: viewModel.showCrowdFundingProgress() -->
    <div ng-if="viewModel.showCrowdFundingProgress()" class="discoverableCard-crowdfundingProgress ng-scope">...</div>
    <!-- end ngIf: viewModel.showCrowdFundingProgress() -->
    <!-- ngIf: viewModel.isStandardCampaign() -->
    <div ng-if="viewModel.isStandardCampaign()" class="discoverableCard-standardProgress">...</div>
  </div>
</div>

```

Styles Computed Layout Event Listeners DOM Breakpoints Properties Accessibility

Filter

```

element.style {
}
@media (min-width: 480px) {
  .discoverableCard-title {
    font-family: "benton-sans", "Helvetica", "sans-serif";
    font-size: 16px;
  }
}

```



div.discoverableCard-title.ng-binding.discoverableCard-lineClamp2
221.98x57



ING



LIVALL EVO21 Smart Helmet

• 270° rear light • 350g ultra lightweight • SOS system

FUNDING

LIVALL EVO21 Smart Helmet: 360 Active...

The world's smartest lighting helmet with 270° rear light, patented fall detection and SOS alert

TRAVEL & OUTDOORS

\$213,411 USD raised 2,134%

30 days left

AirOgo Pilloon Ultralight - Most Versatile Jacket

Ready for any situation anywhere: daily commute, business travels, and your everyday activities

TRAVEL & OUTDOORS

\$6,529 USD raised 131%

30 days left

Tangible Hope Project

Real People. Real Communities. Real Impact. A Documentary Series About Real Hope in America.

WEB SERIES & TV SHOWS

\$2,450 USD raised 25%

30 days left

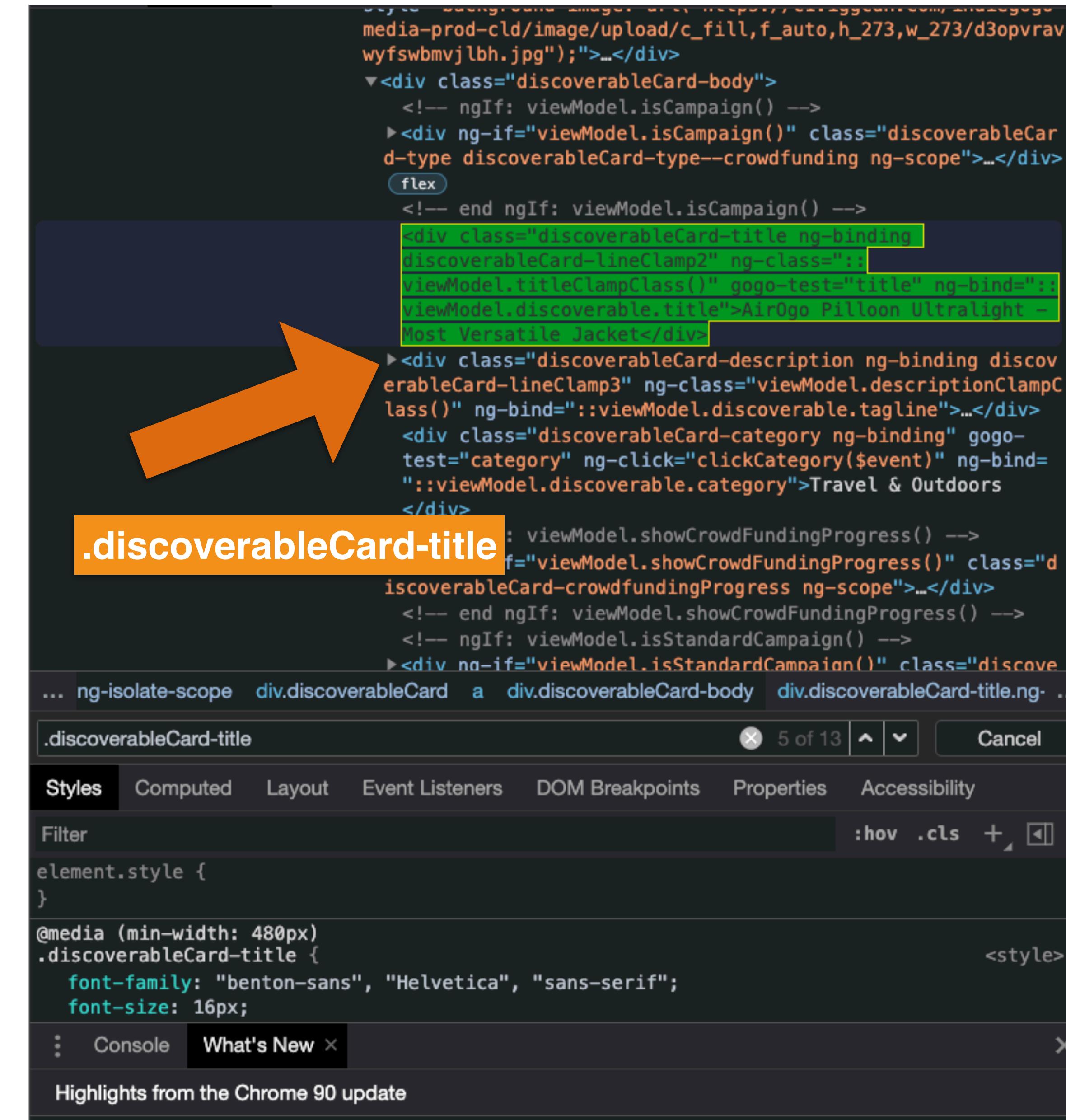
FUNDED IN 15 MINUTES



Funded in 15 Minutes

ONE XPLAYER

FIFD OF



```
media-prod-cld/image/upload/c_fill,f_auto,h_273,w_273/d3opvravwyfswbmvjlbh.jpg");">...</div>
<div class="discoverableCard-body">
  <!-- ngIf: viewModel.isCampaign() -->
  <div ng-if="viewModel.isCampaign()" class="discoverableCard-type discoverableCard-type--crowdfunding ng-scope">...</div>
  flex
  <!-- end ngIf: viewModel.isCampaign() -->
<div class="discoverableCard-title ng-binding" discoverableCard-lineClamp2" ng-class="::viewModel.titleClampClass()" gogo-test="title" ng-bind="::viewModel.discoverable.title">AirOgo Pilloon Ultralight - Most Versatile Jacket</div>
<div class="discoverableCard-description ng-binding discoverableCard-lineClamp3" ng-class="viewModel.descriptionClampClass()" ng-bind="::viewModel.discoverable.tagline">...</div>
<div class="discoverableCard-category ng-binding" gogo-test="category" ng-click="clickCategory($event)" ng-bind="::viewModel.discoverable.category">Travel & Outdoors</div>
<div>: viewModel.showCrowdFundingProgress() -->
  <div ng-if="viewModel.showCrowdFundingProgress()" class="discoverableCard-crowdfundingProgress ng-scope">...</div>
  <!-- end ngIf: viewModel.showCrowdFundingProgress() -->
  <!-- ngIf: viewModel.isStandardCampaign() -->
  <div ng-if="viewModel.isStandardCampaign()" class="discoverableCard-body">
    ... ng-isolate-scope div.discoverableCard a div.discoverableCard-body div.discoverableCard-title.ng-binding
    .discoverableCard-title
    Styles Computed Layout Event Listeners DOM Breakpoints Properties Accessibility
    Filter
    element.style {
    }
    @media (min-width: 480px)
    .discoverableCard-title {
      font-family: "benton-sans", "Helvetica", "sans-serif";
      font-size: 16px;
    }
    :: Console What's New ×
    Highlights from the Chrome 90 update
  
```

Techniques - XPATH

XPath is a syntax for defining parts of an XML document.
XPath uses path expressions to navigate in XML documents.



https://www.w3schools.com/xml/xpath_examples.asp

Expression	Description
<i>nodename</i>	Selects all nodes with the name " <i>nodename</i> "
/	Selects from the root node
//	Selects nodes in the document from the current node that match the selection no matter where they are
.	Selects the current node
..	Selects the parent of the current node
@	Selects attributes

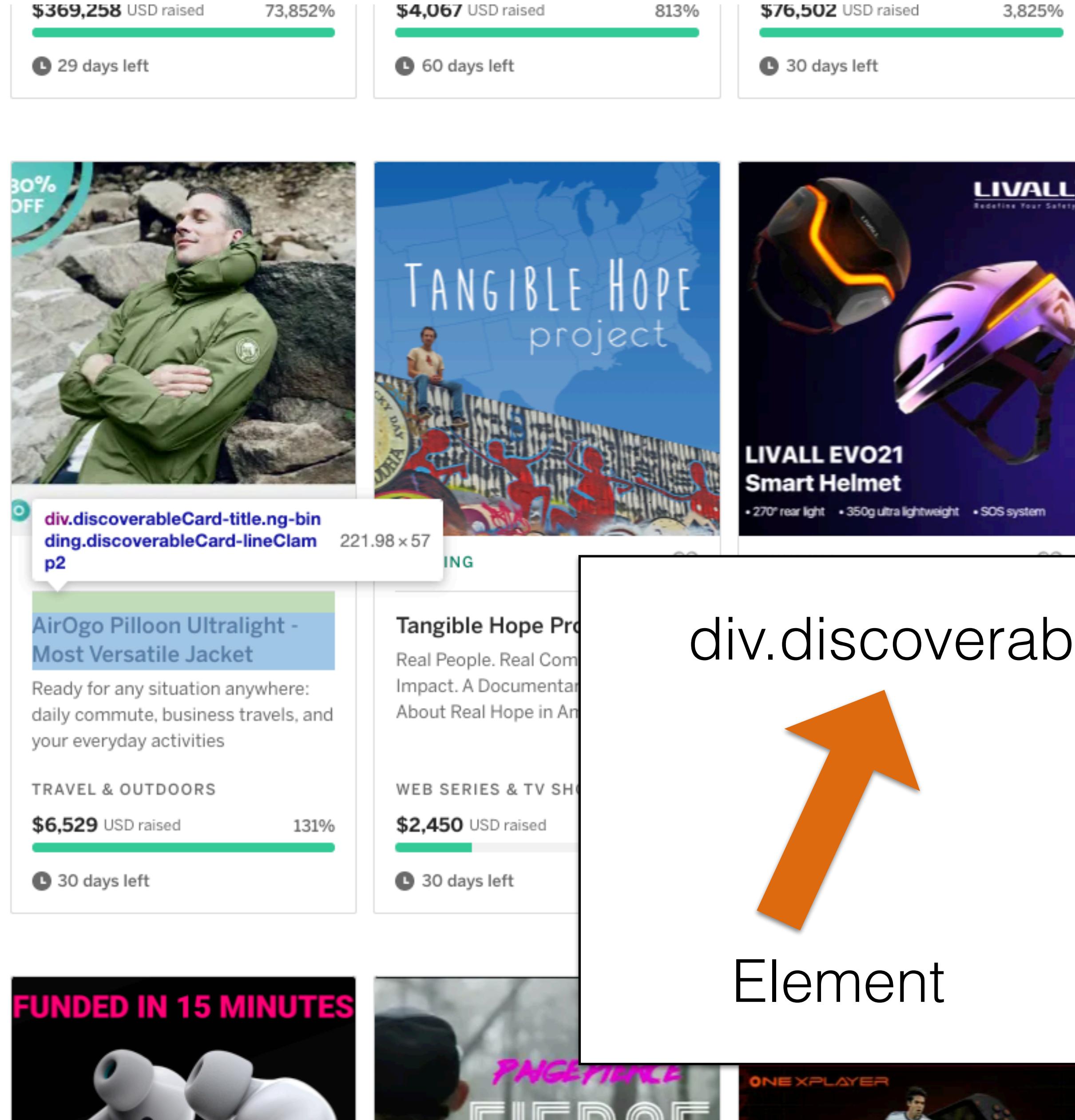
Wildcard	Description
*	Matches any element node
@*	Matches any attribute node
node()	Matches any node of any kind

Techniques - CSS Selector

In CSS, selectors declare which part of the markup a style applies to by **matching tags and attributes** in the markup itself.

Selector overview

- `tagname`: find elements by tag, e.g. `a`
- `ns | tag`: find elements by tag in a namespace, e.g. `fb | name` finds `<fb:name>` elements
- `#id`: find elements by ID, e.g. `#logo`
- `.class`: find elements by class name, e.g. `.masthead`
- `[attribute]`: elements with attribute, e.g. `[href]`
- `[^attr]`: elements with an attribute name prefix, e.g. `[^data-]` finds elements with HTML5 dataset attributes
- `[attr=value]`: elements with attribute value, e.g. `[width=500]` (also quotable, like `[data-name='launch sequence']`)
- `[attr^=value]`, `[attr$=value]`, `[attr*=value]`: elements with attributes that start with, end with, or contain the value, e.g. `[href*=/path/]`
- `[attr~=regex]`: elements with attribute values that match the regular expression; e.g. `img[src~=(?i)\.(png|jpe?g)]`
- `*`: all elements, e.g. `*`



div.discoverableCard-title

Element

Class



Highlights from the Chrome 90 update

```

city, hidden !important, position: fixed !important, height: 1px !important, pointer-events: none !important; user-select: none !important;">
!important; user-select: none !important;">

#document
  <!DOCTYPE html>
  <html>
    <head>...</head>
    <body>
      <script src="fingerprinted/js/m-outer-b07c750...js"></script>
      <iframe src="https://m.stripe.network/inner.html?url=https%3A%2F%2Fwww.indiegogo.com%2F%2F%2Fwww.indiegogo.com%2Fmuid=NA&sid=NA&version=6&preview=false">...</iframe>
    </body>
  </html>
</iframe>
<iframe scrolling="no" frameborder="0" allowtransparency="true" src="https://platform.twitter.com/widgets/widget_iframe.06c6ee5...html?origin=https%3A%2F%2Fwww.indiegogo.com" title="Twitterings iframe" style="display: none; ">...</iframe>
<script type="text/javascript" id="fbq("trackCustom","BrowserLanguage", {language_id:"it-IT")>
</script>
<noscript>...</noscript>
<script type="text/javascript" id="Cookiebot" src="https://consent.cookiebot.com/uc.js?chid=152ad262-955d-4688-b0b3-ea8bee2ea245"></script> == $0
<iframe name="__uspapiLocator" tabindex="-1" role="presentation" aria-hidden="true" title="Blank" style="display: none; position: absolute; width: 1px; height: 1px; top: -9999px; "></i
<iframe tabindex="-1" role="presentation" aria-hidden="true" title="Blank" src="https://cdn.cookiebot.com/sdk/bc-v3.min.html" style="position: absolute; width: 1px; height: 1px; top: -9999px; "></i
... html body.locale-en script#Cookiebot
script#Cookiebot| 1 of 1 Cancel
Styles Computed Layout Event Listeners DOM Breakpoints Properties Accessibility
Filter
element.style {
}
* {
  box-sizing: border-box;
  -webkit-font-smoothing: antialiased;
  -moz-osx-font-smoothing: antialiased;
}
: Console What's New X
<style>

```

Sort by Trending

div.discoverableCard-description.ng-binding.discoverableCard-li 221.98 x 67 neClamp3

Color #6A6A6A
Font 13px benton-sans, Helvetica, sans-serif
Padding 5px 0px

ACCESSIBILITY

Contrast Aa 5.4 ⓘ
Name
Role generic ⓘ
Keyboard-focusable

Enhance the artistic appeal of your footage and immerse your audience in an anamorphic world.

CAMERA GEAR
\$369,258 USD raised 73,852%
29 days left

COMICS
\$4,067 USD raised 813%
60 days left

25.3" E-ink Monitor
The World's First Large Size E-ink Monitor

Paperlike 253 Protect Your Eyes

FUNDING

Paperlike 253: The First 25.3 inch E-ink Monitor

The World Largest E-ink Monitor. 3K HD. Display Like Real Paper. Protect Your Eyes. Unprecedented!

PHONES & ACCESSORIES
\$76,502 USD raised 3,825%
30 days left

LIVALL Redefine Your Safety

30% OFF TANSONE LANE

script#Cookiebot

Element ID

Scopri Dai vita a un progetto

KICKSTARTER

Cerca Q Accedi

Mostra Tutte le categorie luogo Terra ordinati per Magia

Più filtri

Esplora **516.411** progetti

PHANTASMIA'S
Progetti che amiamo

Phantasmias: Fantasy-Horror Anthology
A collection of 27 fantasy and horror stories blurring the line between genres...
di Scribe Cat

Jadelabo J1 - Beautiful, Powerful, and Smart INDEX 3D Printer
Double the performance and productivity with Independent Dual...
di Jadelabo

The Lark's Call: Irish Traditional music on uilleann pipes.
Tom Delany's debut solo album
di Tom Delany

2.794 CA\$ raccolti
111% finanziato
26 giorni rimasti
Antologie ➔ Toronto, Canada

586.499 \$ raccolti
586% finanziato
8 giorni rimasti
Stampa 3D ➔ New Castle, DE

5.385 € raccolti
76% finanziato
18 giorni rimasti
Country e folk ➔ Dingle, Ireland

Elements Console Sources Network Performance Memory >

```
xsf-ugc.imgur.net/assets/033/414/229/0f9959414fe0d6b0c03c3f0ae30cccd65_original.png?ixlib=rb-4.0.2&crop=0,0,256,256&fit=crop&v=1500481286&auto=format&frame=1&q=92&s=d381bf808eb7c79d261792cd9b9aa3" 4.0.25w=80&h=80&fit=crop&v=1500481286&auto=format&frame=1&q=92&s=9f1623672736a77a1c271fe0c633273" 4.0.25w=160&h=160&fit=crop&v=1500481286&auto=format&frame=1&q=92&s=159df837def8e1546964clae8669065 581160863/signature=1621952205_7e0fdcf3c84425071b980946bde520591db7b3b"}}, "location": {"id": 1014118, "name": "Toronto", "country": "CA", "state": "ON", "type": "Town", "is_root": false, "expanded_country": "Canada", "url": "https://api.kickstarter.com/v1/discover?signature=1621952205_7e0fdcf3c84425071b980946bde520591db7b3b"}, "parent_id": null, "link_text": "nearby_projects": "https://api.kickstarter.com/v1/discover?signature=1621952205_7e0fdcf3c84425071b980946bde520591db7b3b", "api": "nearby_projects", "analytics_name": "Anthologies", "slug": "publishing/anthologies", "position": 12, "parent_id": null, "link_text": "link_text": null, "link_url": null, "show_feature_image": false, "background_color": "#f0f0f0", "font_color": "#333333", "text_color": "#333333", "link_text": "link_text": null, "link_url": null, "show_feature_image": false, "background_color": "#f0f0f0", "font_color": "#333333", "text_color": "#333333", "url": "https://www.kickstarter.com/projects/scribecat/phantasmias-fantasy-horror-anthology", "disable_communication": false, "country": "CA", "country_display": "Canada", "percent_funded": 111.7715999999999, "is_liked": false, "is_disliked": false}, "data_ref": "discovery"}
```

1 of 12 ▲ ▼ Cancel

Styles Computed Layout Event Listeners DOM Breakpoints Properties Accessibility

Filter

element.style { }

.grid-container * { box-sizing: inherit; }

base-f09cf3_ac620.css:1

html, body, div, span, anker, object, iframe, p, blockquote, pre, a { base-f09cf3_ac620.css:1 }

: Console What's New

Highlights from the Chrome 90 update

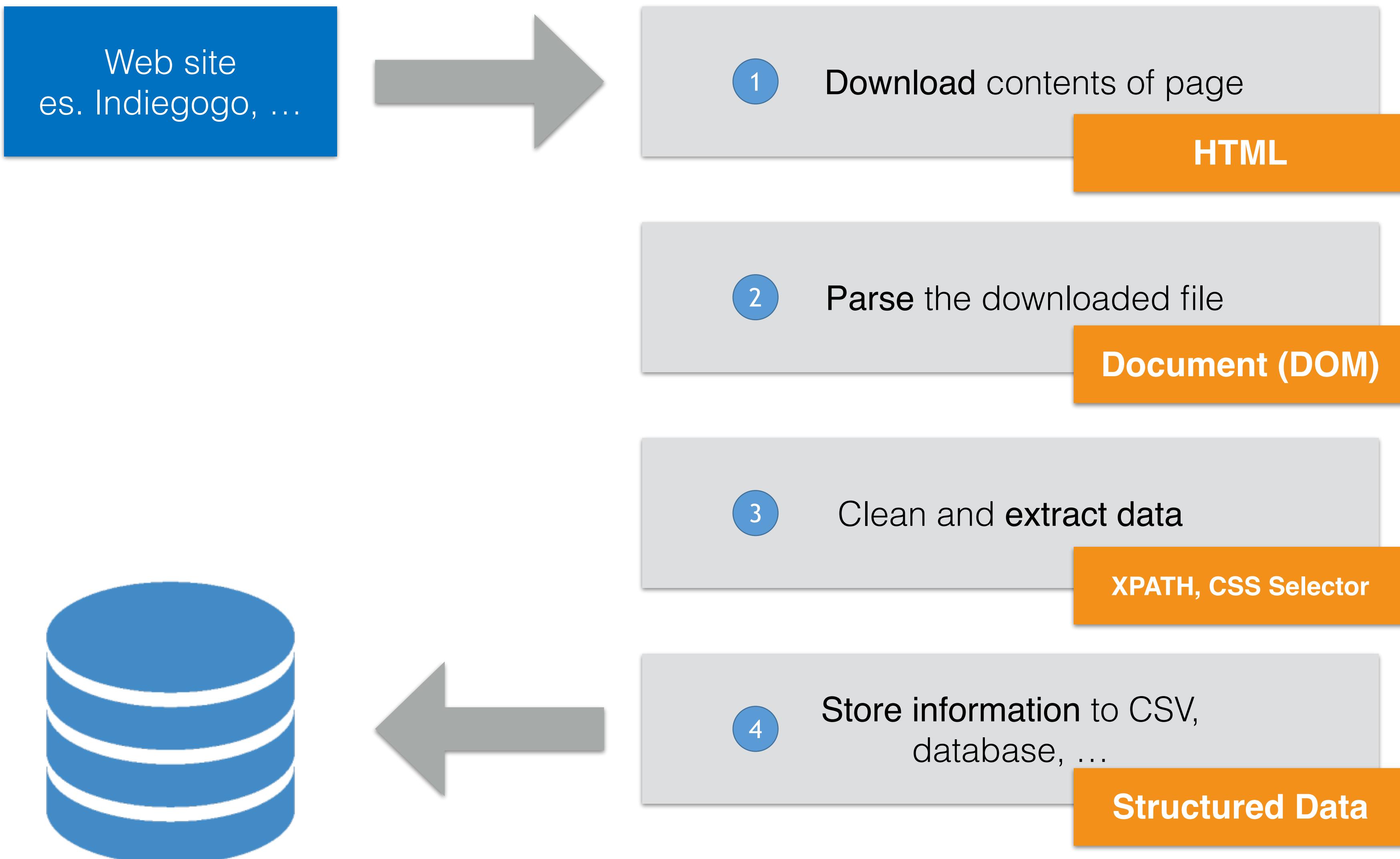
New CSS Flexbox debugging tools

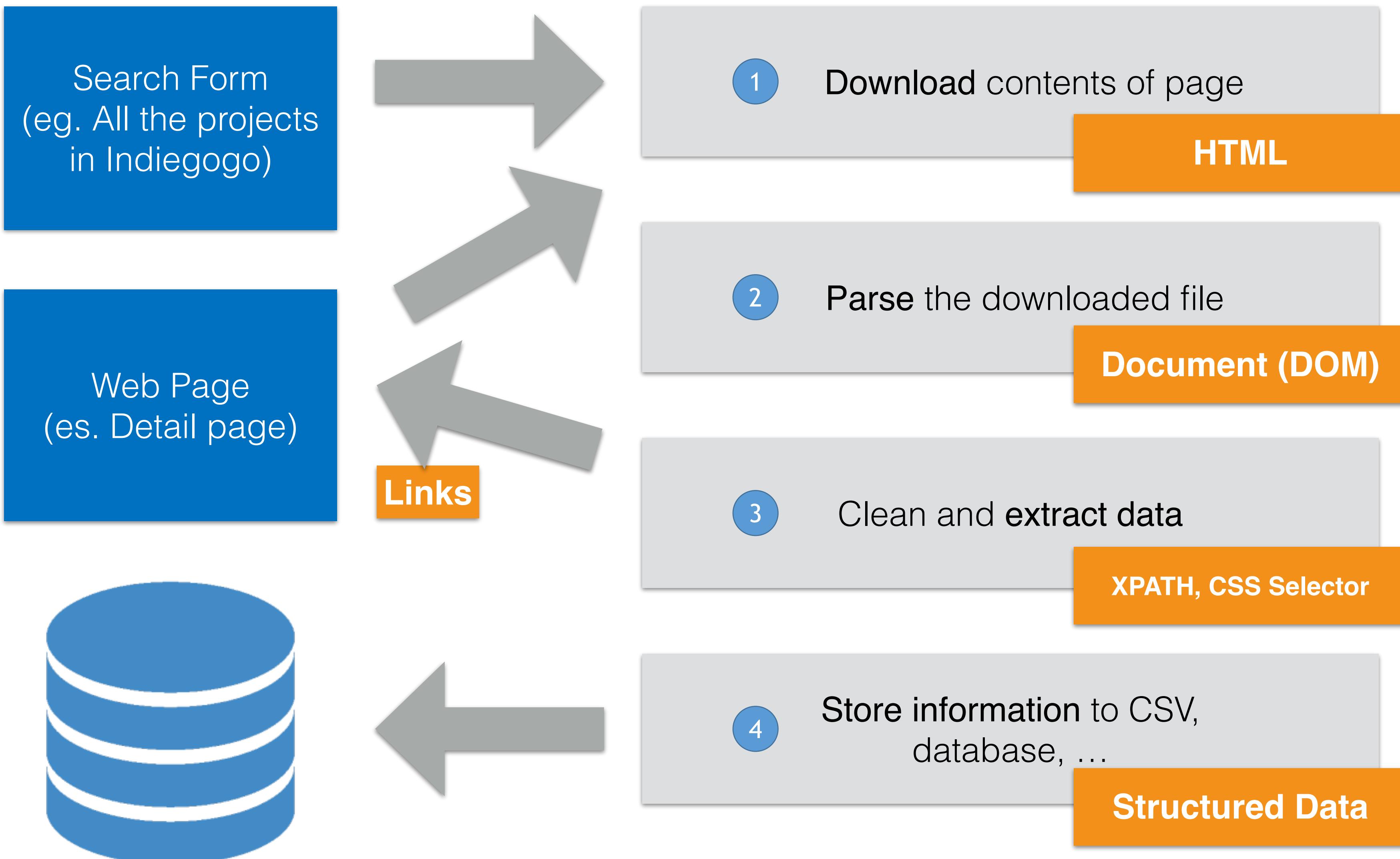
Debug and inspect CSS Flexbox with the new CSS Flexbox debugging tools.

discoverable-card > div > a



By parent relation







Inspecting a website

https://www.indiegogo.com/explore/all?project_type=campaign&project_timing=all&sort=trending



Filter results

CATEGORY

All Categories

Tech & Innovation ▾

Creative Works ▾

Community Projects ▾

PROJECT TIMING

All

Launching soon ?

Just launched

Ending soon

Search for campaigns X

Sort by Trending ▾

75mm

FUNDING

SIRUI 75mm Anamorphic Lens King of Portrait

Enhance the artistic appeal of your footage and immerse your audience in

THE SAINTS

FUNDING

The Saints

A world-renowned archaeologist falls down the rabbit hole and into the crosshairs of dangerous foes.

25.3" E-ink Monitor

The World's First Large Size E-ink Monitor

Paperlike 253

Protect Your Eyes

FUNDING

Paperlike 253: The First 25.3 inch E-ink Monitor

The World Largest E-ink Monitor. 3K HD. Display Like Real Paper. Protect



Inspecting a website



AirOgo® Pilloon Jacket Ultralight

FUNDING



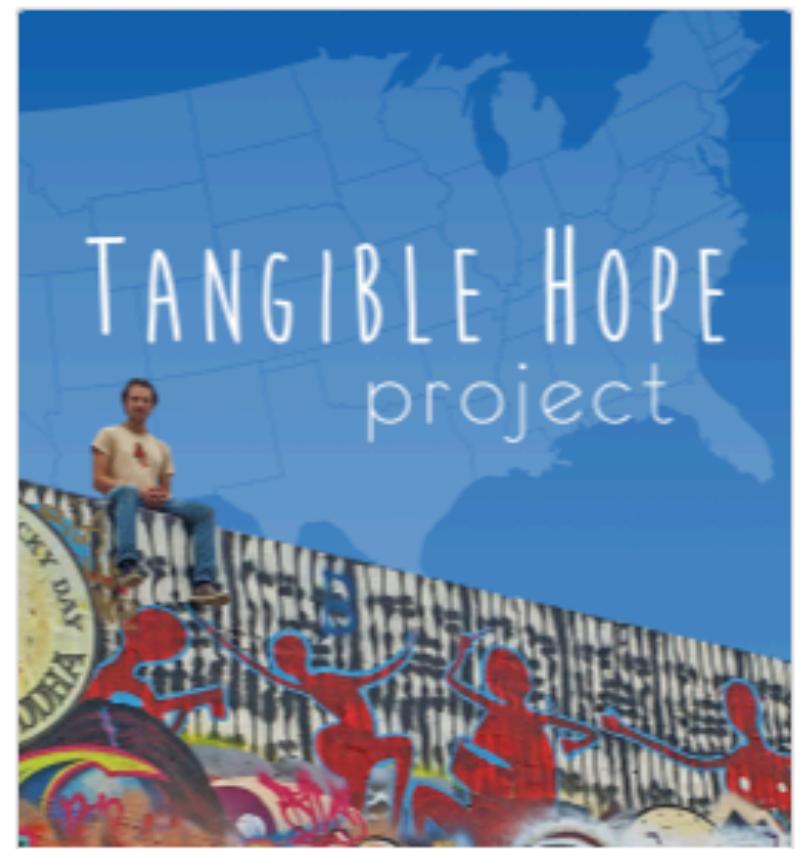
AirOgo Pilloon Ultralight - Most Versatile Jacket

Ready for any situation anywhere: daily commute, business travels, and your everyday activities

TRAVEL & OUTDOORS

\$6,529 USD raised 131%

30 days left



FUNDING



Tangible Hope Project

Real People. Real Communities. Real Impact. A Documentary Series About Real Hope in America.

WEB SERIES & TV SHOWS

\$2,450 USD raised 25%

30 days left



FUNDING



LIVALL EVO21 Smart Helmet: 360 Active...

The world's smartest lighting helmet with 270° rear light, patented fall detection and SOS alert

TRAVEL & OUTDOORS

\$213,411 USD raised 2,134%

18 days left

Most Versatile Jacket</div>

```
><div class="discoverableCard-description ng-binding discoverableCard-lineClamp3" ng-class="viewModel.descriptionClampClass()" ng-bind="viewModel.discoverable.tagline">...</div>
<div class="discoverableCard-category ng-binding" gogo-test="category" ng-click="clickCategory($event)" ng-bind="::viewModel.discoverable.category">Travel & Outdoors</div>
<!-- ngIf: viewModel.showCrowdFundingProgress() -->
<div ng-if="viewModel.showCrowdFundingProgress()" class="discoverableCard-crowdfundingProgress ng-scope">...</div>
<!-- end ngIf: viewModel.showCrowdFundingProgress() -->
<!-- ngIf: viewModel.isStandardCampaign() -->
<div ng-if="viewModel.isStandardCampaign()" class="discoverableCard-crowdfundingTimeLeft ng-scope">...</div> flex
<!-- end ngIf: viewModel.isStandardCampaign() -->
<!-- ngIf: viewModel isInDemandCampaign() -->
<!-- ngIf: viewModel.isPreLaunchCampaign() -->
<!-- ngIf: viewModel.isPreLaunchCampaign() -->
</div>
</a>
```

discoverable-card

```
<!-- end ngRepeat: campaign in campaigns track by campaign.clickthroughUrl -->
<div ng-repeat="campaign in campaigns track by campaign.clickthroughUrl" rd.ng-scope.ng-isolate-scope>
  <div class="discoverableCard">
    <a href="#" class="discoverableCard-link" discoverable-card="campaign">
      <div class="discoverableCard-body">
        ...
```

discoverable-card

Styles Computed Layout Event Listeners DOM Breakpoints Properties Accessibility

Filter :hover .cls + □

```
element.style { }
@media (min-width: 480px) {
  .discoverableCard-title {
    font-size: 1.2em;
    margin-bottom: 0.5em;
  }
}
```

<style>



Inspecting a website



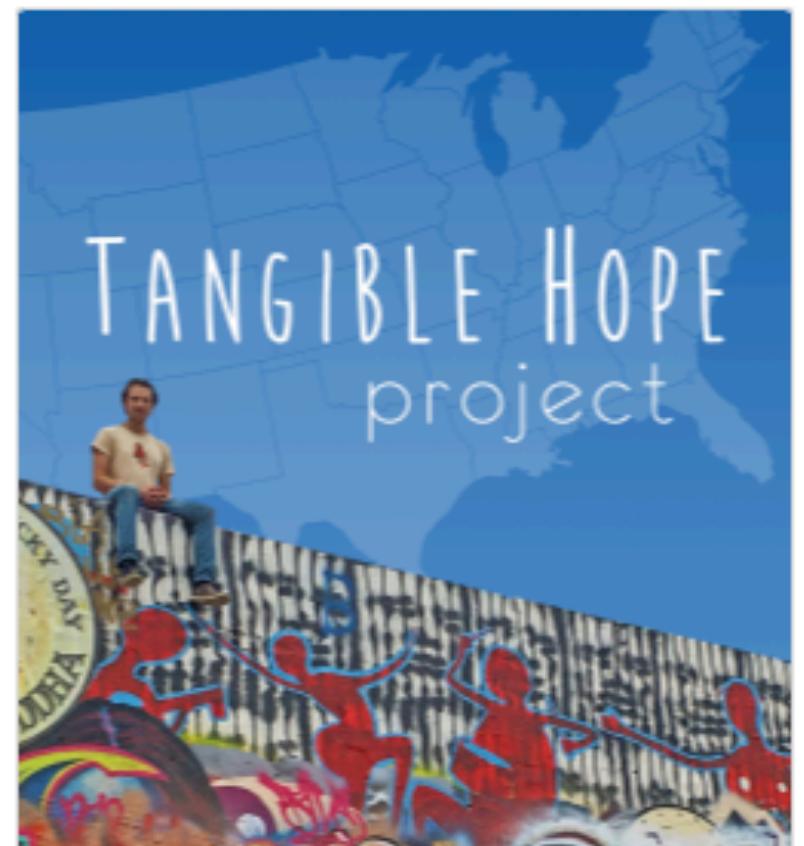
AirOgo Pilloon Ultralight - Most Versatile Jacket

Ready for any situation anywhere: daily commute, business travels, and your everyday activities

TRAVEL & OUTDOORS

\$6,529 USD raised 131%

30 days left



Tangible Hope Project

Real People. Real Communities. Real Impact. A Documentary Series

About Real Hope in America.

WEB SERIES & TV SHOWS

\$2,450 USD raised 25%

30 days left



LIVALL EVO21 Smart Helmet: 360 Active...

The world's smartest lighting helmet with 270° rear light, patented fall detection and SOS alert

TRAVEL & OUTDOORS

\$213,411 USD raised 2,134%

18 days left

Most Versatile Jacket</div>

><div class="discoverableCard-description ng-binding discoverableCard-lineClamp3" ng-class="viewModel.descriptionClampClass()" ng-bind="viewModel.discoverable.tagline">...</div>

<div class="discoverableCard-category ng-binding" gogo-test="category" ng-click="clickCategory(\$event)" ng-bind="::viewModel.discoverable.category">Travel & Outdoors</div>

<!-- ngIf: viewModel.showCrowdFundingProgress() -->

><div ng-if="viewModel.showCrowdFundingProgress()" class="discoverableCard-crowdfundingProgress ng-scope">...</div>

<!-- end ngIf: viewModel.showCrowdFundingProgress() -->

<!-- ngIf: viewModel.isStandardCampaign() -->

><div ng-if="viewModel.isStandardCampaign()" class="discoverableCard-crowdfundingTimeLeft ng-scope">...</div> flex

<!-- end ngIf: viewModel.isStandardCampaign() -->

<!-- ngIf: viewModel isInDemandCampaign() -->

<!-- ngIf: viewModel.isPreLaunchCampaign() -->

<!-- ngIf: viewModel.isPreLaunchCampaign() -->

</div>

</div>

</discoverable-card>

<!-- end ngRepeat: campaign in campaigns track by campaign.clickthroughUrl -->

.discoverableCard-title

... rd.ng-scope.ng-isolate-scope div.discoverableCard a div.discoverableCard-body div.discoverableCard-lineClamp3

discoverable-card

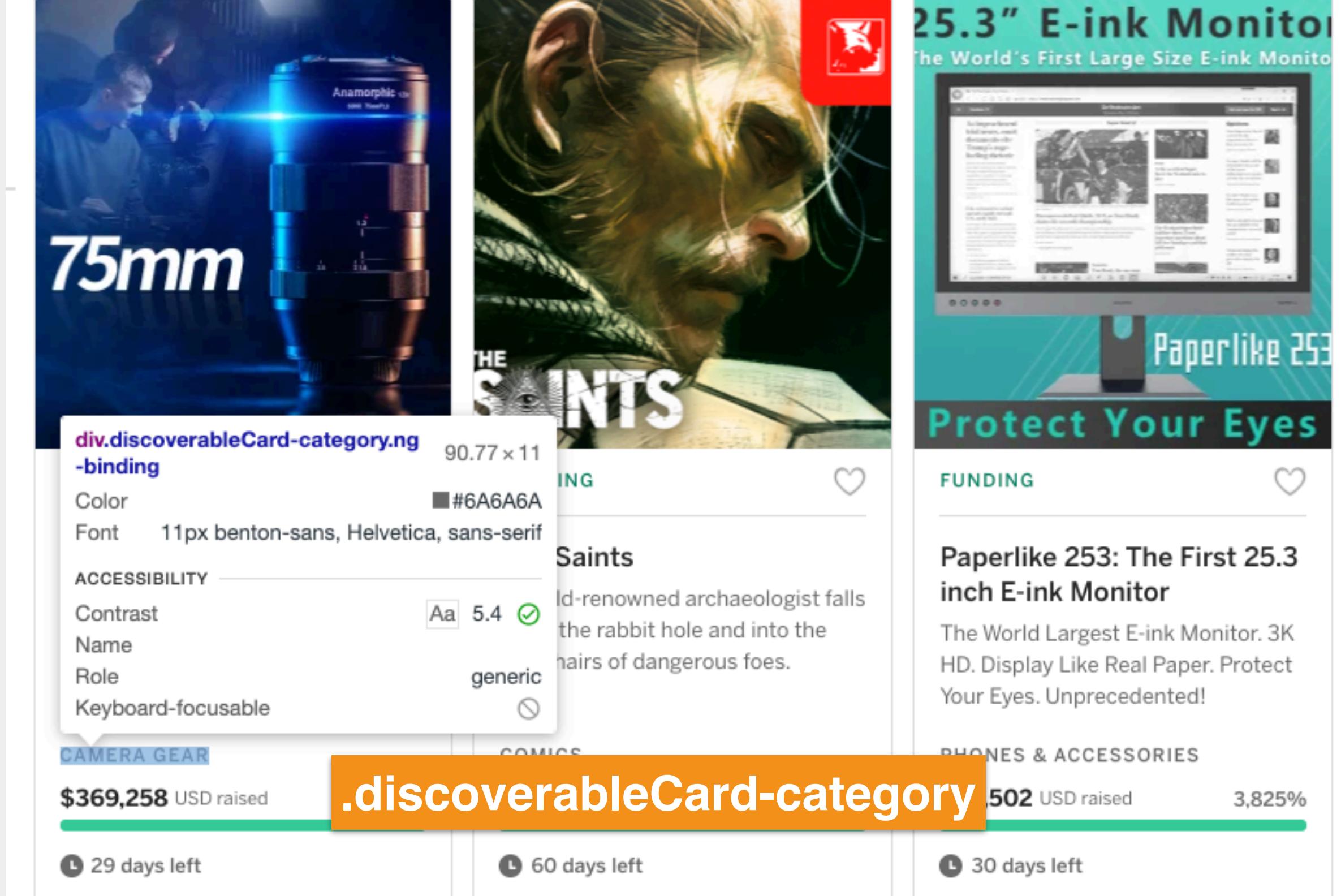
Styles Computed Layout Event Listeners DOM Breakpoints Properties Accessibility

Filter :hover .cls + □

```
element.style { }
}
@media (min-width: 480px)
.discoverableCard-title {
  font-size: 1.2em;
  margin-bottom: 0.5em;
}
```

<style>

Inspecting a website



The Saints

Anamorphic 75mm

THE SAINTS

90.77 x 11

Color #6A6A6A

Font 11px benton-sans, Helvetica, sans-serif

ACCESSIBILITY

Contrast Aa 5.4 ✓

Name generic

Role Keyboard-focusable

CAMERA GEAR

\$369,258 USD raised

29 days left

COMICS

60 days left

PHONES & ACCESSORIES

502 USD raised 3,825%

30 days left

discoverableCard-category

30% OFF

TANGIBLE HOPE

LIVALL Redefine Your Safety

25.3" E-ink Monitor

The World's First Large Size E-ink Monitor

Paperlike 253 Protect Your Eyes

FUNDING

Paperlike 253: The First 25.3 inch E-ink Monitor

The World Largest E-ink Monitor. 3K HD. Display Like Real Paper. Protect Your Eyes. Unprecedented!

Inspecting a website

The screenshot shows a crowdfunding platform with three project cards:

- Cleep Pro - The Smallest 4K Wearable Camera**: 4K Mini Vlogging Camera| EIS Stabilization| Easy Wearable| Ultra-Lightweight Body| 60 Min Recording. CAMERAGEAR. €1,797 EUR raised (36%) over 40 days left.
- EDGE - The First Modular Work From Home Kit**: Your workstation, anywhere. Guaranteed to boost your productivity. PRODUCE Color: #FFFFFF; Font: 14px benton-sans, Helvetica, sans-serif; Background: #E51075; Padding: 11px 18px.
- Koinu : Professional Flat-edged Bowls**: Take advantage of a bowl that has been made for Professionals to be functional! ERAGES. PY raised (82%).

A yellow callout box highlights the "Keyboard-focusable" accessibility item, which is checked. Below the callout is a "SHOW MORE" button. To the right of the callout is an "exploreMore" button.

ACCESSIBILITY

- Contrast: Aa 4.5 (checked)
- Name
- Role
- Keyboard-focusable (checked)

SHOW MORE

exploreMore

ABOUT

About Us

Blog

Trust & Safety

ENTREPRENEURS

How It Works

Indiegogo vs. Kickstarter

Education Center



Find it first on Indiegogo

Discover new and clever products in

Recap

Link

discoverable-card a [href]

Category

.discoverableCard-category

Title

.discoverableCard-title

Description

.discoverableCard-description

BeautifulSoup

A toolkit for dissecting a document and extracting what you need.

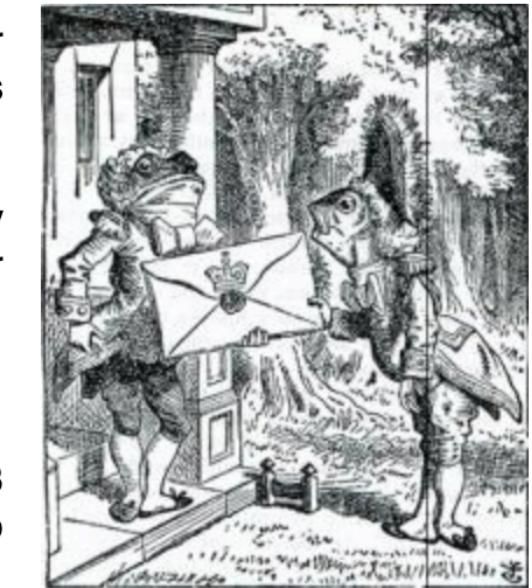
Beautiful Soup Documentation

Beautiful Soup is a Python library for pulling data out of HTML and XML files. It works with your favorite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work.

These instructions illustrate all major features of Beautiful Soup 4, with examples. I show you what the library is good for, how it works, how to use it, how to make it do what you want, and what to do when it violates your expectations.

The examples in this documentation should work the same way in Python 2.7 and Python 3.2.

You might be looking for the documentation for [Beautiful Soup 3](#). If so, you should know that Beautiful Soup 3 is no longer being developed, and that Beautiful Soup 4 is recommended for all new projects. If you want to learn about the differences between Beautiful Soup 3 and Beautiful Soup 4, see [Porting code to BS4](#).



```
pip install --user beautifulsoup4  
pip install --user requests  
pip install --user tqdm
```

Selenium

Selenium is a portable framework for testing web applications.

Scraping with Selenium

<https://www.selenium.dev/>

The screenshot shows the official Selenium website at https://www.selenium.dev/. The page has a green header bar with the text "Selenium automates browsers. That's it!" and "What you do with that power is entirely up to you.". Below the header, there are three main sections: "Getting Started" (with Selenium WebDriver, Selenium IDE, and Selenium Grid), "Downloads" (with Selenium Java, Python, C#, .NET, Java, and Java), and "Projects" (with Java, Python, C#, .NET, and Java). The "Downloads" section includes links to "Setup" and "Source Code". The "Projects" section includes links to "Java", "Python", "C#", ".NET", and "Java". A sidebar on the left lists "Documentation", "Support", "Blog", and "About". A search bar is located at the top right.

Are you ready?



<https://bit.ly/3Bvuq43>

JSOUP

jsoup: Java HTML Parser

`jsoup` is a Java library for working with real-world HTML. It provides a very convenient API for extracting and manipulating data, using the best of DOM, CSS, and jquery-like methods.

`jsoup` implements the **WHATWG HTML5** specification, and parses HTML to the same DOM as modern browsers do.

- scrape and `parse` HTML from a URL, file, or string
- `find` and extract data, using DOM traversal or CSS selectors
- `manipulate` the HTML elements, attributes, and text
- `clean` user-submitted content against a safe white-list, to prevent XSS attacks
- `output` tidy HTML



jsoup is designed to deal with all varieties of HTML found in the wild; from pristine and validating, to invalid tag-soup; jsoup will create a sensible parse tree.



Scrapy

An open source and collaborative framework
for extracting the data you need from websites.
In a fast, simple, yet extensible way.

pip install scrapy

scrapy startproject zalando

scrapy shell <url>

scarpa crawl zalando -o file.csv

Scrapy

Ethics

Supreme Court revives LinkedIn case to protect user data from web scrapers

Zack Whittaker @zackwhittaker 8:54 PM GMT+2 • June 14, 2021

Comment



Facebook: Stolen Data Scrapped from Platform in 2019



Author:

Elizabeth Montalbano

April 7, 2021 / 9:00 am

3 minute read

Write a comment

Share this article:



The flaw that caused the leak of personal data of more than 533 million users over the weekend no longer exists; however, the social media giant still faces an investigation by EU regulators.

Scraping Away at the CFAA—The Supreme Court’s Interpretation of “Exceeds Authorized Access” Limits the Scope of the Statute’s Application to Data Scrapers

June 17, 2021

In a long awaited decision that has a significant application for data scraping, the Supreme Court issued on June 3, 2021 its decision in *Van Buren v. United States*, significantly limiting the scope of the Computer Fraud and Abuse Act by holding that users who access information that they are entitled to obtain but use that information for improper purposes do not violate the statute. The majority opinion adopted a narrow interpretation of the statute that will make it more difficult to pursue both civil and criminal actions based on alleged misappropriation of data.



INFO FOR: CURRENT STUDENTS STAFF ALUMNI BENEFATORS LOG IN
Admissions Research Centres & Institutes News Events People About us Vacancies
Find programmes, pe...

Blog Home > The Legality of Screen Scraping and Its Open Banking Moment
23 Apr 2021
OBLB categories Financial Regulation OBLB Types Research
Han-Wei Liu
Lecturer at Monash University, Australia

MIT Technology Review

Featured Topics Newsletters Events Podcasts

Opinion

Web scraping is a tool, not a crime

Web scraping is one of the most powerful tools journalists have to hold companies and governments accountable.

by Lam Thuy Vo

December 8, 2020

Ethics

- Some websites do not like scraping. Why?
 - Increase load on web servers
 - Private websites
 - Dynamic websites
 - Data ownership
- It is important to "be recognised
 - No fraudulent behaviour (e.g. we insert a sleep(x) in our scripts)
 - Notify
 - Ask for the data (are we sure it is not already available in a structured format?)
- Always check API policies
 - Limits
 - Rate