Retrieving Data from the Web: Common Methods and Applications

Fiona Shen-Bayh

University of California, Berkeley

February 2017

Overview

- Content from the web
    - Finding online content
    - Structure of web pages
    - HTML versus JavaScript
- Web Scraping versus Web Driving

Data from the Web

A variety of data is becoming increasingly available on the web:

- International databases
- Digitized archives
- News content
- Blogs

Data from the Web

What is the most efficient way of gathering and organizing this data?

Web Scraping

We can use computer programs to automatically download or “scrape” online content.

Web Scraping

Web scraping is a method of data collection.

After we have downloaded, cleaned, and organized our data from the web, we can perform any kind of analysis – quantitative, qualitative – that we prefer.

Why Web Scrape?

Anything uploaded to a website can be downloaded: text, pictures, PDFs...

Automating this process is more time efficient than individually downloading files.

Web Scraping Challenges

There is no one-size-fits-all approach; web scraping programs need to be created or adjusted for individual websites.

Websites change, so you have to be ready to adapt your code.

Not all websites can be scraped.

Web Scraping in 3 Steps

1. Figure out the structure of the website you want to scrape.

2. Design a web scraper based on this structure.

3. Scrape the website, clean and organize the data, and save the output to your computer.

Step 1: What are we scraping?

All webpages are structured documents written in HTML language.

These HTML documents are organized into a set of directories. Any text, image, or document on the webpage is contained in one of these directories.

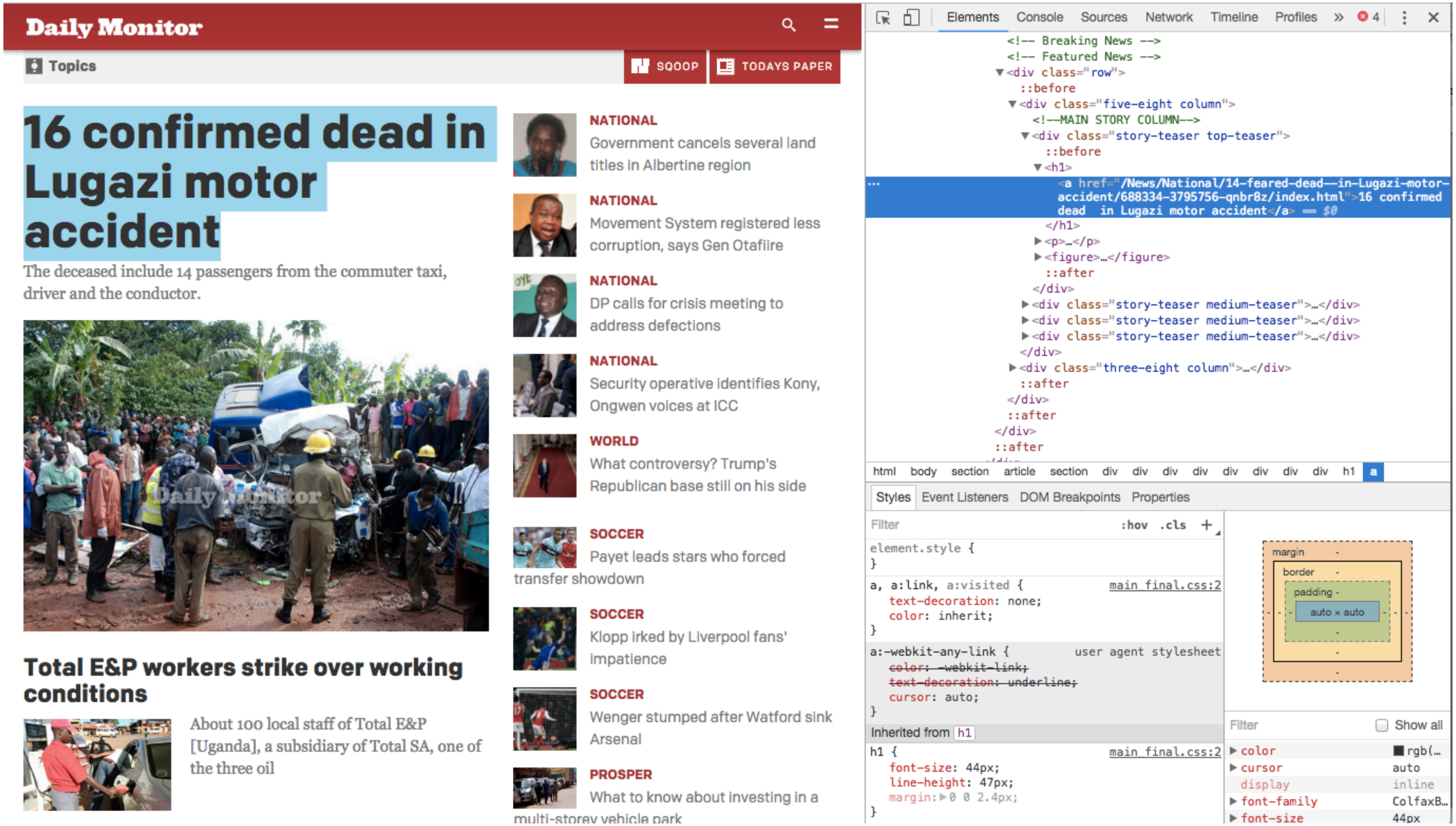
Step 1: What are we scraping?

Where can we find these directories?

Fortunately, there is a handy tool in most web browsers that help us figure this out.

Right-click on any webpage, click *inspect,* and you will see the inner-working HTML structure.

Inspecting an HTML page



16 confirmed dead in Lugazi motor accident

The deceased include 14 passengers from the commuter taxi, driver and the conductor.

Total E&P workers strike over working conditions

About 100 local staff of Total E&P [Uganda], a subsidiary of Total SA, one of the three oil

NATIONAL
Government cancels several land titles in Albertine region

NATIONAL
Movement System registered less corruption, says Gen Otafiire

NATIONAL
DP calls for crisis meeting to address defections

NATIONAL
Security operative identifies Kony, Ongwen voices at ICC

WORLD
What controversy? Trump's Republican base still on his side

SOCCER
Payet leads stars who forced transfer showdown

SOCCER
Klopp irked by Liverpool fans' impatience

SOCCER
Wenger stumped after Watford sink Arsenal

PROSPER
What to know about investing in a multi-storey vehicle park

```
      <!-- Breaking News -->
      <!-- Featured News -->
     ▼<div class="row">
        ::before
       ▼<div class="five-eight column">
          <!--MAIN STORY COLUMN-->
         ▼<div class="story-teaser top-teaser">
            ::before
           ▼<h1>
...           <a href="/News/National/14-feared-dead--in-Lugazi-motor-
              accident/688334-3795756-qnbr8z/index.html">16 confirmed
              dead  in Lugazi motor accident</a> == $0
            </h1>
           ▶<p>…</p>
           ▶<figure>…</figure>
            ::after
          </div>
         ▶<div class="story-teaser medium-teaser">…</div>
         ▶<div class="story-teaser medium-teaser">…</div>
         ▶<div class="story-teaser medium-teaser">…</div>
        </div>
       ▶<div class="three-eight column">…</div>
        ::after
      </div>
      ::after
    </div>
    ::after
```

html body section article section div div div div div div div div h1 a

```
element.style {
}
a, a:link, a:visited {                    main_final.css:2
   text-decoration: none;
   color: inherit;
}
a:-webkit-any-link {               user agent stylesheet
   ~~color: -webkit-link;~~
   ~~text-decoration: underline;~~
   cursor: auto;
}
Inherited from h1
h1 {                                      main_final.css:2
   font-size: 44px;
   line-height: 47px;
   margin:▶0 0 2.4px;
}
```

▶ color	■ rgb(…
▶ cursor	auto
display	inline
▶ font-family	ColfaxB…
▶ font-size	44px

Step 1: What are we scraping?

Using *inspect*, we can locate the specific directory path to the data we want. We will use this information when designing our web scraping program in Python.

Step 2: Design a Web Scraper

We can build our own web scraper in Python, which is a computer programming language.

Python is a highly readable language, where complex concepts can be simply expressed in relatively few lines of code.

Python is an “interpreter” between human programmers and computer operating systems.

Step 2: Design a Web Scraper

We need to use Python modules, which are files or scripts that contain Python definitions and statements.

Think of modules as special functions that can be loaded into the Python interpreter.

Step 2: Design a Web Scraper

We will use two modules to web scrape:

Requests loads a webpage and scrapes all of the HTML as a single, structured object.

Beautiful Soup parses the scraped object, meaning it re-organizes the HTML object into searchable components. Parsing helps us locate the data we want from within the HTML object.

Web Scraping Complications

While all websites are written in HTML, increasingly, websites are being programmed in JavaScript.

What does this mean?

HTML versus JavaScript

HTML language is *static.*

Used for general web development.

e.g. Basic outline of the webpage.

JavaScript language is *dynamic.*

Used for programming applications.

e.g. Animations and other interactive functions on the webpage.

HTML versus JavaScript

Unfortunately, standard Pythonic web scraping methods will not work on JavaScript webpages!

HTML versus JavaScript

Recall what web scraping is actually doing: visiting a webpage and retrieving its static HTML.

Web scraping is NOT interacting with dynamic web applications, meaning it cannot see what has been written in JavaScript.

If content you need was dynamically programmed, it will not be visible when you scrape the static webpage.

Solution: Web Drivers

There is a solution: program a web driver!

A web driver is like a robot that you can program to visit a web browser and interact with the live webpage.

From Python, you can program the web driver to open a new browser, poke around a JavaScript webpage, and download specific content directly to your computer.

Scraping versus Driving

A **scraper** scrapes all of content off of a static web page and then parses the output in Python.

A **driver** first interacts with a live webpage, then downloads specific content, and finally parses the output in Python.

Web Drivers

Web driving involves more preliminary steps than web scraping, but the output is the same.

And like web scraping, web driving can be automated over hundreds or thousands of webpages.

Scraping versus Driving

Webpages written in static HTML can also be downloaded using a web driver. So why do we even bother with web scrapers?

Web scrapers are easier to program and they work beautifully on sites written in HTML.

Web drivers are more complex and can take longer to run. But they work on JavaScript.

Scraping versus Driving

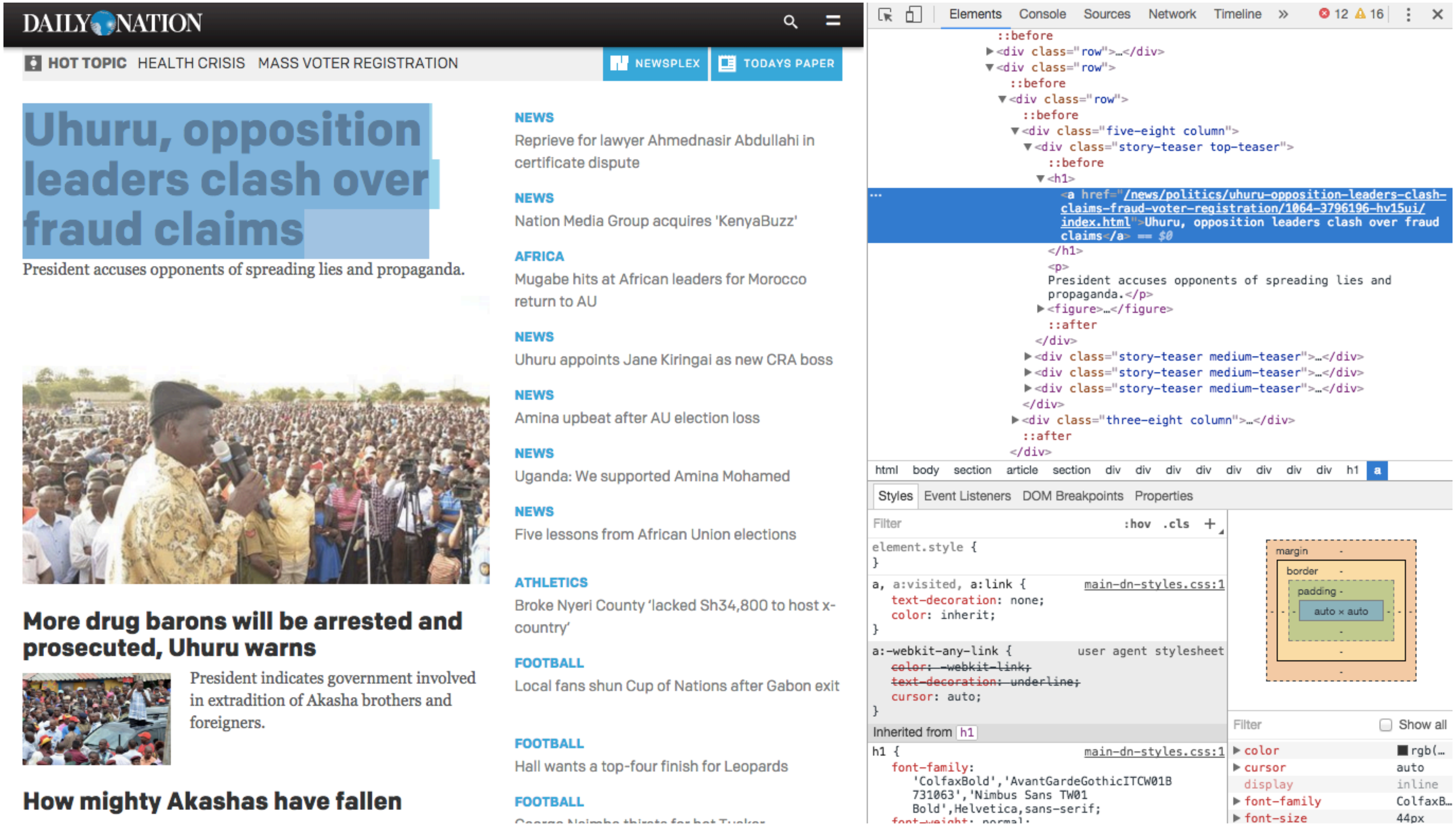
How do we know whether we should scrape or drive a given webpage?

Scrape HTML

Drive JavaScript

But how can we tell when a site is written in HTML or JavaScript?

Inspecting a JavaScript page



Uhuru, opposition leaders clash over fraud claims

President accuses opponents of spreading lies and propaganda.

More drug barons will be arrested and prosecuted, Uhuru warns

President indicates government involved in extradition of Akasha brothers and foreigners.

How mighty Akashas have fallen

NEWS
Reprieve for lawyer Ahmednasir Abdullahi in certificate dispute

NEWS
Nation Media Group acquires 'KenyaBuzz'

AFRICA
Mugabe hits at African leaders for Morocco return to AU

NEWS
Uhuru appoints Jane Kiringai as new CRA boss

NEWS
Amina upbeat after AU election loss

NEWS
Uganda: We supported Amina Mohamed

NEWS
Five lessons from African Union elections

ATHLETICS
Broke Nyeri County 'lacked Sh34,800 to host x-country'

FOOTBALL
Local fans shun Cup of Nations after Gabon exit

FOOTBALL
Hall wants a top-four finish for Leopards

FOOTBALL

```
          ::before
        ▶<div class="row">…</div>
        ▼<div class="row">
           ::before
          ▼<div class="row">
             ::before
            ▼<div class="five-eight column">
              ▼<div class="story-teaser top-teaser">
                 ::before
                ▼<h1>
...                <a href="/news/politics/uhuru-opposition-leaders-clash-
                   claims-fraud-voter-registration/1064-3796196-hv15ui/
                   index.html">Uhuru, opposition leaders clash over fraud
                   claims</a> == $0
                 </h1>
                 <p>
                 President accuses opponents of spreading lies and
                 propaganda.</p>
                ▶<figure>…</figure>
                 ::after
               </div>
              ▶<div class="story-teaser medium-teaser">…</div>
              ▶<div class="story-teaser medium-teaser">…</div>
              ▶<div class="story-teaser medium-teaser">…</div>
             </div>
            ▶<div class="three-eight column">…</div>
             ::after
           </div>
```

```
element.style {
}
a, a:visited, a:link {                  main-dn-styles.css:1
   text-decoration: none;
   color: inherit;
}
a:-webkit-any-link {                    user agent stylesheet
   ~~color: -webkit-link;~~
   ~~text-decoration: underline;~~
   cursor: auto;
}
Inherited from h1
h1 {                                    main-dn-styles.css:1
   font-family:
      'ColfaxBold','AvantGardeGothicITCW01B
      731063','Nimbus Sans TW01
      Bold',Helvetica,sans-serif;
   font-weight: normal;
```

```
▶ color          ■rgb(…
▶ cursor          auto
  display         inline
▶ font-family     ColfaxB…
▶ font-size       44px
```

Web Scrapers and Drivers

Let’s try it out!