web_scraping_wiki

May 17, 2019

Table of Contents

Autor: Matheus de Vasconcellos Barroso

1 Web scraping com Python

1.1 Introdução

Aprenderemos alguns conceitos básicos de **web scraping** e como utilizar o Python para essa tarefa. Mais precisamente esse material foi preparado utilizando um Jupyter Notebook e algumas extensões úteis (`nbextensions`). O exemplo que utilizaremos será a definição de **web scraping** pela Wikipedia:

In [48]:
```
#from IPython.display import display, HTML
from IPython.display import IFrame

IFrame(src = 'https://en.wikipedia.org/wiki/Web_scraping', width = 900, height=650)
```

Out[48]:
```
<IPython.lib.display.IFrame at 0x902e0f0>
```

Imagine como seria útil desenvolver uma técnica para ler o primeiro ou **n** primeiros parágrafos de uma página da Wikipedia para testar algorítmos de sumarização? Utilizando as técnicas de raspagem de dados podemos obter de forma simples os três primeiros parágrafos da url:

In [50]:
```
import requests
from bs4 import BeautifulSoup

page = requests.get("https://en.wikipedia.org/wiki/Web_scraping")
soup = BeautifulSoup(page.content, 'html.parser')
for item in soup.find_all('p')[:3]: print(item.get_text())
```

```
Web scraping, web harvesting, or web data extraction is data scraping used for extracting data

Web scraping a web page involves fetching it and extracting from it.[1][2] Fetching is the dow

Web scraping is used for contact scraping, and as a component of applications used for web ind
```

Para alcançar esse objetivo final serão abordados alguns tópicos necessários para compreender como funciona esse processo: + HTML + CSS (Não será abordado) + JS (Não será abordado)

Após um melhor entedimento teórico utilizaremos algumas bibliotecas do Python para realizar essa tarefa como `requests` e `bs4`. Algumas importante, mas que não serão utilizadas: `scrapy` e `selenium`

2 Conceitos básicos

2.1 HTML

HyperText Markup Language (HTML) ou Linguagem de Marcação de Hipertexto é uma linguaguem que server para a construção de páginas web. Ela serve para informar ao navegador como exibir o conteúdo da página.

O HTML é formado por tags, é importante compreendê-las e reconhecê-las já que fascilitam o trabalho de raspagem de dados na web. Algumas tags principais: - **html**: informar ao navegador aonde temos código em HTML - **head**: contêm informações como o título da página - **body**: é aonde o conteúdo principal da página está inserido, usualmente é onde o scraping ocorre. - **p**: delimita um parágrafo - **a**: para links - **div**: aponta uma região na página, útil para dividir o conteúdo - **b**: texto em negrito - **i**: texto em itálico - **table**: cria uma tablea - **form**: cria um formulário

2.2 CSS

Cascading Style Sheets (CSS) é um mecanismo para adicionar estilos (cores, fontes, espaçamento, etc.) a um documento web

2.3 JS

JavaScript (JS) é uma linguagem de programação que adiciona interatividade às paginas web.

2.4 Hands On

2.4.1 Explicando como extraimos a explicação da Wikipedia

Primeiro precisamos importas as bibliotecas:

```
In [ ]: import requests
        from bs4 import BeautifulSoup
```

Posteriormente, precisamos obter o conteúdo da página (HTML, CSS, JS). Podemos utilizar o método `get`do módulo **requests** para a *url* desejada e atribuí-lo à variável 'page':

```
In [51]: page = requests.get("https://en.wikipedia.org/wiki/Web_scraping")
```

Se a conexão for bem sucedida teremos um status 200 para a página:

```
In [52]: page.status_code
```

```
Out[52]: 200
```

Podemos obter o conteúdo da página:

```
In [53]: page.content
```

```
Out[53]: b'<!DOCTYPE html>\n<html class="client-nojs" dir="ltr" lang="en">\n<head>\n<meta charset="utf-8"/>\n<title>Web scraping - Wikipedia</title>\n<script>document.documentElement.className = document.documentElement.className.replac
<script>(window.RLQ=window.RLQ||[]).push(function(){mw.config.set({"wgCanonicalNamespa
});RLPAGEMODULES=["ext.cite.ux-enhancements","site","mediawiki.page.startup","mediawi
<link href="/w/load.php?lang=en&amp;modules=ext.3d.styles%7Cext.cite.styles%7Cext.uls
<script async="" src="/w/load.php?lang=en&amp;modules=startup&amp;only=scripts&amp;ski
<meta content="" name="ResourceLoaderDynamicStyles"/>
<link href="/w/load.php?lang=en&amp;modules=ext.gadget.charinsert-styles&amp;only=styl
<link href="/w/load.php?lang=en&amp;modules=site.styles&amp;only=styles&amp;skin=vecto
<meta content="MediaWiki 1.34.0-wmf.4" name="generator"/>
<meta content="origin" name="referrer"/>
<meta content="origin-when-crossorigin" name="referrer"/>
<meta content="origin-when-cross-origin" name="referrer"/>
<link href="android-app://org.wikipedia/http/en.m.wikipedia.org/wiki/Web_scraping" rel
<link href="/w/index.php?title=Web_scraping&amp;action=edit" rel="alternate" title="Ed
<link href="/w/index.php?title=Web_scraping&amp;action=edit" rel="edit" title="Edit th
<link href="/static/apple-touch/wikipedia.png" rel="apple-touch-icon"/>
<link href="/static/favicon/wikipedia.ico" rel="shortcut icon"/>
<link href="/w/opensearch_desc.php" rel="search" title="Wikipedia (en)" type="applicat
<link href="//en.wikipedia.org/w/api.php?action=rsd" rel="EditURI" type="application/r
<link href="//creativecommons.org/licenses/by-sa/3.0/" rel="license"/>
<link href="https://en.wikipedia.org/wiki/Web_scraping" rel="canonical"/>
<link href="//login.wikimedia.org" rel="dns-prefetch"/>
<link href="//meta.wikimedia.org" rel="dns-prefetch"/>
```

```
In [59]: from bs4 import BeautifulSoup

         soup = BeautifulSoup(page.content, 'html.parser')
         soup
```

```
Out[59]: <!DOCTYPE html>

         <html class="client-nojs" dir="ltr" lang="en">
         <head>
         <meta charset="utf-8"/>
         <title>Web scraping - Wikipedia</title>
         <script>document.documentElement.className = document.documentElement.className.replac
         <script>(window.RLQ=window.RLQ||[]).push(function(){mw.config.set({"wgCanonicalNamespa
         });RLPAGEMODULES=["ext.cite.ux-enhancements","site","mediawiki.page.startup","mediawi
         <link href="/w/load.php?lang=en&amp;modules=ext.3d.styles%7Cext.cite.styles%7Cext.uls
         <script async="" src="/w/load.php?lang=en&amp;modules=startup&amp;only=scripts&amp;ski
         <meta content="" name="ResourceLoaderDynamicStyles"/>
         <link href="/w/load.php?lang=en&amp;modules=ext.gadget.charinsert-styles&amp;only=styl
         <link href="/w/load.php?lang=en&amp;modules=site.styles&amp;only=styles&amp;skin=vecto
         <meta content="MediaWiki 1.34.0-wmf.4" name="generator"/>
         <meta content="origin" name="referrer"/>
         <meta content="origin-when-crossorigin" name="referrer"/>
         <meta content="origin-when-cross-origin" name="referrer"/>
         <link href="android-app://org.wikipedia/http/en.m.wikipedia.org/wiki/Web_scraping" rel
         <link href="/w/index.php?title=Web_scraping&amp;action=edit" rel="alternate" title="Ed
         <link href="/w/index.php?title=Web_scraping&amp;action=edit" rel="edit" title="Edit th
         <link href="/static/apple-touch/wikipedia.png" rel="apple-touch-icon"/>
         <link href="/static/favicon/wikipedia.ico" rel="shortcut icon"/>
         <link href="/w/opensearch_desc.php" rel="search" title="Wikipedia (en)" type="applicat
         <link href="//en.wikipedia.org/w/api.php?action=rsd" rel="EditURI" type="application/r
         <link href="//creativecommons.org/licenses/by-sa/3.0/" rel="license"/>
         <link href="https://en.wikipedia.org/wiki/Web_scraping" rel="canonical"/>
         <link href="//login.wikimedia.org" rel="dns-prefetch"/>
         <link href="//meta.wikimedia.org" rel="dns-prefetch"/>
```

```
<!--[if lt IE 9]><script src="/w/load.php?lang=qqx&amp;modules=html5shiv&amp;only=scri
</head>
<body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject mw-editable p
<div class="noprint" id="mw-page-base"></div>
<div class="noprint" id="mw-head-base"></div>
<div class="mw-body" id="content" role="main">
<a id="top"></a>
<div class="mw-body-content" id="siteNotice"><!-- CentralNotice --></div>
<div class="mw-indicators mw-body-content">
</div>
<h1 class="firstHeading" id="firstHeading" lang="en">Web scraping</h1>
<div class="mw-body-content" id="bodyContent">
<div class="noprint" id="siteSub">From Wikipedia, the free encyclopedia</div>
<div id="contentSub"></div>
<div id="jump-to-nav"></div>
<a class="mw-jump-link" href="#mw-head">Jump to navigation</a>
<a class="mw-jump-link" href="#p-search">Jump to search</a>
<div class="mw-content-ltr" dir="ltr" id="mw-content-text" lang="en"><div class="mw-pa
<div class="hatnote navigation-not-searchable" role="note">For broader coverage of thi
<p><b>Web scraping</b>, <b>web harvesting</b>, or <b>web data extraction</b> is <a hre
</p><p>Web scraping a web page involves fetching it and extracting from it.<sup class=
</p><p>Web scraping is used for <a href="/wiki/Contact_scraping" title="Contact scrapi
</p><p><a href="/wiki/Web_page" title="Web page">Web pages</a> are built using text-ba
</p><p>Newer forms of web scraping involve listening to data feeds from web servers.
</p><p>There are methods that some websites use to prevent web scraping, such as detec
</p>
<div class="toc" id="toc"><input class="toctogglecheckbox" id="toctogglecheckbox" role
<ul>
<li class="toclevel-1 tocsection-1"><a href="#History"><span class="tocnumber">1</span
<li class="toclevel-1 tocsection-2"><a href="#Techniques"><span class="tocnumber">2</s
<ul>
<li class="toclevel-2 tocsection-3"><a href="#Human_copy-and-paste"><span class="tocnu
<li class="toclevel-2 tocsection-4"><a href="#Text_pattern_matching"><span class="tocn
<li class="toclevel-2 tocsection-5"><a href="#HTTP_programming"><span class="tocnumber
<li class="toclevel-2 tocsection-6"><a href="#HTML_parsing"><span class="tocnumber">2
<li class="toclevel-2 tocsection-7"><a href="#DOM_parsing"><span class="tocnumber">2.5
<li class="toclevel-2 tocsection-8"><a href="#Vertical_aggregation"><span class="tocnu
<li class="toclevel-2 tocsection-9"><a href="#Semantic_annotation_recognizing"><span c
<li class="toclevel-2 tocsection-10"><a href="#Computer_vision_web-page_analysis"><spa
</ul>
</li>
<li class="toclevel-1 tocsection-11"><a href="#Software"><span class="tocnumber">3</sp
<ul>
<li class="toclevel-2 tocsection-12"><a href="#Example_tools"><span class="tocnumber">
<ul>
<li class="toclevel-3 tocsection-13"><a href="#Javascript_tools"><span class="tocnumbe
<li class="toclevel-3 tocsection-14"><a href="#Web_crawling_frameworks"><span class="t
</ul>
```

```
</li>
</ul>
</li>
<li class="toclevel-1 tocsection-15"><a href="#Legal_issues"><span class="tocnumber">4
<ul>
<li class="toclevel-2 tocsection-16"><a href="#United_States"><span class="tocnumber">
<li class="toclevel-2 tocsection-17"><a href="#The_EU"><span class="tocnumber">4.2</sp
<li class="toclevel-2 tocsection-18"><a href="#Australia"><span class="tocnumber">4.3
</ul>
</li>
<li class="toclevel-1 tocsection-19"><a href="#Methods_to_prevent_web_scraping"><span
<li class="toclevel-1 tocsection-20"><a href="#See_also"><span class="tocnumber">6</sp
<li class="toclevel-1 tocsection-21"><a href="#References"><span class="tocnumber">7</
</ul>
</div>
<h2><span class="mw-headline" id="History">History</span><span class="mw-editsection">
<table class="box-Unreferenced_section plainlinks metadata ambox ambox-content ambox-U
<p>The history of the web scraping is actually much longer, dating back significantly
</p>
<ul><li>After the birth of <a href="/wiki/History_of_the_World_Wide_Web" title="Histor
<li>In 1993,December, the First <b>crawler-based web search engine</b> - <a href="/wik
<li>In 2000, the <b>first Web API and API crawler</b> came. <a href="/wiki/Application
<li>In 2004, <a href="/wiki/Beautiful_Soup_(HTML_parser)" title="Beautiful Soup (HTML
<h2><span class="mw-headline" id="Techniques">Techniques</span><span class="mw-editsec
<p>Web scraping is the process of automatically mining data or collecting information
</p>
<h3><span class="mw-headline" id="Human_copy-and-paste">Human copy-and-paste</span><sp
<p>Sometimes even the best web-scraping technology cannot replace a humans manual exam
</p>
<h3><span class="mw-headline" id="Text_pattern_matching">Text pattern matching</span><
<p>A simple yet powerful approach to extract information from web pages can be based o
</p>
<h3><span class="mw-headline" id="HTTP_programming">HTTP programming</span><span class
<p><a href="/wiki/Static_web_page" title="Static web page">Static</a> and <a href="/wi
</p>
<h3><span class="mw-headline" id="HTML_parsing">HTML parsing</span><span class="mw-edi
<p>Many websites have large collections of pages generated dynamically from an underly
</p>
<h3><span class="mw-headline" id="DOM_parsing">DOM parsing</span><span class="mw-edits
<div class="hatnote navigation-not-searchable" role="note">Further information: <a hre
<p>By embedding a full-fledged web browser, such as the <a href="/wiki/Internet_Explor
</p>
<h3><span class="mw-headline" id="Vertical_aggregation">Vertical aggregation</span><sp
<p>There are several companies that have developed vertical specific harvesting platfo
</p>
<h3><span class="mw-headline" id="Semantic_annotation_recognizing">Semantic annotation
<p>The pages being scraped may embrace <a href="/wiki/Metadata" title="Metadata">metad
</p>
```

<h3>Computer vision w
There are efforts using <a href="/wiki/Machine_learning" title="Machine learning">m

<h2>Software<span class="mw-editsectio
There are many software tools available that can be used to customize web-scraping

<h3>Example tools<span class="mw-e
cURL command line tool and library for
Data Toolbar web scraping
Diffbot uses computer vision and ma
Heritrix gets pages (lots of them)
HtmlUnit headless browser that can
HTTrack free and open source Web craw
iMacros a browser extension to record
<a href="/wiki/Selenium_(software)" title="Selenium (software)">Selenium (software
Jaxer
Mozenda is a WYSIWYG software that c
nokogiri
OutWit Hub Web scraping applica
watir
Wget computer program that retrieves conten
<a href="/wiki/WSO2_Mashup_Server" title="WSO2 Mashup Server">WSO2 Mashup Server</
<a href="/wiki/Yahoo!_Query_Language" title="Yahoo! Query Language">Yahoo! Query L
<h4>Javascript tools<span class
Greasemonkey
Node.js
PhantomJS scripted, <a href="/wi
jQuery
<h4>Web crawling frameworks</sp
These can be used to build web scrapers.

Scrapy
<h2>Legal issues<span class="mw-edi
<table class="box-Globalize plainlinks metadata ambox ambox-content ambox-globalize" r
The legality of web scraping varies across the world. In general, web scraping may

<h3>United States<span class="mw-e
<table class="plainlinks metadata ambox ambox-move" role="presentation"><tbody><tr><td
In the United States, website owners can use three major <a href="/wiki/Cause_of_ac
U.S. courts have acknowledged that users of "scrapers" or "robots" may be held
One of the first major tests of <a class="mw-redirect" href="/wiki/Screen_scrap
<a href="/wiki/Southwest_Airlines" title="Southwest Airlines">Southwest Airline
In 2012, a startup called 3Taps scraped classified housing ads from Craigslist. Craigs
Although these are early scraping decisions, and the theories of liability are
While the law in this area becomes more settled, entities contemplating using s
In the plaintiff's web site during the period of this trial the terms of use li
In <i><a href="/wiki/Facebook,_Inc._v._Power_Ventures,_Inc." title="Facebook, I
Internet Archive collects and distributes significant number of publicly availa

```
</p>
<h3><span class="mw-headline" id="The_EU">The EU</span><span class="mw-editsection"><s
<p>In February 2006, the Danish Maritime and Commercial Court (Copenhagen) ruled that
</p><p>In a February 2010 case complicated by matters of jurisdiction, Ireland's High
<sup class="reference" id="cite_ref-22"><a href="#cite_note-22">[22]</a></sup> The dec
</p>
<h3><span class="mw-headline" id="Australia">Australia</span><span class="mw-editsecti
<p>In Australia, the <a href="/wiki/Spam_Act_2003" title="Spam Act 2003">Spam Act 2003
</p>
<h2><span class="mw-headline" id="Methods_to_prevent_web_scraping">Methods to prevent
<p>The administrator of a website can use various measures to stop or slow a bot. Some
</p>
<ul><li>Blocking an <a href="/wiki/IP_address" title="IP address">IP address</a> eithe
<li>Disabling any <a href="/wiki/Web_service" title="Web service">web service</a> <a h
<li>Bots sometimes declare who they are (using <a href="/wiki/User_agent" title="User
<li>Bots can be blocked by monitoring excess traffic</li>
<li>Bots can sometimes be blocked with tools to verify that it is a real person access
<li>Commercial anti-bot services: Companies offer anti-bot and anti-scraping services
<ul><li>Locating bots with a <a href="/wiki/Honeypot_(computing)" title="Honeypot (com
<li><a href="/wiki/Obfuscation" title="Obfuscation">Obfuscation</a> using <a class="mw
<li>Because bots rely on consistency in the front-end code of a target website, adding
<li>Websites can declare if crawling is allowed or not in the <a href="/wiki/Robots_ex
<h2><span class="mw-headline" id="See_also">See also</span><span class="mw-editsection
<div class="div-col columns column-width" style="-moz-column-width: 22em; -webkit-colu
<ul><li><a class="mw-redirect" href="/wiki/Archive.is" title="Archive.is">Archive.is</
<li><a href="/wiki/Comparison_of_feed_aggregators" title="Comparison of feed aggregato
<li><a href="/wiki/Data_scraping" title="Data scraping">Data scraping</a></li>
<li><a href="/wiki/Data_wrangling" title="Data wrangling">Data wrangling</a></li>
<li><a href="/wiki/Importer_(computing)" title="Importer (computing)">Importer</a></li
<li><a href="/wiki/Job_wrapping" title="Job wrapping">Job wrapping</a></li>
<li><a href="/wiki/Knowledge_extraction" title="Knowledge extraction">Knowledge extrac
<li><a href="/wiki/OpenSocial" title="OpenSocial">OpenSocial</a></li>
<li><a href="/wiki/Scraper_site" title="Scraper site">Scraper site</a></li>
<li><a href="/wiki/Fake_news_website" title="Fake news website">Fake news website</a><
<li><a href="/wiki/Blog_scraping" title="Blog scraping">Blog scraping</a></li>
<li><a href="/wiki/Spamdexing" title="Spamdexing">Spamdexing</a></li>
<li><a href="/wiki/Domain_name_drop_list" title="Domain name drop list">Domain name dr
<li><a href="/wiki/Text_corpus" title="Text corpus">Text corpus</a></li>
<li><a href="/wiki/Web_archiving" title="Web archiving">Web archiving</a></li>
<li><a class="mw-redirect" href="/wiki/Blog_network" title="Blog network">Blog network
<li><a class="mw-redirect" href="/wiki/Search_Engine_Scraping" title="Search Engine Sc
<li><a href="/wiki/Category:Web_crawlers" title="Category:Web crawlers">Web crawlers</
</div>
<h2><span class="mw-headline" id="References">References</span><span class="mw-editsec
<div class="reflist columns references-column-width" style="-moz-column-width: 20em; -
<ol class="references">
<li id="cite_note-Boeing2016JPER-1"><span class="mw-cite-backlink">^ <a href="#cite_re
</li>
```

<li id="cite_note-2"><b>^</b

<li id="cite_note-3"><b>^</b

<li id="cite_note-4"><b>^</b

<li id="cite_note-5"><b>^</b

<li id="cite_note-6"><b>^</b

<li id="cite_note-7"><b>^</b

<li id="cite_note-8"><b>^</b

<li id="cite_note-9"><b>^</b

<li id="cite_note-10"><b>^<

<li id="cite_note-11"><b>^<

<li id="cite_note-12"><b>^<

<li id="cite_note-impervawp2011-13"><b><a href="#cite_

<li id="cite_note-14"><b>^<

<li id="cite_note-15"><b>^<

<li id="cite_note-16"><b>^<

<li id="cite_note-17"><b>^<

<li id="cite_note-18"><b>^<

<li id="cite_note-19"><b>^<

<li id="cite_note-20"><b>^<

<li id="cite_note-21"><b>^<

<li id="cite_note-22"><b>^<

<li id="cite_note-23"><b>^<

<li id="cite_note-24"><b>^<

<li id="cite_note-25"><b>^<


```
<li id="cite_note-26"><span class="mw-cite-backlink"><b><a href="#cite_ref-26">^</a></
</li>
</ol></div>
<!--
NewPP limit report
Parsed by mw1264
Cached time: 20190516030922
Cache expiry: 2592000
Dynamic content: false
CPU time usage: 0.492 seconds
Real time usage: 0.660 seconds
Preprocessor visited node count: 1805/1000000
Preprocessor generated node count: 0/1500000
Postexpand include size: 65153/2097152 bytes
Template argument size: 772/2097152 bytes
Highest expansion depth: 15/40
Expensive parser function count: 8/500
Unstrip recursion depth: 1/20
Unstrip postexpand size: 70826/5000000 bytes
Number of Wikibase entities loaded: 3/400
Lua time usage: 0.279/10.000 seconds
Lua memory usage: 6.63 MB/50 MB
-->
<!--
Transclusion expansion time report (%,ms,calls,template)
100.00%  548.078      1 -total
 59.50%  326.090      1 Template:Reflist
 38.08%  208.730      5 Template:Cite_journal
 19.94%  109.301      1 Template:More_citations_needed
 18.46%  101.199      4 Template:Ambox
 16.32%   89.451     18 Template:Cite_web
  6.60%   36.163      1 Template:Find_sources_mainspace
  5.52%   30.244      1 Template:Split_section
  4.82%   26.437      1 Template:Split_portions
  2.82%   15.468      1 Template:US-centric
-->
<!-- Saved in parser cache with key enwiki:pcache:idhash:2696619-0!canonical and times
 -->
</div><noscript><img alt="" height="1" src="//en.wikipedia.org/wiki/Special:CentralAu
<div class="printfooter">Retrieved from "<a dir="ltr" href="https://en.wikipedia.org/w
<div class="catlinks" data-mw="interface" id="catlinks"><div class="mw-normal-catlinks
<div class="visualClear"></div>
</div>
</div>
<div id="mw-navigation">
<h2>Navigation menu</h2>
<div id="mw-head">
<div aria-labelledby="p-personal-label" id="p-personal" role="navigation">
```

```
<h3 id="p-personal-label">Personal tools</h3>
<ul>
<li id="pt-anonuserpage">Not logged in</li><li id="pt-anontalk"><a accesskey="n" href=
</div>
<div id="left-navigation">
<div aria-labelledby="p-namespaces-label" class="vectorTabs" id="p-namespaces" role="
<h3 id="p-namespaces-label">Namespaces</h3>
<ul>
<li class="selected" id="ca-nstab-main"><span><a accesskey="c" href="/wiki/Web_scrapi
</div>
<div aria-labelledby="p-variants-label" class="vectorMenu emptyPortlet" id="p-variants
<input aria-labelledby="p-variants-label" class="vectorMenuCheckbox" type="checkbox"/
<h3 id="p-variants-label">
<span>Variants</span>
</h3>
<ul class="menu">
</ul>
</div>
</div>
<div id="right-navigation">
<div aria-labelledby="p-views-label" class="vectorTabs" id="p-views" role="navigation"
<h3 id="p-views-label">Views</h3>
<ul>
<li class="collapsible selected" id="ca-view"><span><a href="/wiki/Web_scraping">Read<
</div>
<div aria-labelledby="p-cactions-label" class="vectorMenu emptyPortlet" id="p-cactions
<input aria-labelledby="p-cactions-label" class="vectorMenuCheckbox" type="checkbox"/
<h3 id="p-cactions-label"><span>More</span></h3>
<ul class="menu">
</ul>
</div>
<div id="p-search" role="search">
<h3>
<label for="searchInput">Search</label>
</h3>
<form action="/w/index.php" id="searchform">
<div id="simpleSearch">
<input accesskey="f" id="searchInput" name="search" placeholder="Search Wikipedia" ti
</form>
</div>
</div>
</div>
<div id="mw-panel">
<div id="p-logo" role="banner"><a class="mw-wiki-logo" href="/wiki/Main_Page" title="
<div aria-labelledby="p-navigation-label" class="portal" id="p-navigation" role="navig
<h3 id="p-navigation-label">Navigation</h3>
<div class="body">
<ul>
```

<li id="n-mainpage-description"><a accesskey="z" href="/wiki/Main_Page" title="Visit

<div aria-labelledby="p-interaction-label" class="portal" id="p-interaction" role="nav
<h3 id="p-interaction-label">Interaction</h3>

<li id="n-help"><a href="/wiki/Help:Contents" title="Guidance on how to use and edit

<h3 id="p-tb-label">Tools</h3>

<li id="t-whatlinkshere"><a accesskey="j" href="/wiki/Special:WhatLinksHere/Web_scrapi

<div aria-labelledby="p-coll-print_export-label" class="portal" id="p-coll-print_expor
<h3 id="p-coll-print_export-label">Print/export</h3>

<li id="coll-create_a_book"><a href="/w/index.php?title=Special:Book&bookcmd=book_

<h3 id="p-lang-label">Languages</h3>

<li class="interlanguage-link interwiki-ar"><a class="interlanguage-link-target" href=
<span class="wb-langlinks-edit wb-langli

<ul id="footer-info">
<li id="footer-info-lastmod"> This page was last edited on 15 May 2019, at 09:54<span
<li id="footer-info-copyright">Text is available under the <a href="//en.wikipedia.org
additional terms may apply. By using this site, you agree to the <a href="//foundatio

<ul id="footer-places">
<li id="footer-places-privacy"><a class="extiw" href="https://foundation.wikimedia.org
<li id="footer-places-about"><a href="/wiki/Wikipedia:About" title="Wikipedia:About">A
<li id="footer-places-disclaimer"><a href="/wiki/Wikipedia:General_disclaimer" title="
<li id="footer-places-contact"><a href="//en.wikipedia.org/wiki/Wikipedia:Contact_us">
<li id="footer-places-developers"><a href="https://www.mediawiki.org/wiki/Special:MyLa
<li id="footer-places-cookiestatement"><a href="https://foundation.wikimedia.org/wiki/
<li id="footer-places-mobileview"><a class="noprint stopMobileRedirectToggle" href="/


```
        <ul class="noprint" id="footer-icons">
        <li id="footer-copyrightico">
        <a href="https://wikimediafoundation.org/"><img alt="Wikimedia Foundation" height="31"
        <li id="footer-poweredbyico">
        <a href="//www.mediawiki.org/"><img alt="Powered by MediaWiki" height="31" src="/stati
        </ul>
        <div style="clear: both;"></div>
        </div>
        <script>(window.RLQ=window.RLQ||[]).push(function(){mw.config.set({"wgPageParseReport"
        <script type="application/ld+json">{"@context":"https:\/\/schema.org","@type":"Article
        <script>(window.RLQ=window.RLQ||[]).push(function(){mw.config.set({"wgBackendResponseT
        </body>
        </html>
```

É possível melhorar ainda mais usando o método `prettify`

```
In [57]: print(soup.prettify())
```

```
<!DOCTYPE html>
<html class="client-nojs" dir="ltr" lang="en">
 <head>
  <meta charset="utf-8"/>
  <title>
   Web scraping - Wikipedia
  </title>
  <script>
   document.documentElement.className = document.documentElement.className.replace( /(^|\s)clie
  </script>
  <script>
   (window.RLQ=window.RLQ||[]).push(function(){mw.config.set({"wgCanonicalNamespace":"","wgCano
});RLPAGEMODULES=["ext.cite.ux-enhancements","site","mediawiki.page.startup","mediawiki.page.re
  </script>
  <link href="/w/load.php?lang=en&amp;modules=ext.3d.styles%7Cext.cite.styles%7Cext.uls.interla
  <script async="" src="/w/load.php?lang=en&amp;modules=startup&amp;only=scripts&amp;skin=vecto
  </script>
  <meta content="" name="ResourceLoaderDynamicStyles"/>
  <link href="/w/load.php?lang=en&amp;modules=ext.gadget.charinsert-styles&amp;only=styles&amp
  <link href="/w/load.php?lang=en&amp;modules=site.styles&amp;only=styles&amp;skin=vector" rel=
  <meta content="MediaWiki 1.34.0-wmf.4" name="generator"/>
  <meta content="origin" name="referrer"/>
  <meta content="origin-when-crossorigin" name="referrer"/>
  <meta content="origin-when-cross-origin" name="referrer"/>
  <link href="android-app://org.wikipedia/http/en.m.wikipedia.org/wiki/Web_scraping" rel="alter
  <link href="/w/index.php?title=Web_scraping&amp;action=edit" rel="alternate" title="Edit this
  <link href="/w/index.php?title=Web_scraping&amp;action=edit" rel="edit" title="Edit this page
  <link href="/static/apple-touch/wikipedia.png" rel="apple-touch-icon"/>
  <link href="/static/favicon/wikipedia.ico" rel="shortcut icon"/>
  <link href="/w/opensearch_desc.php" rel="search" title="Wikipedia (en)" type="application/ope
```

```
 <link href="//en.wikipedia.org/w/api.php?action=rsd" rel="EditURI" type="application/rsd+xml"
 <link href="//creativecommons.org/licenses/by-sa/3.0/" rel="license"/>
 <link href="https://en.wikipedia.org/wiki/Web_scraping" rel="canonical"/>
 <link href="//login.wikimedia.org" rel="dns-prefetch"/>
 <link href="//meta.wikimedia.org" rel="dns-prefetch"/>
 <!--[if lt IE 9]><script src="/w/load.php?lang=qqx&amp;modules=html5shiv&amp;only=scripts&amp
</head>
<body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject mw-editable page-Web_
 <div class="noprint" id="mw-page-base">
 </div>
 <div class="noprint" id="mw-head-base">
 </div>
 <div class="mw-body" id="content" role="main">
  <a id="top">
  </a>
  <div class="mw-body-content" id="siteNotice">
   <!-- CentralNotice -->
  </div>
  <div class="mw-indicators mw-body-content">
  </div>
  <h1 class="firstHeading" id="firstHeading" lang="en">
   Web scraping
  </h1>
  <div class="mw-body-content" id="bodyContent">
   <div class="noprint" id="siteSub">
    From Wikipedia, the free encyclopedia
   </div>
   <div id="contentSub">
   </div>
   <div id="jump-to-nav">
   </div>
   <a class="mw-jump-link" href="#mw-head">
    Jump to navigation
   </a>
   <a class="mw-jump-link" href="#p-search">
    Jump to search
   </a>
   <div class="mw-content-ltr" dir="ltr" id="mw-content-text" lang="en">
    <div class="mw-parser-output">
     <table class="box-More_citations_needed plainlinks metadata ambox ambox-content ambox-Re
      <tbody>
       <tr>
        <td class="mbox-image">
         <div style="width:52px">
          <a class="image" href="/wiki/File:Question_book-new.svg">
           <img alt="" data-file-height="399" data-file-width="512" decoding="async" height="
          </a>
         </div>
```

```
</td>
<td class="mbox-text">
 <div class="mbox-text-span">
  This article
  <b>
   needs additional citations for
   <a href="/wiki/Wikipedia:Verifiability" title="Wikipedia:Verifiability">
    verification
   </a>
  </b>
  .
  <span class="hide-when-compact">
   Please help
   <a class="external text" href="//en.wikipedia.org/w/index.php?title=Web_scraping&a
    improve this article
   </a>
   by
   <a href="/wiki/Help:Introduction_to_referencing_with_Wiki_Markup/1" title="Help:In
    adding citations to reliable sources
   </a>
   . Unsourced material may be challenged and removed.
   <br/>
   <small>
    <span class="plainlinks">
     <i>
      Find sources:
     </i>
     <a class="external text" href="//www.google.com/search?as_eq=wikipedia&amp;q=%22
      "Web scraping"
     </a>

     <a class="external text" href="//www.google.com/search?tbm=nws&amp;q=%22Web+scra
      news
     </a>
     <b>
      ů
     </b>
     <a class="external text" href="//www.google.com/search?&amp;q=%22Web+scraping%22
      newspapers
     </a>
     <b>
      ů
     </b>
     <a class="external text" href="//www.google.com/search?tbs=bks:1&amp;q=%22Web+sc
      books
     </a>
     <b>
      ů
```

```
        </b>
        <a class="external text" href="//scholar.google.com/scholar?q=%22Web+scraping%22
         scholar
        </a>
        <b>
         ů
        </b>
        <a class="external text" href="https://www.jstor.org/action/doBasicSearch?Query=
         JSTOR
        </a>
       </span>
      </small>
     </span>
     <small class="date-container">
      <i>
       (
       <span class="date">
        June 2017
       </span>
       )
      </i>
     </small>
     <small class="hide-when-compact">
      <i>
       (
       <a href="/wiki/Help:Maintenance_template_removal" title="Help:Maintenance templat
        Learn how and when to remove this template message
       </a>
       )
      </i>
     </small>
    </div>
   </td>
  </tr>
 </tbody>
</table>
<div class="hatnote navigation-not-searchable" role="note">
 For broader coverage of this topic, see
 <a href="/wiki/Data_scraping" title="Data scraping">
  Data scraping
 </a>
 .
</div>
<p>
 <b>
  Web scraping
 </b>
 ,
```

```
<b>
 web harvesting
</b>
, or
<b>
 web data extraction
</b>
is
<a href="/wiki/Data_scraping" title="Data scraping">
 data scraping
</a>
used for
<a href="/wiki/Data_extraction" title="Data extraction">
 extracting data
</a>
from
<a href="/wiki/Website" title="Website">
 websites
</a>
.
<sup class="reference" id="cite_ref-Boeing2016JPER_1-0">
 <a href="#cite_note-Boeing2016JPER-1">
  [1]
 </a>
</sup>
Web scraping software may access the World Wide Web directly using the
<a href="/wiki/Hypertext_Transfer_Protocol" title="Hypertext Transfer Protocol">
 Hypertext Transfer Protocol
</a>
, or through a web browser. While web scraping can be done manually by a software user,
<a href="/wiki/Internet_bot" title="Internet bot">
 bot
</a>
or
<a href="/wiki/Web_crawler" title="Web crawler">
 web crawler
</a>
. It is a form of copying, in which specific data is gathered and copied from the web, t
<a href="/wiki/Database" title="Database">
 database
</a>
or spreadsheet, for later
<a href="/wiki/Data_retrieval" title="Data retrieval">
 retrieval
</a>
or
<a href="/wiki/Data_analysis" title="Data analysis">
 analysis
```

```
 </a>
 .
</p>
<p>
 Web scraping a web page involves fetching it and extracting from it.
 <sup class="reference" id="cite_ref-Boeing2016JPER_1-1">
  <a href="#cite_note-Boeing2016JPER-1">
   [1]
  </a>
 </sup>
 <sup class="reference" id="cite_ref-2">
  <a href="#cite_note-2">
   [2]
  </a>
 </sup>
 Fetching is the downloading of a page (which a browser does when you view the page). The
 <a href="/wiki/Parsing" title="Parsing">
  parsed
 </a>
 , searched, reformatted, its data copied into a spreadsheet, and so on. Web scrapers typ
</p>
<p>
 Web scraping is used for
 <a href="/wiki/Contact_scraping" title="Contact scraping">
  contact scraping
 </a>
 , and as a component of applications used for
 <a href="/wiki/Web_indexing" title="Web indexing">
  web indexing
 </a>
 ,
 <a href="/wiki/Web_mining" title="Web mining">
  web mining
 </a>
 and
 <a href="/wiki/Data_mining" title="Data mining">
  data mining
 </a>
 , online price change monitoring and
 <a href="/wiki/Comparison_shopping_website" title="Comparison shopping website">
  price comparison
 </a>
 , product review scraping (to watch the competition), gathering real estate listings, we
 <a href="/wiki/Change_detection_and_notification" title="Change detection and notificati
  website change detection
 </a>
 , research, tracking online presence and reputation,
 <a class="mw-redirect" href="/wiki/Web_mashup" title="Web mashup">
```

```
 web mashup
</a>
and,
<a href="/wiki/Web_data_integration" title="Web data integration">
 web data integration
</a>
.
</p>
<p>
<a href="/wiki/Web_page" title="Web page">
 Web pages
</a>
are built using text-based mark-up languages (
<a href="/wiki/HTML" title="HTML">
 HTML
</a>
and
<a href="/wiki/XHTML" title="XHTML">
 XHTML
</a>
), and frequently contain a wealth of useful data in text form.  However, most web pages
<a class="mw-redirect" href="/wiki/End-user_(computer_science)" title="End-user (compute
 end-users
</a>
and not for ease of automated use. Because of this, tool kits that scrape web content we
<a class="mw-redirect" href="/wiki/Application_Programming_Interface" title="Application
 Application Programming Interface
</a>
(API) to extract data from a web site. Companies like
<a class="mw-redirect" href="/wiki/Amazon_AWS" title="Amazon AWS">
 Amazon AWS
</a>
and
<a href="/wiki/Google" title="Google">
 Google
</a>
provide web scraping tools, services and public data available free of cost to end users
</p>
<p>
Newer forms of web scraping involve listening to data feeds from web servers.  For examp
<a href="/wiki/JSON" title="JSON">
 JSON
</a>
is commonly used as a transport storage mechanism between the client and the web server
</p>
<p>
There are methods that some websites use to prevent web scraping, such as detecting and
<a href="/wiki/Document_Object_Model" title="Document Object Model">
```

```
 DOM
</a>
parsing,
<a href="/wiki/Computer_vision" title="Computer vision">
 computer vision
</a>
and
<a href="/wiki/Natural_language_processing" title="Natural language processing">
 natural language processing
</a>
to simulate human browsing to enable gathering web page content for offline parsing.
</p>
<div class="toc" id="toc">
<input class="toctogglecheckbox" id="toctogglecheckbox" role="button" style="display:nor
<div class="toctitle" dir="ltr" lang="en">
 <h2>
  Contents
 </h2>
 <span class="toctogglespan">
  <label class="toctogglelabel" for="toctogglecheckbox">
  </label>
 </span>
</div>
<ul>
 <li class="toclevel-1 tocsection-1">
  <a href="#History">
   <span class="tocnumber">
    1
   </span>
   <span class="toctext">
    History
   </span>
  </a>
 </li>
 <li class="toclevel-1 tocsection-2">
  <a href="#Techniques">
   <span class="tocnumber">
    2
   </span>
   <span class="toctext">
    Techniques
   </span>
  </a>
  <ul>
   <li class="toclevel-2 tocsection-3">
    <a href="#Human_copy-and-paste">
     <span class="tocnumber">
      2.1
```

```
  </span>
  <span class="toctext">
   Human copy-and-paste
  </span>
 </a>
</li>
<li class="toclevel-2 tocsection-4">
 <a href="#Text_pattern_matching">
  <span class="tocnumber">
   2.2
  </span>
  <span class="toctext">
   Text pattern matching
  </span>
 </a>
</li>
<li class="toclevel-2 tocsection-5">
 <a href="#HTTP_programming">
  <span class="tocnumber">
   2.3
  </span>
  <span class="toctext">
   HTTP programming
  </span>
 </a>
</li>
<li class="toclevel-2 tocsection-6">
 <a href="#HTML_parsing">
  <span class="tocnumber">
   2.4
  </span>
  <span class="toctext">
   HTML parsing
  </span>
 </a>
</li>
<li class="toclevel-2 tocsection-7">
 <a href="#DOM_parsing">
  <span class="tocnumber">
   2.5
  </span>
  <span class="toctext">
   DOM parsing
  </span>
 </a>
</li>
<li class="toclevel-2 tocsection-8">
 <a href="#Vertical_aggregation">
```

```
    <span class="tocnumber">
     2.6
    </span>
    <span class="toctext">
     Vertical aggregation
    </span>
   </a>
  </li>
  <li class="toclevel-2 tocsection-9">
   <a href="#Semantic_annotation_recognizing">
    <span class="tocnumber">
     2.7
    </span>
    <span class="toctext">
     Semantic annotation recognizing
    </span>
   </a>
  </li>
  <li class="toclevel-2 tocsection-10">
   <a href="#Computer_vision_web-page_analysis">
    <span class="tocnumber">
     2.8
    </span>
    <span class="toctext">
     Computer vision web-page analysis
    </span>
   </a>
  </li>
 </ul>
</li>
<li class="toclevel-1 tocsection-11">
 <a href="#Software">
  <span class="tocnumber">
   3
  </span>
  <span class="toctext">
   Software
  </span>
 </a>
 <ul>
  <li class="toclevel-2 tocsection-12">
   <a href="#Example_tools">
    <span class="tocnumber">
     3.1
    </span>
    <span class="toctext">
     Example tools
    </span>
```

```
   </a>
   <ul>
    <li class="toclevel-3 tocsection-13">
     <a href="#Javascript_tools">
      <span class="tocnumber">
       3.1.1
      </span>
      <span class="toctext">
       Javascript tools
      </span>
     </a>
    </li>
    <li class="toclevel-3 tocsection-14">
     <a href="#Web_crawling_frameworks">
      <span class="tocnumber">
       3.1.2
      </span>
      <span class="toctext">
       Web crawling frameworks
      </span>
     </a>
    </li>
   </ul>
  </li>
 </ul>
</li>
<li class="toclevel-1 tocsection-15">
 <a href="#Legal_issues">
  <span class="tocnumber">
   4
  </span>
  <span class="toctext">
   Legal issues
  </span>
 </a>
 <ul>
  <li class="toclevel-2 tocsection-16">
   <a href="#United_States">
    <span class="tocnumber">
     4.1
    </span>
    <span class="toctext">
     United States
    </span>
   </a>
  </li>
  <li class="toclevel-2 tocsection-17">
   <a href="#The_EU">
```

```
    <span class="tocnumber">
     4.2
    </span>
    <span class="toctext">
     The EU
    </span>
   </a>
  </li>
  <li class="toclevel-2 tocsection-18">
   <a href="#Australia">
    <span class="tocnumber">
     4.3
    </span>
    <span class="toctext">
     Australia
    </span>
   </a>
  </li>
 </ul>
</li>
<li class="toclevel-1 tocsection-19">
 <a href="#Methods_to_prevent_web_scraping">
  <span class="tocnumber">
   5
  </span>
  <span class="toctext">
   Methods to prevent web scraping
  </span>
 </a>
</li>
<li class="toclevel-1 tocsection-20">
 <a href="#See_also">
  <span class="tocnumber">
   6
  </span>
  <span class="toctext">
   See also
  </span>
 </a>
</li>
<li class="toclevel-1 tocsection-21">
 <a href="#References">
  <span class="tocnumber">
   7
  </span>
  <span class="toctext">
   References
  </span>
```

```
   </a>
  </li>
 </ul>
</div>
<h2>
 <span class="mw-headline" id="History">
  History
 </span>
 <span class="mw-editsection">
  <span class="mw-editsection-bracket">
   [
  </span>
  <a href="/w/index.php?title=Web_scraping&amp;action=edit&amp;section=1" title="Edit se
   edit
  </a>
  <span class="mw-editsection-bracket">
   ]
  </span>
 </span>
</h2>
<table class="box-Unreferenced_section plainlinks metadata ambox ambox-content ambox-Unr
 <tbody>
  <tr>
   <td class="mbox-image">
    <div style="width:52px">
     <a class="image" href="/wiki/File:Question_book-new.svg">
      <img alt="" data-file-height="399" data-file-width="512" decoding="async" height="
     </a>
    </div>
   </td>
   <td class="mbox-text">
    <div class="mbox-text-span">
     This section
     <b>
      does not
      <a href="/wiki/Wikipedia:Citing_sources" title="Wikipedia:Citing sources">
       cite
      </a>
      any
      <a href="/wiki/Wikipedia:Verifiability" title="Wikipedia:Verifiability">
       sources
      </a>
     </b>
     .
     <span class="hide-when-compact">
      Please help
      <a class="external text" href="//en.wikipedia.org/w/index.php?title=Web_scraping&a
       improve this section
```

```
      </a>
      by
      <a href="/wiki/Help:Introduction_to_referencing_with_Wiki_Markup/1" title="Help:Int
       adding citations to reliable sources
      </a>
      . Unsourced material may be challenged and
      <a href="/wiki/Wikipedia:Verifiability#Burden_of_evidence" title="Wikipedia:Verifia
       removed
      </a>
      .
     </span>
     <small class="date-container">
      <i>
       (
       <span class="date">
        October 2018
       </span>
       )
      </i>
     </small>
     <small class="hide-when-compact">
      <i>
       (
       <a href="/wiki/Help:Maintenance_template_removal" title="Help:Maintenance templat
        Learn how and when to remove this template message
       </a>
       )
      </i>
     </small>
    </div>
   </td>
  </tr>
 </tbody>
</table>
<p>
 The history of the web scraping is actually much longer, dating back significantly to t
</p>
<ul>
 <li>
  After the birth of
  <a href="/wiki/History_of_the_World_Wide_Web" title="History of the World Wide Web">
   <b>
    World Wide Web
   </b>
  </a>
  in 1989, the First web robot -
  <b>
   <a href="/wiki/World_Wide_Web_Wanderer" title="World Wide Web Wanderer">
```

```
   World Wide Web Wanderer
  </a>
 </b>
 was created in 1993,June, which was intended only to measure the size of the web.
</li>
<li>
 In 1993,December, the First
 <b>
  crawler-based web search engine
 </b>
 -
 <a href="/wiki/JumpStation" title="JumpStation">
  JumpStation
 </a>
 . As there were not so many websites available on the web, search engines at that time
</li>
<li>
 In 2000, the
 <b>
  first Web API and API crawler
 </b>
 came.
 <a href="/wiki/Application_programming_interface" title="Application programming inter
  API
 </a>
 stands for
 <b>
  Application Programming Interface
 </b>
 . It is an interface that makes it much easier to develop a program by providing the bu
 <a href="/wiki/Salesforce.com" title="Salesforce.com">
  Salesforce
 </a>
 and
 <a href="/wiki/EBay" title="EBay">
  eBay
 </a>
 launched their own API, with which programmers were enabled to access and download some
</li>
<li>
 In 2004,
 <a href="/wiki/Beautiful_Soup_(HTML_parser)" title="Beautiful Soup (HTML parser)">
  <b>
   Beautiful Soup
  </b>
 </a>
 was released. It is a library designed for Python. As not all websites offer APIs, prog
</li>
```

```
</ul>
<h2>
 <span class="mw-headline" id="Techniques">
  Techniques
 </span>
 <span class="mw-editsection">
  <span class="mw-editsection-bracket">
   [
  </span>
  <a href="/w/index.php?title=Web_scraping&amp;action=edit&amp;section=2" title="Edit se
   edit
  </a>
  <span class="mw-editsection-bracket">
   ]
  </span>
 </span>
</h2>
<p>
 Web scraping is the process of automatically mining data or collecting information from
 <a class="mw-redirect" href="/wiki/Semantic_web" title="Semantic web">
  semantic web
 </a>
 vision, an ambitious initiative that still requires breakthroughs in text processing, s
 <a class="mw-redirect" href="/wiki/Human-computer_interaction" title="Human-computer in
  human-computer interactions
 </a>
 . Current web scraping solutions range from the ad-hoc, requiring human effort, to fully
</p>
<h3>
 <span class="mw-headline" id="Human_copy-and-paste">
  Human copy-and-paste
 </span>
 <span class="mw-editsection">
  <span class="mw-editsection-bracket">
   [
  </span>
  <a href="/w/index.php?title=Web_scraping&amp;action=edit&amp;section=3" title="Edit se
   edit
  </a>
  <span class="mw-editsection-bracket">
   ]
  </span>
 </span>
</h3>
<p>
 Sometimes even the best web-scraping technology cannot replace a humans manual examinat
</p>
<h3>
```

```
 <span class="mw-headline" id="Text_pattern_matching">
  Text pattern matching
 </span>
 <span class="mw-editsection">
  <span class="mw-editsection-bracket">
   [
  </span>
  <a href="/w/index.php?title=Web_scraping&amp;action=edit&amp;section=4" title="Edit se
   edit
  </a>
  <span class="mw-editsection-bracket">
   ]
  </span>
 </span>
</h3>
<p>
 A simple yet powerful approach to extract information from web pages can be based on th
 <a href="/wiki/Grep" title="Grep">
  grep
 </a>
 command or
 <a href="/wiki/Regular_expression" title="Regular expression">
  regular expression
 </a>
 -matching facilities of programming languages (for instance
 <a href="/wiki/Perl" title="Perl">
  Perl
 </a>
 or
 <a href="/wiki/Python_(programming_language)" title="Python (programming language)">
  Python
 </a>
 ).
</p>
<h3>
 <span class="mw-headline" id="HTTP_programming">
  HTTP programming
 </span>
 <span class="mw-editsection">
  <span class="mw-editsection-bracket">
   [
  </span>
  <a href="/w/index.php?title=Web_scraping&amp;action=edit&amp;section=5" title="Edit se
   edit
  </a>
  <span class="mw-editsection-bracket">
   ]
  </span>
```

```
 </span>
</h3>
<p>
 <a href="/wiki/Static_web_page" title="Static web page">
  Static
 </a>
 and
 <a href="/wiki/Dynamic_web_page" title="Dynamic web page">
  dynamic web pages
 </a>
 can be retrieved by posting HTTP requests to the remote web server using
 <a class="mw-redirect" href="/wiki/Socket_programming" title="Socket programming">
  socket programming
 </a>
 .
</p>
<h3>
 <span class="mw-headline" id="HTML_parsing">
  HTML parsing
 </span>
 <span class="mw-editsection">
  <span class="mw-editsection-bracket">
   [
  </span>
  <a href="/w/index.php?title=Web_scraping&amp;action=edit&amp;section=6" title="Edit se
   edit
  </a>
  <span class="mw-editsection-bracket">
   ]
  </span>
 </span>
</h3>
<p>
 Many websites have large collections of pages generated dynamically from an underlying
 <a href="/wiki/Wrapper_(data_mining)" title="Wrapper (data mining)">
  wrapper
 </a>
 . Wrapper generation algorithms assume that input pages of a wrapper induction system c
 <sup class="reference" id="cite_ref-3">
  <a href="#cite_note-3">
   [3]
  </a>
 </sup>
 Moreover, some
 <a href="/wiki/Semi-structured_data" title="Semi-structured data">
  semi-structured data
 </a>
 query languages, such as
```

```
 <a href="/wiki/XQuery" title="XQuery">
  XQuery
 </a>
 and the HTQL, can be used to parse HTML pages and to retrieve and transform page content
</p>
<h3>
 <span class="mw-headline" id="DOM_parsing">
  DOM parsing
 </span>
 <span class="mw-editsection">
  <span class="mw-editsection-bracket">
   [
  </span>
  <a href="/w/index.php?title=Web_scraping&amp;action=edit&amp;section=7" title="Edit sec
   edit
  </a>
  <span class="mw-editsection-bracket">
   ]
  </span>
 </span>
</h3>
<div class="hatnote navigation-not-searchable" role="note">
 Further information:
 <a href="/wiki/Document_Object_Model" title="Document Object Model">
  Document Object Model
 </a>
</div>
<p>
 By embedding a full-fledged web browser, such as the
 <a href="/wiki/Internet_Explorer" title="Internet Explorer">
  Internet Explorer
 </a>
 or the
 <a href="/wiki/Mozilla" title="Mozilla">
  Mozilla
 </a>
 browser control, programs can retrieve the dynamic content generated by client-side scri
</p>
<h3>
 <span class="mw-headline" id="Vertical_aggregation">
  Vertical aggregation
 </span>
 <span class="mw-editsection">
  <span class="mw-editsection-bracket">
   [
  </span>
  <a href="/w/index.php?title=Web_scraping&amp;action=edit&amp;section=8" title="Edit sec
   edit
```

```
  </a>
  <span class="mw-editsection-bracket">
   ]
  </span>
 </span>
</h3>
<p>
 There are several companies that have developed vertical specific harvesting platforms.
 <a class="mw-redirect" href="/wiki/Long_Tail" title="Long Tail">
  Long Tail
 </a>
 of sites that common aggregators find complicated or too labor-intensive to harvest cont
</p>
<h3>
 <span class="mw-headline" id="Semantic_annotation_recognizing">
  Semantic annotation recognizing
 </span>
 <span class="mw-editsection">
  <span class="mw-editsection-bracket">
   [
  </span>
  <a href="/w/index.php?title=Web_scraping&amp;action=edit&amp;section=9" title="Edit sec
   edit
  </a>
  <span class="mw-editsection-bracket">
   ]
  </span>
 </span>
</h3>
<p>
 The pages being scraped may embrace
 <a href="/wiki/Metadata" title="Metadata">
  metadata
 </a>
 or semantic markups and annotations, which can be used to locate specific data snippets
 <a href="/wiki/Microformat" title="Microformat">
  Microformat
 </a>
 does, this technique can be viewed as a special case of DOM parsing. In another case, th
 <sup class="reference" id="cite_ref-4">
  <a href="#cite_note-4">
   [4]
  </a>
 </sup>
 are stored and managed separately from the web pages, so the scrapers can retrieve data
</p>
<h3>
 <span class="mw-headline" id="Computer_vision_web-page_analysis">
```

```
  Computer vision web-page analysis
 </span>
 <span class="mw-editsection">
  <span class="mw-editsection-bracket">
   [
  </span>
  <a href="/w/index.php?title=Web_scraping&amp;action=edit&amp;section=10" title="Edit se
   edit
  </a>
  <span class="mw-editsection-bracket">
   ]
  </span>
 </span>
</h3>
<p>
 There are efforts using
 <a href="/wiki/Machine_learning" title="Machine learning">
  machine learning
 </a>
 and
 <a href="/wiki/Computer_vision" title="Computer vision">
  computer vision
 </a>
 that attempt to identify and extract information from web pages by interpreting pages vi
 <sup class="reference" id="cite_ref-5">
  <a href="#cite_note-5">
   [5]
  </a>
 </sup>
</p>
<h2>
 <span class="mw-headline" id="Software">
  Software
 </span>
 <span class="mw-editsection">
  <span class="mw-editsection-bracket">
   [
  </span>
  <a href="/w/index.php?title=Web_scraping&amp;action=edit&amp;section=11" title="Edit se
   edit
  </a>
  <span class="mw-editsection-bracket">
   ]
  </span>
 </span>
</h2>
<p>
 There are many software tools available that can be used to customize web-scraping solut
```

```
</p>
<h3>
 <span class="mw-headline" id="Example_tools">
  Example tools
 </span>
 <span class="mw-editsection">
  <span class="mw-editsection-bracket">
   [
  </span>
  <a href="/w/index.php?title=Web_scraping&amp;action=edit&amp;section=12" title="Edit se
   edit
  </a>
  <span class="mw-editsection-bracket">
   ]
  </span>
 </span>
</h3>
<ul>
 <li>
  <a href="/wiki/CURL" title="CURL">
   cURL
  </a>
   command line tool and library for transferring (including getting) data with URLs supp
 </li>
 <li>
  <a href="/wiki/Data_Toolbar" title="Data Toolbar">
   Data Toolbar
  </a>
    web scraping add-on for Internet Explorer, Mozilla Firefox, and Google Chrome Web bro
 </li>
 <li>
  <a href="/wiki/Diffbot" title="Diffbot">
   Diffbot
  </a>
    uses computer vision and machine learning to automatically extract data from web page
 </li>
 <li>
  <a href="/wiki/Heritrix" title="Heritrix">
   Heritrix
  </a>
   gets pages (lots of them). It is a web crawler designed for web archiving, written by
  <a href="/wiki/Wayback_Machine" title="Wayback Machine">
   Wayback Machine
  </a>
  ).
 </li>
 <li>
  <a href="/wiki/HtmlUnit" title="HtmlUnit">
```

```
  HtmlUnit
 </a>
  headless browser that can be used for retrieving web pages, web scraping, and more.
</li>
<li>
 <a href="/wiki/HTTrack" title="HTTrack">
  HTTrack
 </a>
  free and open source Web crawler and offline browser, designed to download websites.
</li>
<li>
 <a href="/wiki/IMacros" title="IMacros">
  iMacros
 </a>
  a browser extension to record, code, share and replay browser automation (javascript)
</li>
<li>
 <a href="/wiki/Selenium_(software)" title="Selenium (software)">
  Selenium (software)
 </a>
  a portable software-testing framework for web applications
</li>
<li>
 <a href="/wiki/Aptana#Aptana_Jaxer" title="Aptana">
  Jaxer
 </a>
</li>
<li>
 <a href="/wiki/Mozenda" title="Mozenda">
  Mozenda
 </a>
   is a WYSIWYG software that offers cloud, onsite, and data wrangling services.
</li>
<li>
 <a href="/wiki/Nokogiri_(software)" title="Nokogiri (software)">
  nokogiri
 </a>
</li>
<li>
 <a href="/wiki/OutWit_Hub" title="OutWit Hub">
  OutWit Hub
 </a>
  Web scraping application including built-in data, image, document extractors and edito
</li>
<li>
 <a href="/wiki/Watir" title="Watir">
  watir
 </a>
```

```
 </li>
 <li>
  <a href="/wiki/Wget" title="Wget">
   Wget
  </a>
   computer program that retrieves content from web servers. It is part of the GNU Proje
 </li>
 <li>
  <a href="/wiki/WSO2_Mashup_Server" title="WSO2 Mashup Server">
   WSO2 Mashup Server
  </a>

 </li>
 <li>
  <a href="/wiki/Yahoo!_Query_Language" title="Yahoo! Query Language">
   Yahoo! Query Language
  </a>
  (YQL)
 </li>
</ul>
<h4>
 <span class="mw-headline" id="Javascript_tools">
  Javascript tools
 </span>
 <span class="mw-editsection">
  <span class="mw-editsection-bracket">
   [
  </span>
  <a href="/w/index.php?title=Web_scraping&amp;action=edit&amp;section=13" title="Edit s
   edit
  </a>
  <span class="mw-editsection-bracket">
   ]
  </span>
 </span>
</h4>
<ul>
 <li>
  <a href="/wiki/Greasemonkey" title="Greasemonkey">
   Greasemonkey
  </a>
 </li>
 <li>
  <a href="/wiki/Node.js" title="Node.js">
   Node.js
  </a>
 </li>
```

```
 <li>
  <a href="/wiki/PhantomJS" title="PhantomJS">
   PhantomJS
  </a>
   scripted,
  <a href="/wiki/Headless_browser" title="Headless browser">
   headless browser
  </a>
  used for automating web page interaction.
 </li>
 <li>
  <a href="/wiki/JQuery" title="JQuery">
   jQuery
  </a>
 </li>
</ul>
<h4>
 <span class="mw-headline" id="Web_crawling_frameworks">
  Web crawling frameworks
 </span>
 <span class="mw-editsection">
  <span class="mw-editsection-bracket">
   [
  </span>
  <a href="/w/index.php?title=Web_scraping&amp;action=edit&amp;section=14" title="Edit se
   edit
  </a>
  <span class="mw-editsection-bracket">
   ]
  </span>
 </span>
</h4>
<p>
 These can be used to build web scrapers.
</p>
<ul>
 <li>
  <a href="/wiki/Scrapy" title="Scrapy">
   Scrapy
  </a>
 </li>
</ul>
<h2>
 <span class="mw-headline" id="Legal_issues">
  Legal issues
 </span>
 <span class="mw-editsection">
  <span class="mw-editsection-bracket">
```

```
   [
  </span>
  <a href="/w/index.php?title=Web_scraping&amp;action=edit&amp;section=15" title="Edit se
   edit
  </a>
  <span class="mw-editsection-bracket">
   ]
  </span>
 </span>
</h2>
<table class="box-Globalize plainlinks metadata ambox ambox-content ambox-globalize" role
 <tbody>
  <tr>
   <td class="mbox-image">
    <div style="width:52px">
     <img alt="Globe icon." data-file-height="290" data-file-width="350" decoding="async"
    </div>
   </td>
   <td class="mbox-text">
    <div class="mbox-text-span">
     The examples and perspective in this section
     <b>
      deal primarily with the United States and do not represent a
      <a href="/wiki/Wikipedia:WikiProject_Countering_systemic_bias" title="Wikipedia:Wi
       worldwide view
      </a>
      of the subject
     </b>
     .
     <span class="hide-when-compact">
      You may
      <a class="external text" href="//en.wikipedia.org/w/index.php?title=Web_scraping&am
       improve this section
      </a>
      , discuss the issue on the
      <a href="/wiki/Talk:Web_scraping" title="Talk:Web scraping">
       talk page
      </a>
      , or
      <a href="/wiki/Wikipedia:Article_wizard" title="Wikipedia:Article wizard">
       create a new article
      </a>
      , as appropriate.
     </span>
     <small class="date-container">
      <i>
       (
       <span class="date">
```

```
        October 2015
       </span>
       )
      </i>
     </small>
     <small class="hide-when-compact">
      <i>
       (
       <a href="/wiki/Help:Maintenance_template_removal" title="Help:Maintenance templat
        Learn how and when to remove this template message
       </a>
       )
      </i>
     </small>
    </div>
   </td>
  </tr>
 </tbody>
</table>
<p>
 The legality of web scraping varies across the world. In general, web scraping may be ag
 <a class="mw-redirect" href="/wiki/Terms_of_use" title="Terms of use">
  terms of use
 </a>
 of some websites, but the enforceability of these terms is unclear.
 <sup class="reference" id="cite_ref-6">
  <a href="#cite_note-6">
   [6]
  </a>
 </sup>
</p>
<h3>
 <span class="mw-headline" id="United_States">
  United States
 </span>
 <span class="mw-editsection">
  <span class="mw-editsection-bracket">
   [
  </span>
  <a href="/w/index.php?title=Web_scraping&amp;action=edit&amp;section=16" title="Edit se
   edit
  </a>
  <span class="mw-editsection-bracket">
   ]
  </span>
 </span>
</h3>
<table class="plainlinks metadata ambox ambox-move" role="presentation">
```

```
 <tbody>
  <tr>
   <td class="mbox-image">
    <div style="width:52px">
     <img alt="" data-file-height="100" data-file-width="300" decoding="async" height="1
    </div>
   </td>
   <td class="mbox-text">
    <div class="mbox-text-span">
     It has been suggested that this section be
     <a href="/wiki/Wikipedia:Splitting" title="Wikipedia:Splitting">
      split
     </a>
     out  into another article titled
     <i>
      <a class="new" href="/w/index.php?title=Web_scraping_in_the_United_States&amp;actio
       Web scraping in the United States
      </a>
     </i>
     . (
     <a href="/wiki/Talk:Web_scraping#Split_section" title="Talk:Web scraping">
      Discuss
     </a>
     )
     <small>
      <i>
       (July 2018)
      </i>
     </small>
    </div>
   </td>
  </tr>
 </tbody>
</table>
<p>
 In the United States, website owners can use three major
 <a href="/wiki/Cause_of_action" title="Cause of action">
  legal claims
 </a>
 to prevent undesired web scraping: (1) copyright infringement (compilation), (2) violati
 <a href="/wiki/Computer_Fraud_and_Abuse_Act" title="Computer Fraud and Abuse Act">
  Computer Fraud and Abuse Act
 </a>
 (CFAA), and (3)
 <a href="/wiki/Trespass_to_chattels" title="Trespass to chattels">
  trespass to chattel
 </a>
 .
```

```
 <sup class="reference" id="cite_ref-7">
  <a href="#cite_note-7">
   [7]
  </a>
 </sup>
 However, the effectiveness of these claims relies upon meeting various criteria, and the
 <a href="/wiki/Feist_Publications,_Inc.,_v._Rural_Telephone_Service_Co." title="Feist Pu
  <i>
   Feist Publications v. Rural Telephone Service
  </i>
 </a>
 that duplication of facts is allowable.
</p>
<p>
 U.S. courts have acknowledged that users of "scrapers" or "robots" may be held liable fo
 <a href="/wiki/Trespass_to_chattels" title="Trespass to chattels">
  trespass to chattels
 </a>
 ,
 <sup class="reference" id="cite_ref-8">
  <a href="#cite_note-8">
   [8]
  </a>
 </sup>
 <sup class="reference" id="cite_ref-9">
  <a href="#cite_note-9">
   [9]
  </a>
 </sup>
 which involves a computer system itself being considered personal property upon which th
 <i>
  <a href="/wiki/EBay_v._Bidder%27s_Edge" title="EBay v. Bidder's Edge">
   eBay v. Bidder's Edge
  </a>
 </i>
 , resulted in an injunction ordering Bidder's Edge to stop accessing, collecting, and in
 <a href="/wiki/Auction_sniping" title="Auction sniping">
  auction sniping
 </a>
 . However, in order to succeed on a claim of trespass to
 <a href="/wiki/Personal_property" title="Personal property">
  chattels
 </a>
 , the
 <a href="/wiki/Plaintiff" title="Plaintiff">
  plaintiff
 </a>
 must demonstrate that the
```

```
 <a href="/wiki/Defendant" title="Defendant">
  defendant
 </a>
 intentionally and without authorization interfered with the plaintiff's possessory inter
 <sup class="reference" id="cite_ref-10">
  <a href="#cite_note-10">
   [10]
  </a>
 </sup>
</p>
<p>
 One of the first major tests of
 <a class="mw-redirect" href="/wiki/Screen_scraping" title="Screen scraping">
  screen scraping
 </a>
 involved
 <a href="/wiki/American_Airlines" title="American Airlines">
  American Airlines
 </a>
 (AA), and a firm called FareChase.
 <sup class="reference" id="cite_ref-11">
  <a href="#cite_note-11">
   [11]
  </a>
 </sup>
 AA successfully obtained an
 <a href="/wiki/Injunction" title="Injunction">
  injunction
 </a>
 from a Texas trial court, stopping FareChase from selling software that enables users to
 <sup class="reference" id="cite_ref-12">
  <a href="#cite_note-12">
   [12]
  </a>
 </sup>
</p>
<p>
 <a href="/wiki/Southwest_Airlines" title="Southwest Airlines">
  Southwest Airlines
 </a>
 has also challenged screen-scraping practices, and has involved both FareChase and anoth
 <a class="mw-redirect" href="/wiki/US_Copyright_law" title="US Copyright law">
  US Copyright law
 </a>
 , and that under copyright, the pieces of information being scraped would not be subject
 <a href="/wiki/Supreme_Court_of_the_United_States" title="Supreme Court of the United St
  Supreme Court of the United States
 </a>
```

```
 , FareChase was eventually shuttered by parent company
 <a href="/wiki/Yahoo!" title="Yahoo!">
  Yahoo!
 </a>
 , and Outtask was purchased by travel expense company Concur.
 <sup class="reference" id="cite_ref-impervawp2011_13-0">
  <a href="#cite_note-impervawp2011-13">
   [13]
  </a>
 </sup>
 In 2012, a startup called 3Taps scraped classified housing ads from Craigslist. Craigsli
 <i>
  <a class="mw-redirect" href="/wiki/Craigslist_v._3Taps" title="Craigslist v. 3Taps">
   Craigslist v. 3Taps
  </a>
 </i>
 . The court held that the cease-and-desist letter and IP blocking was sufficient for Cra
 <a href="/wiki/Computer_Fraud_and_Abuse_Act" title="Computer Fraud and Abuse Act">
  Computer Fraud and Abuse Act
 </a>
 .
</p>
<p>
 Although these are early scraping decisions, and the theories of liability are not unifo
 <sup class="reference" id="cite_ref-14">
  <a href="#cite_note-14">
   [14]
  </a>
 </sup>
</p>
<p>
 While the law in this area becomes more settled, entities contemplating using scraping p
 <i>
  <a class="mw-redirect" href="/wiki/Cvent,_Inc." title="Cvent, Inc.">
   Cvent, Inc.
  </a>
  v.
  <a href="/wiki/Eventbrite" title="Eventbrite">
   Eventbrite, Inc.
  </a>
 </i>
 In the United States district court for the eastern district of Virginia, the court rule
 <a href="/wiki/Browse_wrap" title="Browse wrap">
  browse wrap
 </a>
 contract or license to be enforced.
 <sup class="reference" id="cite_ref-15">
  <a href="#cite_note-15">
```

```
  [15]
 </a>
</sup>
In a 2014, filed in the
<a href="/wiki/United_States_District_Court_for_the_Eastern_District_of_Pennsylvania" ti
 United States District Court for the Eastern District of Pennsylvania
</a>
,
<sup class="reference" id="cite_ref-16">
 <a href="#cite_note-16">
  [16]
 </a>
</sup>
e-commerce site
<a href="/wiki/QVC" title="QVC">
 QVC
</a>
objected to the Pinterest-like shopping aggregator Resultlys `scraping of QVCs site for
<sup class="reference" id="cite_ref-17">
 <a href="#cite_note-17">
  [17]
 </a>
</sup>
QVC's complaint alleges that the defendant disguised its web crawler to mask its source
</p>
<p>
In the plaintiff's web site during the period of this trial the terms of use link is dis
<sup class="reference" id="cite_ref-18">
 <a href="#cite_note-18">
  [18]
 </a>
</sup>
</p>
<p>
In
<i>
 <a href="/wiki/Facebook,_Inc._v._Power_Ventures,_Inc." title="Facebook, Inc. v. Power V
  Facebook, Inc. v. Power Ventures, Inc.
 </a>
</i>
, a district court ruled in 2012 that Power Ventures could not scrape Facebook pages on
<a href="/wiki/Electronic_Frontier_Foundation" title="Electronic Frontier Foundation">
 Electronic Frontier Foundation
</a>
filed a brief in 2015 asking that it be overturned.
<sup class="reference" id="cite_ref-19">
 <a href="#cite_note-19">
  [19]
```

```
  </a>
 </sup>
 <sup class="reference" id="cite_ref-20">
  <a href="#cite_note-20">
   [20]
  </a>
 </sup>
 In
 <i>
  <a href="/wiki/Associated_Press_v._Meltwater_U.S._Holdings,_Inc." title="Associated Pre
   Associated Press v. Meltwater U.S. Holdings, Inc.
  </a>
 </i>
 , a court in the US held Meltwater liable for scraping and republishing news informatio
</p>
<p>
 Internet Archive collects and distributes significant number of publicly available webpa
</p>
<h3>
 <span class="mw-headline" id="The_EU">
  The EU
 </span>
 <span class="mw-editsection">
  <span class="mw-editsection-bracket">
   [
  </span>
  <a href="/w/index.php?title=Web_scraping&amp;action=edit&amp;section=17" title="Edit se
   edit
  </a>
  <span class="mw-editsection-bracket">
   ]
  </span>
 </span>
</h3>
<p>
 In February 2006, the Danish Maritime and Commercial Court (Copenhagen) ruled that syste
 <sup class="reference" id="cite_ref-21">
  <a href="#cite_note-21">
   [21]
  </a>
 </sup>
</p>
<p>
 In a February 2010 case complicated by matters of jurisdiction, Ireland's High Court del
 <a href="/wiki/Inchoate_offense" title="Inchoate offense">
  inchoate
 </a>
 state of developing case law. In the case of
```

```
 <i>
  Ryanair Ltd v Billigfluege.de GmbH
 </i>
 , Ireland's High Court ruled
 <a href="/wiki/Ryanair" title="Ryanair">
  Ryanair's
 </a>
 "click-wrap" agreement to be legally binding. In contrast to the findings of the United
 <sup class="reference" id="cite_ref-22">
  <a href="#cite_note-22">
   [22]
  </a>
 </sup>
 The decision is under appeal in Ireland's Supreme Court.
 <sup class="reference" id="cite_ref-23">
  <a href="#cite_note-23">
   [23]
  </a>
 </sup>
</p>
<h3>
 <span class="mw-headline" id="Australia">
  Australia
 </span>
 <span class="mw-editsection">
  <span class="mw-editsection-bracket">
   [
  </span>
  <a href="/w/index.php?title=Web_scraping&amp;action=edit&amp;section=18" title="Edit se
   edit
  </a>
  <span class="mw-editsection-bracket">
   ]
  </span>
 </span>
</h3>
<p>
 In Australia, the
 <a href="/wiki/Spam_Act_2003" title="Spam Act 2003">
  Spam Act 2003
 </a>
 outlaws some forms of web harvesting, although this only applies to email addresses.
 <sup class="reference" id="cite_ref-24">
  <a href="#cite_note-24">
   [24]
  </a>
 </sup>
 <sup class="reference" id="cite_ref-25">
```

```
  <a href="#cite_note-25">
   [25]
  </a>
 </sup>
</p>
<h2>
 <span class="mw-headline" id="Methods_to_prevent_web_scraping">
  Methods to prevent web scraping
 </span>
 <span class="mw-editsection">
  <span class="mw-editsection-bracket">
   [
  </span>
  <a href="/w/index.php?title=Web_scraping&amp;action=edit&amp;section=19" title="Edit se
   edit
  </a>
  <span class="mw-editsection-bracket">
   ]
  </span>
 </span>
</h2>
<p>
 The administrator of a website can use various measures to stop or slow a bot. Some tech
</p>
<ul>
 <li>
  Blocking an
  <a href="/wiki/IP_address" title="IP address">
   IP address
  </a>
  either manually or based on criteria such as
  <a href="/wiki/Geolocation" title="Geolocation">
   geolocation
  </a>
  and
  <a href="/wiki/DNSBL" title="DNSBL">
   DNSRBL
  </a>
  . This will also block all browsing from that address.
 </li>
 <li>
  Disabling any
  <a href="/wiki/Web_service" title="Web service">
   web service
  </a>
  <a href="/wiki/Application_programming_interface" title="Application programming interf
   API
  </a>
```

```
 that the website's system might expose.
</li>
<li>
 Bots sometimes declare who they are (using
 <a href="/wiki/User_agent" title="User agent">
  user agent
 </a>
 <a href="/wiki/String_(computer_science)" title="String (computer science)">
  strings
 </a>
 ) and can be blocked on that basis using
 <a href="/wiki/Robots_exclusion_standard" title="Robots exclusion standard">
  robots.txt
 </a>
 ; '
 <a href="/wiki/Googlebot" title="Googlebot">
  googlebot
 </a>
 ' is an example. Other bots make no distinction between themselves and a human using a
</li>
<li>
 Bots can be blocked by monitoring excess traffic
</li>
<li>
 Bots can sometimes be blocked with tools to verify that it is a real person accessing t
 <a href="/wiki/CAPTCHA" title="CAPTCHA">
  CAPTCHA
 </a>
 . Bots are sometimes coded to explicitly break specific CAPTCHA patterns or may employ
</li>
<li>
 Commercial anti-bot services: Companies offer anti-bot and anti-scraping services for w
 <a href="/wiki/Application_firewall" title="Application firewall">
  application firewalls
 </a>
 have limited bot detection capabilities as well. However, many such solutions are not v
 <sup class="reference" id="cite_ref-26">
  <a href="#cite_note-26">
   [26]
  </a>
 </sup>
 .
</li>
</ul>
<ul>
 <li>
  Locating bots with a
  <a href="/wiki/Honeypot_(computing)" title="Honeypot (computing)">
```

```
   honeypot
  </a>
  or other method to identify the IP addresses of automated crawlers.
 </li>
 <li>
  <a href="/wiki/Obfuscation" title="Obfuscation">
   Obfuscation
  </a>
  using
  <a class="mw-redirect" href="/wiki/CSS_sprite" title="CSS sprite">
   CSS sprites
  </a>
  to display such data as phone numbers or email addresses, at the cost of
  <a href="/wiki/Web_accessibility" title="Web accessibility">
   accessibility
  </a>
  to
  <a href="/wiki/Screen_reader" title="Screen reader">
   screen reader
  </a>
  users.
 </li>
 <li>
  Because bots rely on consistency in the front-end code of a target website, adding smal
 </li>
 <li>
  Websites can declare if crawling is allowed or not in the
  <a href="/wiki/Robots_exclusion_standard" title="Robots exclusion standard">
   robots.txt
  </a>
  file and allow partial access, limit the crawl rate, specify the optimal time to crawl
 </li>
</ul>
<h2>
 <span class="mw-headline" id="See_also">
  See also
 </span>
 <span class="mw-editsection">
  <span class="mw-editsection-bracket">
   [
  </span>
  <a href="/w/index.php?title=Web_scraping&amp;action=edit&amp;section=20" title="Edit se
   edit
  </a>
  <span class="mw-editsection-bracket">
   ]
  </span>
 </span>
```

```
</h2>
<div class="div-col columns column-width" style="-moz-column-width: 22em; -webkit-column-
 <ul>
  <li>
   <a class="mw-redirect" href="/wiki/Archive.is" title="Archive.is">
    Archive.is
   </a>
  </li>
  <li>
   <a href="/wiki/Comparison_of_feed_aggregators" title="Comparison of feed aggregators">
    Comparison of feed aggregators
   </a>
  </li>
  <li>
   <a href="/wiki/Data_scraping" title="Data scraping">
    Data scraping
   </a>
  </li>
  <li>
   <a href="/wiki/Data_wrangling" title="Data wrangling">
    Data wrangling
   </a>
  </li>
  <li>
   <a href="/wiki/Importer_(computing)" title="Importer (computing)">
    Importer
   </a>
  </li>
  <li>
   <a href="/wiki/Job_wrapping" title="Job wrapping">
    Job wrapping
   </a>
  </li>
  <li>
   <a href="/wiki/Knowledge_extraction" title="Knowledge extraction">
    Knowledge extraction
   </a>
  </li>
  <li>
   <a href="/wiki/OpenSocial" title="OpenSocial">
    OpenSocial
   </a>
  </li>
  <li>
   <a href="/wiki/Scraper_site" title="Scraper site">
    Scraper site
   </a>
  </li>
```

```
  <li>
   <a href="/wiki/Fake_news_website" title="Fake news website">
    Fake news website
   </a>
  </li>
  <li>
   <a href="/wiki/Blog_scraping" title="Blog scraping">
    Blog scraping
   </a>
  </li>
  <li>
   <a href="/wiki/Spamdexing" title="Spamdexing">
    Spamdexing
   </a>
  </li>
  <li>
   <a href="/wiki/Domain_name_drop_list" title="Domain name drop list">
    Domain name drop list
   </a>
  </li>
  <li>
   <a href="/wiki/Text_corpus" title="Text corpus">
    Text corpus
   </a>
  </li>
  <li>
   <a href="/wiki/Web_archiving" title="Web archiving">
    Web archiving
   </a>
  </li>
  <li>
   <a class="mw-redirect" href="/wiki/Blog_network" title="Blog network">
    Blog network
   </a>
  </li>
  <li>
   <a class="mw-redirect" href="/wiki/Search_Engine_Scraping" title="Search Engine Scrapi
    Search Engine Scraping
   </a>
  </li>
  <li>
   <a href="/wiki/Category:Web_crawlers" title="Category:Web crawlers">
    Web crawlers
   </a>
  </li>
 </ul>
</div>
<h2>
```

```
 <span class="mw-headline" id="References">
  References
 </span>
 <span class="mw-editsection">
  <span class="mw-editsection-bracket">
   [
  </span>
  <a href="/w/index.php?title=Web_scraping&amp;action=edit&amp;section=21" title="Edit se
   edit
  </a>
  <span class="mw-editsection-bracket">
   ]
  </span>
 </span>
</h2>
<div class="reflist columns references-column-width" style="-moz-column-width: 20em; -web
 <ol class="references">
  <li id="cite_note-Boeing2016JPER-1">
   <span class="mw-cite-backlink">
    ^
    <a href="#cite_ref-Boeing2016JPER_1-0">
     <sup>
      <i>
       <b>
        a
       </b>
      </i>
     </sup>
    </a>
    <a href="#cite_ref-Boeing2016JPER_1-1">
     <sup>
      <i>
       <b>
        b
       </b>
      </i>
     </sup>
    </a>
   </span>
   <span class="reference-text">
    <cite class="citation journal">
     Boeing, G.; Waddell, P. (2016). "New Insights into Rental Housing Markets across the
     <i>
      Journal of Planning Education and Research
     </i>
     (0739456X16664789).
     <a href="/wiki/ArXiv" title="ArXiv">
      arXiv
```

```
   </a>
   :
   <span class="cs1-lock-free" title="Freely accessible">
    <a class="external text" href="//arxiv.org/abs/1605.05397" rel="nofollow">
     1605.05397
    </a>
   </span>
   .
   <a href="/wiki/Digital_object_identifier" title="Digital object identifier">
    doi
   </a>
   :
   <a class="external text" href="//doi.org/10.1177%2F0739456X16664789" rel="nofollow"
    10.1177/0739456X16664789
   </a>
   .
  </cite>
  <span class="Z3988" title="ctx_ver=Z39.88-2004&amp;rft_val_fmt=info%3Aofi%2Ffmt%3Ake
  </span>
  <style data-mw-deduplicate="TemplateStyles:r886058088">
   .mw-parser-output cite.citation{font-style:inherit}.mw-parser-output .citation q{quo
  </style>
 </span>
</li>
<li id="cite_note-2">
 <span class="mw-cite-backlink">
  <b>
   <a href="#cite_ref-2">
    ^
   </a>
  </b>
 </span>
 <span class="reference-text">
  <cite class="citation journal">
   Vargiu &amp;amp; Urru (2013). "Exploiting web scraping in a collaborative filtering- bas
   <i>
    Artificial Intelligence Research
   </i>
   .
   <b>
    2
   </b>
   (1).
   <a href="/wiki/Digital_object_identifier" title="Digital object identifier">
    doi
   </a>
   :
   <a class="external text" href="//doi.org/10.5430%2Fair.v2n1p44" rel="nofollow">
```

```
    10.5430/air.v2n1p44
   </a>
   .
  </cite>
  <span class="Z3988" title="ctx_ver=Z39.88-2004&amp;rft_val_fmt=info%3Aofi%2Ffmt%3Ake
  </span>
  <link href="mw-data:TemplateStyles:r886058088" rel="mw-deduplicated-inline-style"/>
 </span>
</li>
<li id="cite_note-3">
 <span class="mw-cite-backlink">
  <b>
   <a href="#cite_ref-3">
    ^
   </a>
  </b>
 </span>
 <span class="reference-text">
  <cite class="citation journal">
   Song, Ruihua; Microsoft Research (Sep 14, 2007).
   <a class="external text" href="https://pdfs.semanticscholar.org/4fb4/3c5a212df751e84
    "Joint Optimization of Wrapper Generation and Template Detection"
   </a>
   <span class="cs1-format">
    (PDF)
   </span>
   .
   <i>
    The 13th International Conference on Knowledge Discovery and Data Mining
   </i>
   .
  </cite>
  <span class="Z3988" title="ctx_ver=Z39.88-2004&amp;rft_val_fmt=info%3Aofi%2Ffmt%3Ake
  </span>
  <link href="mw-data:TemplateStyles:r886058088" rel="mw-deduplicated-inline-style"/>
 </span>
</li>
<li id="cite_note-4">
 <span class="mw-cite-backlink">
  <b>
   <a href="#cite_ref-4">
    ^
   </a>
  </b>
 </span>
 <span class="reference-text">
  <a class="external text" href="http://www.gooseeker.com/en/node/knowledgebase/freefo
   Semantic annotation based web scraping
```

```
  </a>
 </span>
</li>
<li id="cite_note-5">
 <span class="mw-cite-backlink">
  <b>
   <a href="#cite_ref-5">
    ^
   </a>
  </b>
 </span>
 <span class="reference-text">
  <cite class="citation web">
   Roush, Wade (2012-07-25).
   <a class="external text" href="http://www.xconomy.com/san-francisco/2012/07/25/diffl
    "Diffbot Is Using Computer Vision to Reinvent the Semantic Web"
   </a>
   . www.xconomy.com
   <span class="reference-accessdate">
    . Retrieved
    <span class="nowrap">
     2013-03-15
    </span>
   </span>
   .
  </cite>
  <span class="Z3988" title="ctx_ver=Z39.88-2004&amp;rft_val_fmt=info%3Aofi%2Ffmt%3Ake
  </span>
  <link href="mw-data:TemplateStyles:r886058088" rel="mw-deduplicated-inline-style"/>
 </span>
</li>
<li id="cite_note-6">
 <span class="mw-cite-backlink">
  <b>
   <a href="#cite_ref-6">
    ^
   </a>
  </b>
 </span>
 <span class="reference-text">
  <cite class="citation web">
   <a class="external text" href="https://web.archive.org/web/20020308222536/http://ww
    "FAQ about linkingă Are website terms of use binding contracts?"
   </a>
   . www.chillingeffects.org. 2007-08-20. Archived from
   <a class="external text" href="http://www.chillingeffects.org/linking/faq.cgi#QID596
    the original
   </a>
```

```
   on 2002-03-08
   <span class="reference-accessdate">
    . Retrieved
    <span class="nowrap">
     2007-08-20
    </span>
   </span>
   .
  </cite>
  <span class="Z3988" title="ctx_ver=Z39.88-2004&amp;rft_val_fmt=info%3Aofi%2Ffmt%3Ake
  </span>
  <link href="mw-data:TemplateStyles:r886058088" rel="mw-deduplicated-inline-style"/>
 </span>
</li>
<li id="cite_note-7">
 <span class="mw-cite-backlink">
  <b>
   <a href="#cite_ref-7">
    ^
   </a>
  </b>
 </span>
 <span class="reference-text">
  <cite class="citation journal">
   Kenneth, Hirschey, Jeffrey (2014-01-01).
   <a class="external text" href="http://scholarship.law.berkeley.edu/btlj/vol29/iss4/
    "Symbiotic Relationships: Pragmatic Acceptance of Data Scraping"
   </a>
   .
   <i>
    Berkeley Technology Law Journal
   </i>
   .
   <b>
    29
   </b>
   (4).
   <a href="/wiki/Digital_object_identifier" title="Digital object identifier">
    doi
   </a>
   :
   <a class="external text" href="//doi.org/10.15779%2FZ38B39B" rel="nofollow">
    10.15779/Z38B39B
   </a>
   .
   <a href="/wiki/International_Standard_Serial_Number" title="International Standard S
    ISSN
   </a>
```

```
   <a class="external text" href="//www.worldcat.org/issn/1086-3818" rel="nofollow">
    1086-3818
   </a>
   .
  </cite>
  <span class="Z3988" title="ctx_ver=Z39.88-2004&amp;rft_val_fmt=info%3Aofi%2Ffmt%3Ake
  </span>
  <link href="mw-data:TemplateStyles:r886058088" rel="mw-deduplicated-inline-style"/>
 </span>
</li>
<li id="cite_note-8">
 <span class="mw-cite-backlink">
  <b>
   <a href="#cite_ref-8">
    ^
   </a>
  </b>
 </span>
 <span class="reference-text">
  <cite class="citation web">
   <a class="external text" href="http://www.tomwbell.com/NetLaw/Ch06.html" rel="nofol
    "Internet Law, Ch. 06: Trespass to Chattels"
   </a>
   . www.tomwbell.com. 2007-08-20
   <span class="reference-accessdate">
    . Retrieved
    <span class="nowrap">
     2007-08-20
    </span>
   </span>
   .
  </cite>
  <span class="Z3988" title="ctx_ver=Z39.88-2004&amp;rft_val_fmt=info%3Aofi%2Ffmt%3Ake
  </span>
  <link href="mw-data:TemplateStyles:r886058088" rel="mw-deduplicated-inline-style"/>
 </span>
</li>
<li id="cite_note-9">
 <span class="mw-cite-backlink">
  <b>
   <a href="#cite_ref-9">
    ^
   </a>
  </b>
 </span>
 <span class="reference-text">
  <cite class="citation web">
   <a class="external text" href="https://web.archive.org/web/20020308222536/http://ww
```

```
    "What are the "trespass to chattels" claims some companies or website owners have l
   </a>
   . www.chillingeffects.org. 2007-08-20. Archived from
   <a class="external text" href="http://www.chillingeffects.org/linking/faq.cgi#QID460
    the original
   </a>
   on 2002-03-08
   <span class="reference-accessdate">
    . Retrieved
    <span class="nowrap">
     2007-08-20
    </span>
   </span>
   .
  </cite>
  <span class="Z3988" title="ctx_ver=Z39.88-2004&amp;rft_val_fmt=info%3Aofi%2Ffmt%3Akev
  </span>
  <link href="mw-data:TemplateStyles:r886058088" rel="mw-deduplicated-inline-style"/>
 </span>
</li>
<li id="cite_note-10">
 <span class="mw-cite-backlink">
  <b>
   <a href="#cite_ref-10">
    ^
   </a>
  </b>
 </span>
 <span class="reference-text">
  <cite class="citation web">
   <a class="external text" href="http://www.tomwbell.com/NetLaw/Ch07/Ticketmaster.html
    "Ticketmaster Corp. v. Tickets.com, Inc"
   </a>
   . 2007-08-20
   <span class="reference-accessdate">
    . Retrieved
    <span class="nowrap">
     2007-08-20
    </span>
   </span>
   .
  </cite>
  <span class="Z3988" title="ctx_ver=Z39.88-2004&amp;rft_val_fmt=info%3Aofi%2Ffmt%3Akev
  </span>
  <link href="mw-data:TemplateStyles:r886058088" rel="mw-deduplicated-inline-style"/>
 </span>
</li>
<li id="cite_note-11">
```

```
 <span class="mw-cite-backlink">
  <b>
   <a href="#cite_ref-11">
    ^
   </a>
  </b>
 </span>
 <span class="reference-text">
  <cite class="citation web">
   <a class="external text" href="https://web.archive.org/web/20110723131832/http://ww
    "American Airlines v. FareChase"
   </a>
   <span class="cs1-format">
    (PDF)
   </span>
   . 2007-08-20. Archived from
   <a class="external text" href="http://www.fornova.net/documents/AAFareChase.pdf" re
    the original
   </a>
   <span class="cs1-format">
    (PDF)
   </span>
   on 2011-07-23
   <span class="reference-accessdate">
    . Retrieved
    <span class="nowrap">
     2007-08-20
    </span>
   </span>
   .
  </cite>
  <span class="Z3988" title="ctx_ver=Z39.88-2004&amp;rft_val_fmt=info%3Aofi%2Ffmt%3Ake
  </span>
  <link href="mw-data:TemplateStyles:r886058088" rel="mw-deduplicated-inline-style"/>
 </span>
</li>
<li id="cite_note-12">
 <span class="mw-cite-backlink">
  <b>
   <a href="#cite_ref-12">
    ^
   </a>
  </b>
 </span>
 <span class="reference-text">
  <cite class="citation web">
   <a class="external text" href="http://www.thefreelibrary.com/American+Airlines,+Far
    "American Airlines, FareChase Settle Suit"
```

```
   </a>
   . The Free Library. 2003-06-13
   <span class="reference-accessdate">
    . Retrieved
    <span class="nowrap">
     2012-02-26
    </span>
   </span>
   .
  </cite>
  <span class="Z3988" title="ctx_ver=Z39.88-2004&amp;rft_val_fmt=info%3Aofi%2Ffmt%3Ake
  </span>
  <link href="mw-data:TemplateStyles:r886058088" rel="mw-deduplicated-inline-style"/>
 </span>
</li>
<li id="cite_note-impervawp2011-13">
 <span class="mw-cite-backlink">
  <b>
   <a href="#cite_ref-impervawp2011_13-0">
    ^
   </a>
  </b>
 </span>
 <span class="reference-text">
  Imperva (2011).
  <a class="external text" href="http://www.imperva.com/docs/WP_Detecting_and_Blocking_
   Detecting and Blocking Site Scraping Attacks
  </a>
  . Imperva white paper..
 </span>
</li>
<li id="cite_note-14">
 <span class="mw-cite-backlink">
  <b>
   <a href="#cite_ref-14">
    ^
   </a>
  </b>
 </span>
 <span class="reference-text">
  <cite class="citation web">
   Adler, Kenneth A. (2003-07-29).
   <a class="external text" href="http://library.findlaw.com/2003/Jul/29/132944.html" r
    "Controversy Surrounds 'Screen Scrapers': Software Helps Users Access Web Sites Bu
   </a>
   <span class="reference-accessdate">
    . Retrieved
    <span class="nowrap">
```

```
     2010-10-27
    </span>
   </span>
   .
  </cite>
  <span class="Z3988" title="ctx_ver=Z39.88-2004&amp;rft_val_fmt=info%3Aofi%2Ffmt%3Ake
  </span>
  <link href="mw-data:TemplateStyles:r886058088" rel="mw-deduplicated-inline-style"/>
 </span>
</li>
<li id="cite_note-15">
 <span class="mw-cite-backlink">
  <b>
   <a href="#cite_ref-15">
    ^
   </a>
  </b>
 </span>
 <span class="reference-text">
  <cite class="citation web">
   <a class="external text" href="http://www.fornova.net/documents/Cvent.pdf" rel="nofo
    "QVC Inc. v. Resultly LLC, No. 14-06714 (E.D. Pa. filed Nov. 24, 2014)"
   </a>
   <span class="cs1-format">
    (PDF)
   </span>
   . 2014-11-24
   <span class="reference-accessdate">
    . Retrieved
    <span class="nowrap">
     2015-11-05
    </span>
   </span>
   .
  </cite>
  <span class="Z3988" title="ctx_ver=Z39.88-2004&amp;rft_val_fmt=info%3Aofi%2Ffmt%3Ake
  </span>
  <link href="mw-data:TemplateStyles:r886058088" rel="mw-deduplicated-inline-style"/>
 </span>
</li>
<li id="cite_note-16">
 <span class="mw-cite-backlink">
  <b>
   <a href="#cite_ref-16">
    ^
   </a>
  </b>
 </span>
```

```
 <span class="reference-text">
  <cite class="citation web">
   <a class="external text" href="https://www.scribd.com/doc/249068700/LinkedIn-v-Resul
    "QVC Inc. v. Resultly LLC, No. 14-06714 (E.D. Pa. filed Nov. 24, 2014)"
   </a>
   .
   <i>
    United States District Court for the Eastern District of Pennsylvania
   </i>
   <span class="reference-accessdate">
    . Retrieved
    <span class="nowrap">
     5 November
    </span>
    2015
   </span>
   .
  </cite>
  <span class="Z3988" title="ctx_ver=Z39.88-2004&amp;rft_val_fmt=info%3Aofi%2Ffmt%3Akev
  </span>
  <link href="mw-data:TemplateStyles:r886058088" rel="mw-deduplicated-inline-style"/>
 </span>
</li>
<li id="cite_note-17">
 <span class="mw-cite-backlink">
  <b>
   <a href="#cite_ref-17">
    ^
   </a>
  </b>
 </span>
 <span class="reference-text">
  <cite class="citation journal">
   Neuburger, Jeffrey D (5 December 2014).
   <a class="external text" href="http://newmedialaw.proskauer.com/2014/12/05/qvc-sues-
    "QVC Sues Shopping App for Web Scraping That Allegedly Triggered Site Outage"
   </a>
   .
   <i>
    The National Law Review
   </i>
   . Proskauer Rose LLP
   <span class="reference-accessdate">
    . Retrieved
    <span class="nowrap">
     5 November
    </span>
    2015
```

```
   </span>
   .
  </cite>
  <span class="Z3988" title="ctx_ver=Z39.88-2004&amp;rft_val_fmt=info%3Aofi%2Ffmt%3Akev
  </span>
  <link href="mw-data:TemplateStyles:r886058088" rel="mw-deduplicated-inline-style"/>
 </span>
</li>
<li id="cite_note-18">
 <span class="mw-cite-backlink">
  <b>
   <a href="#cite_ref-18">
    ^
   </a>
  </b>
 </span>
 <span class="reference-text">
  <cite class="citation web">
   <a class="external text" href="http://www.fornova.net/documents/pblog-bna-com.pdf" r
    "Did Iqbal/Twombly Raise the Bar for Browsewrap Claims?"
   </a>
   <span class="cs1-format">
    (PDF)
   </span>
   . 2010-09-17
   <span class="reference-accessdate">
    . Retrieved
    <span class="nowrap">
     2010-10-27
    </span>
   </span>
   .
  </cite>
  <span class="Z3988" title="ctx_ver=Z39.88-2004&amp;rft_val_fmt=info%3Aofi%2Ffmt%3Akev
  </span>
  <link href="mw-data:TemplateStyles:r886058088" rel="mw-deduplicated-inline-style"/>
 </span>
</li>
<li id="cite_note-19">
 <span class="mw-cite-backlink">
  <b>
   <a href="#cite_ref-19">
    ^
   </a>
  </b>
 </span>
 <span class="reference-text">
  <cite class="citation web">
```

```
   <a class="external text" href="https://www.techdirt.com/articles/20090605/222820514
    "Can Scraping Non-Infringing Content Become Copyright Infringement... Because Of H
   </a>
   .
   <i>
    Techdirt
   </i>
   . 2009-06-10
   <span class="reference-accessdate">
    . Retrieved
    <span class="nowrap">
     2016-05-24
    </span>
   </span>
   .
  </cite>
  <span class="Z3988" title="ctx_ver=Z39.88-2004&amp;rft_val_fmt=info%3Aofi%2Ffmt%3Ake
  </span>
  <link href="mw-data:TemplateStyles:r886058088" rel="mw-deduplicated-inline-style"/>
 </span>
</li>
<li id="cite_note-20">
 <span class="mw-cite-backlink">
  <b>
   <a href="#cite_ref-20">
    ^
   </a>
  </b>
 </span>
 <span class="reference-text">
  <cite class="citation web">
   <a class="external text" href="https://www.eff.org/cases/facebook-v-power-ventures"
    "Facebook v. Power Ventures"
   </a>
   .
   <i>
    Electronic Frontier Foundation
   </i>
   <span class="reference-accessdate">
    . Retrieved
    <span class="nowrap">
     2016-05-24
    </span>
   </span>
   .
  </cite>
  <span class="Z3988" title="ctx_ver=Z39.88-2004&amp;rft_val_fmt=info%3Aofi%2Ffmt%3Ake
  </span>
```

```
  <link href="mw-data:TemplateStyles:r886058088" rel="mw-deduplicated-inline-style"/>
 </span>
</li>
<li id="cite_note-21">
 <span class="mw-cite-backlink">
  <b>
   <a href="#cite_ref-21">
    ^
   </a>
  </b>
 </span>
 <span class="reference-text">
  <cite class="citation web">
   <a class="external text" href="https://web.archive.org/web/20071012005033/http://ww
    "UDSKRIFT AF SØ- &amp;amp; HANDELSRETTENS DOMBOG"
   </a>
   <span class="cs1-format">
    (PDF)
   </span>
   (in Danish). bvhd.dk. 2006-02-24. Archived from
   <a class="external text" href="http://www.bvhd.dk/uploads/tx_mocarticles/S_-_og_Han
    the original
   </a>
   <span class="cs1-format">
    (PDF)
   </span>
   on 2007-10-12
   <span class="reference-accessdate">
    . Retrieved
    <span class="nowrap">
     2007-05-30
    </span>
   </span>
   .
  </cite>
  <span class="Z3988" title="ctx_ver=Z39.88-2004&amp;amp;rft_val_fmt=info%3Aofi%2Ffmt%3Ake
  </span>
  <link href="mw-data:TemplateStyles:r886058088" rel="mw-deduplicated-inline-style"/>
 </span>
</li>
<li id="cite_note-22">
 <span class="mw-cite-backlink">
  <b>
   <a href="#cite_ref-22">
    ^
   </a>
  </b>
 </span>
```

<cite class="citation web">
<a class="external text" href="http://www.bailii.org/ie/cases/IEHC/2010/H47.html" re
"High Court of Ireland Decisions >> Ryanair Ltd -v- Billigfluege.de GMBH 201

. British and Irish Legal Information Institute. 2010-02-26

. Retrieved

2012-04-19

.
</cite>
<span class="Z3988" title="ctx_ver=Z39.88-2004&rft_val_fmt=info%3Aofi%2Ffmt%3Ake

<link href="mw-data:TemplateStyles:r886058088" rel="mw-deduplicated-inline-style"/>

<li id="cite_note-23">

^

<cite class="citation web">
Matthews, Áine (June 2010).
<a class="external text" href="http://www.lkshields.ie/htmdocs/publications/newslet
"Intellectual Property: Website Terms of Use"

.

Issue 26: June 2010

. LK Shields Solicitors Update. p.ã03

. Retrieved

2012-04-19

.
</cite>
<span class="Z3988" title="ctx_ver=Z39.88-2004&rft_val_fmt=info%3Aofi%2Ffmt%3Ake

```
  <link href="mw-data:TemplateStyles:r886058088" rel="mw-deduplicated-inline-style"/>
 </span>
</li>
<li id="cite_note-24">
 <span class="mw-cite-backlink">
  <b>
   <a href="#cite_ref-24">
    ^
   </a>
  </b>
 </span>
 <span class="reference-text">
  <cite class="citation web">
   National Office for the Information Economy (February 2004).
   <a class="external text" href="https://www.lloyds.com/~/media/5880dae185914b2487bed
    "Spam Act 2003: An overview for business"
   </a>
   . Australian Communications Authority. p.ã6
   <span class="reference-accessdate">
    . Retrieved
    <span class="nowrap">
     2017-12-07
    </span>
   </span>
   .
  </cite>
  <span class="Z3988" title="ctx_ver=Z39.88-2004&amp;rft_val_fmt=info%3Aofi%2Ffmt%3Ake
  </span>
  <link href="mw-data:TemplateStyles:r886058088" rel="mw-deduplicated-inline-style"/>
 </span>
</li>
<li id="cite_note-25">
 <span class="mw-cite-backlink">
  <b>
   <a href="#cite_ref-25">
    ^
   </a>
  </b>
 </span>
 <span class="reference-text">
  <cite class="citation web">
   National Office for the Information Economy (February 2004).
   <a class="external text" href="http://www.webstartdesign.com.au/spam_business_practi
    "Spam Act 2003: A practical guide for business"
   </a>
   <span class="cs1-format">
    (PDF)
   </span>
```

```
          . Australian Communications Authority. p.ã20
          <span class="reference-accessdate">
           . Retrieved
           <span class="nowrap">
            2017-12-07
           </span>
          </span>
          .
         </cite>
         <span class="Z3988" title="ctx_ver=Z39.88-2004&amp;rft_val_fmt=info%3Aofi%2Ffmt%3Ake
         </span>
         <link href="mw-data:TemplateStyles:r886058088" rel="mw-deduplicated-inline-style"/>
        </span>
       </li>
       <li id="cite_note-26">
        <span class="mw-cite-backlink">
         <b>
          <a href="#cite_ref-26">
           ^
          </a>
         </b>
        </span>
        <span class="reference-text">
         Mayank Dhiman
         <a class="external text" href="https://s3.us-west-2.amazonaws.com/research-papers-my
          Breaking Fraud &amp;amp; Bot Detection Solutions
         </a>
         <i>
          OWASP AppSec Cali' 2018
         </i>
         Retrieved February 10, 2018.
        </span>
       </li>
      </ol>
     </div>
     <!--
NewPP limit report
Parsed by mw1264
Cached time: 20190516030922
Cache expiry: 2592000
Dynamic content: false
CPU time usage: 0.492 seconds
Real time usage: 0.660 seconds
Preprocessor visited node count: 1805/1000000
Preprocessor generated node count: 0/1500000
Postexpand include size: 65153/2097152 bytes
Template argument size: 772/2097152 bytes
Highest expansion depth: 15/40
```

```
Expensive parser function count: 8/500
Unstrip recursion depth: 1/20
Unstrip postexpand size: 70826/5000000 bytes
Number of Wikibase entities loaded: 3/400
Lua time usage: 0.279/10.000 seconds
Lua memory usage: 6.63 MB/50 MB
-->
      <!--
Transclusion expansion time report (%,ms,calls,template)
100.00%  548.078      1 -total
 59.50%  326.090      1 Template:Reflist
 38.08%  208.730      5 Template:Cite_journal
 19.94%  109.301      1 Template:More_citations_needed
 18.46%  101.199      4 Template:Ambox
 16.32%   89.451     18 Template:Cite_web
  6.60%   36.163      1 Template:Find_sources_mainspace
  5.52%   30.244      1 Template:Split_section
  4.82%   26.437      1 Template:Split_portions
  2.82%   15.468      1 Template:US-centric
-->
      <!-- Saved in parser cache with key enwiki:pcache:idhash:2696619-0!canonical and timestar
 -->
     </div>
     <noscript>
      <img alt="" height="1" src="//en.wikipedia.org/wiki/Special:CentralAutoLogin/start?type=
     </noscript>
    </div>
    <div class="printfooter">
     Retrieved from "
     <a dir="ltr" href="https://en.wikipedia.org/w/index.php?title=Web_scraping&amp;oldid=89718
      https://en.wikipedia.org/w/index.php?title=Web_scraping&amp;oldid=897184947
     </a>
     "
    </div>
    <div class="catlinks" data-mw="interface" id="catlinks">
     <div class="mw-normal-catlinks" id="mw-normal-catlinks">
      <a href="/wiki/Help:Category" title="Help:Category">
       Categories
      </a>
      :
      <ul>
       <li>
        <a href="/wiki/Category:Web_scraping" title="Category:Web scraping">
         Web scraping
        </a>
       </li>
      </ul>
     </div>
```

```
 <div class="mw-hidden-catlinks mw-hidden-cats-hidden" id="mw-hidden-catlinks">
  Hidden categories:
  <ul>
   <li>
    <a href="/wiki/Category:CS1_Danish-language_sources_(da)" title="Category:CS1 Danish-la
     CS1 Danish-language sources (da)
    </a>
   </li>
   <li>
    <a href="/wiki/Category:Articles_needing_additional_references_from_June_2017" title="C
     Articles needing additional references from June 2017
    </a>
   </li>
   <li>
    <a href="/wiki/Category:All_articles_needing_additional_references" title="Category:All
     All articles needing additional references
    </a>
   </li>
   <li>
    <a href="/wiki/Category:Articles_needing_additional_references_from_October_2018" title
     Articles needing additional references from October 2018
    </a>
   </li>
   <li>
    <a href="/wiki/Category:Articles_with_limited_geographic_scope_from_October_2015" title
     Articles with limited geographic scope from October 2015
    </a>
   </li>
   <li>
    <a href="/wiki/Category:USA-centric" title="Category:USA-centric">
     USA-centric
    </a>
   </li>
   <li>
    <a href="/wiki/Category:Articles_to_be_split_from_July_2018" title="Category:Articles t
     Articles to be split from July 2018
    </a>
   </li>
   <li>
    <a href="/wiki/Category:All_articles_to_be_split" title="Category:All articles to be sp
     All articles to be split
    </a>
   </li>
  </ul>
 </div>
</div>
<div class="visualClear">
</div>
```

```
 </div>
</div>
<div id="mw-navigation">
 <h2>
  Navigation menu
 </h2>
 <div id="mw-head">
  <div aria-labelledby="p-personal-label" id="p-personal" role="navigation">
   <h3 id="p-personal-label">
    Personal tools
   </h3>
   <ul>
    <li id="pt-anonuserpage">
     Not logged in
    </li>
    <li id="pt-anontalk">
     <a accesskey="n" href="/wiki/Special:MyTalk" title="Discussion about edits from this IP
      Talk
     </a>
    </li>
    <li id="pt-anoncontribs">
     <a accesskey="y" href="/wiki/Special:MyContributions" title="A list of edits made from t
      Contributions
     </a>
    </li>
    <li id="pt-createaccount">
     <a href="/w/index.php?title=Special:CreateAccount&amp;returnto=Web+scraping" title="You
      Create account
     </a>
    </li>
    <li id="pt-login">
     <a accesskey="o" href="/w/index.php?title=Special:UserLogin&amp;returnto=Web+scraping" t
      Log in
     </a>
    </li>
   </ul>
  </div>
  <div id="left-navigation">
   <div aria-labelledby="p-namespaces-label" class="vectorTabs" id="p-namespaces" role="navig
    <h3 id="p-namespaces-label">
     Namespaces
    </h3>
    <ul>
     <li class="selected" id="ca-nstab-main">
      <span>
       <a accesskey="c" href="/wiki/Web_scraping" title="View the content page [c]">
        Article
       </a>
```

```
    </span>
   </li>
   <li id="ca-talk">
    <span>
     <a accesskey="t" href="/wiki/Talk:Web_scraping" rel="discussion" title="Discussion ab
      Talk
     </a>
    </span>
   </li>
  </ul>
 </div>
 <div aria-labelledby="p-variants-label" class="vectorMenu emptyPortlet" id="p-variants" r
  <input aria-labelledby="p-variants-label" class="vectorMenuCheckbox" type="checkbox"/>
  <h3 id="p-variants-label">
   <span>
    Variants
   </span>
  </h3>
  <ul class="menu">
  </ul>
 </div>
</div>
<div id="right-navigation">
 <div aria-labelledby="p-views-label" class="vectorTabs" id="p-views" role="navigation">
  <h3 id="p-views-label">
   Views
  </h3>
  <ul>
   <li class="collapsible selected" id="ca-view">
    <span>
     <a href="/wiki/Web_scraping">
      Read
     </a>
    </span>
   </li>
   <li class="collapsible" id="ca-edit">
    <span>
     <a accesskey="e" href="/w/index.php?title=Web_scraping&amp;action=edit" title="Edit t
      Edit
     </a>
    </span>
   </li>
   <li class="collapsible" id="ca-history">
    <span>
     <a accesskey="h" href="/w/index.php?title=Web_scraping&amp;action=history" title="Pas
      View history
     </a>
    </span>
```

```
    </li>
   </ul>
  </div>
  <div aria-labelledby="p-cactions-label" class="vectorMenu emptyPortlet" id="p-cactions" r
   <input aria-labelledby="p-cactions-label" class="vectorMenuCheckbox" type="checkbox"/>
   <h3 id="p-cactions-label">
    <span>
     More
    </span>
   </h3>
   <ul class="menu">
   </ul>
  </div>
  <div id="p-search" role="search">
   <h3>
    <label for="searchInput">
     Search
    </label>
   </h3>
   <form action="/w/index.php" id="searchform">
    <div id="simpleSearch">
     <input accesskey="f" id="searchInput" name="search" placeholder="Search Wikipedia" titl
     <input name="title" type="hidden" value="Special:Search"/>
     <input class="searchButton mw-fallbackSearchButton" id="mw-searchButton" name="fulltext
     <input class="searchButton" id="searchButton" name="go" title="Go to a page with this e
    </div>
   </form>
  </div>
 </div>
</div>
<div id="mw-panel">
 <div id="p-logo" role="banner">
  <a class="mw-wiki-logo" href="/wiki/Main_Page" title="Visit the main page">
  </a>
 </div>
 <div aria-labelledby="p-navigation-label" class="portal" id="p-navigation" role="navigation
  <h3 id="p-navigation-label">
   Navigation
  </h3>
  <div class="body">
   <ul>
    <li id="n-mainpage-description">
     <a accesskey="z" href="/wiki/Main_Page" title="Visit the main page [z]">
      Main page
     </a>
    </li>
    <li id="n-contents">
     <a href="/wiki/Portal:Contents" title="Guides to browsing Wikipedia">
```

```
     Contents
    </a>
   </li>
   <li id="n-featuredcontent">
    <a href="/wiki/Portal:Featured_content" title="Featured content  the best of Wikipedia"
     Featured content
    </a>
   </li>
   <li id="n-currentevents">
    <a href="/wiki/Portal:Current_events" title="Find background information on current eve
     Current events
    </a>
   </li>
   <li id="n-randompage">
    <a accesskey="x" href="/wiki/Special:Random" title="Load a random article [x]">
     Random article
    </a>
   </li>
   <li id="n-sitesupport">
    <a href="https://donate.wikimedia.org/wiki/Special:FundraiserRedirector?utm_source=dona
     Donate to Wikipedia
    </a>
   </li>
   <li id="n-shoplink">
    <a href="//shop.wikimedia.org" title="Visit the Wikipedia store">
     Wikipedia store
    </a>
   </li>
  </ul>
 </div>
</div>
<div aria-labelledby="p-interaction-label" class="portal" id="p-interaction" role="navigati
 <h3 id="p-interaction-label">
  Interaction
 </h3>
 <div class="body">
  <ul>
   <li id="n-help">
    <a href="/wiki/Help:Contents" title="Guidance on how to use and edit Wikipedia">
     Help
    </a>
   </li>
   <li id="n-aboutsite">
    <a href="/wiki/Wikipedia:About" title="Find out about Wikipedia">
     About Wikipedia
    </a>
   </li>
   <li id="n-portal">
```

```
    <a href="/wiki/Wikipedia:Community_portal" title="About the project, what you can do, w
     Community portal
    </a>
   </li>
   <li id="n-recentchanges">
    <a accesskey="r" href="/wiki/Special:RecentChanges" title="A list of recent changes in
     Recent changes
    </a>
   </li>
   <li id="n-contactpage">
    <a href="//en.wikipedia.org/wiki/Wikipedia:Contact_us" title="How to contact Wikipedia"
     Contact page
    </a>
   </li>
  </ul>
 </div>
</div>
<div aria-labelledby="p-tb-label" class="portal" id="p-tb" role="navigation">
 <h3 id="p-tb-label">
  Tools
 </h3>
 <div class="body">
  <ul>
   <li id="t-whatlinkshere">
    <a accesskey="j" href="/wiki/Special:WhatLinksHere/Web_scraping" title="List of all Eng
     What links here
    </a>
   </li>
   <li id="t-recentchangeslinked">
    <a accesskey="k" href="/wiki/Special:RecentChangesLinked/Web_scraping" rel="nofollow" t
     Related changes
    </a>
   </li>
   <li id="t-upload">
    <a accesskey="u" href="/wiki/Wikipedia:File_Upload_Wizard" title="Upload files [u]">
     Upload file
    </a>
   </li>
   <li id="t-specialpages">
    <a accesskey="q" href="/wiki/Special:SpecialPages" title="A list of all special pages
     Special pages
    </a>
   </li>
   <li id="t-permalink">
    <a href="/w/index.php?title=Web_scraping&amp;oldid=897184947" title="Permanent link to
     Permanent link
    </a>
   </li>
```

```
   <li id="t-info">
    <a href="/w/index.php?title=Web_scraping&amp;action=info" title="More information about
     Page information
    </a>
   </li>
   <li id="t-wikibase">
    <a accesskey="g" href="https://www.wikidata.org/wiki/Special:EntityPage/Q665452" title=
     Wikidata item
    </a>
   </li>
   <li id="t-cite">
    <a href="/w/index.php?title=Special:CiteThisPage&amp;page=Web_scraping&amp;id=89718494
     Cite this page
    </a>
   </li>
  </ul>
 </div>
</div>
<div aria-labelledby="p-coll-print_export-label" class="portal" id="p-coll-print_export" ro
 <h3 id="p-coll-print_export-label">
  Print/export
 </h3>
 <div class="body">
  <ul>
   <li id="coll-create_a_book">
    <a href="/w/index.php?title=Special:Book&amp;bookcmd=book_creator&amp;referer=Web+scrap
     Create a book
    </a>
   </li>
   <li id="coll-download-as-rdf2latex">
    <a href="/w/index.php?title=Special:ElectronPdf&amp;page=Web+scraping&amp;action=show-d
     Download as PDF
    </a>
   </li>
   <li id="t-print">
    <a accesskey="p" href="/w/index.php?title=Web_scraping&amp;printable=yes" title="Printa
     Printable version
    </a>
   </li>
  </ul>
 </div>
</div>
<div aria-labelledby="p-lang-label" class="portal" id="p-lang" role="navigation">
 <h3 id="p-lang-label">
  Languages
 </h3>
 <div class="body">
  <ul>
```

```
<li class="interlanguage-link interwiki-ar">
 <a class="interlanguage-link-target" href="https://ar.wikipedia.org/wiki/%D8%A7%D8%B3%

 </a>
</li>
<li class="interlanguage-link interwiki-ca">
 <a class="interlanguage-link-target" href="https://ca.wikipedia.org/wiki/Web_scraping"
  Català
 </a>
</li>
<li class="interlanguage-link interwiki-de">
 <a class="interlanguage-link-target" href="https://de.wikipedia.org/wiki/Screen_Scrapi
  Deutsch
 </a>
</li>
<li class="interlanguage-link interwiki-es">
 <a class="interlanguage-link-target" href="https://es.wikipedia.org/wiki/Web_scraping"
  Español
 </a>
</li>
<li class="interlanguage-link interwiki-eu">
 <a class="interlanguage-link-target" href="https://eu.wikipedia.org/wiki/Web_scraping"
  Euskara
 </a>
</li>
<li class="interlanguage-link interwiki-fr">
 <a class="interlanguage-link-target" href="https://fr.wikipedia.org/wiki/Web_scraping"
  Français
 </a>
</li>
<li class="interlanguage-link interwiki-id">
 <a class="interlanguage-link-target" href="https://id.wikipedia.org/wiki/Web_scraping"
  Bahasa Indonesia
 </a>
</li>
<li class="interlanguage-link interwiki-is">
 <a class="interlanguage-link-target" href="https://is.wikipedia.org/wiki/Vefs%C3%B6fnu
  Íslenska
 </a>
</li>
<li class="interlanguage-link interwiki-it">
 <a class="interlanguage-link-target" href="https://it.wikipedia.org/wiki/Web_scraping"
  Italiano
 </a>
</li>
<li class="interlanguage-link interwiki-lv">
 <a class="interlanguage-link-target" href="https://lv.wikipedia.org/wiki/Rasmo%C5%A1ana
  Latvieu
```

```
      </a>
     </li>
     <li class="interlanguage-link interwiki-nl">
      <a class="interlanguage-link-target" href="https://nl.wikipedia.org/wiki/Scrapen" hrefl
       Nederlands
      </a>
     </li>
     <li class="interlanguage-link interwiki-ja">
      <a class="interlanguage-link-target" href="https://ja.wikipedia.org/wiki/%E3%82%A6%E3%8

      </a>
     </li>
     <li class="interlanguage-link interwiki-sr">
      <a class="interlanguage-link-target" href="https://sr.wikipedia.org/wiki/Web_scraping"
        / srpski
      </a>
     </li>
     <li class="interlanguage-link interwiki-tr">
      <a class="interlanguage-link-target" href="https://tr.wikipedia.org/wiki/Web_kaz%C4%B1n
       Türkçe
      </a>
     </li>
     <li class="interlanguage-link interwiki-uk">
      <a class="interlanguage-link-target" href="https://uk.wikipedia.org/wiki/Web_scraping"

      </a>
     </li>
     <li class="interlanguage-link interwiki-zh">
      <a class="interlanguage-link-target" href="https://zh.wikipedia.org/wiki/%E7%BD%91%E9%A

      </a>
     </li>
    </ul>
    <div class="after-portlet after-portlet-lang">
     <span class="wb-langlinks-edit wb-langlinks-link">
      <a class="wbc-editpage" href="https://www.wikidata.org/wiki/Special:EntityPage/Q665452#
       Edit links
      </a>
     </span>
    </div>
   </div>
  </div>
 </div>
</div>
<div id="footer" role="contentinfo">
 <ul id="footer-info">
  <li id="footer-info-lastmod">
   This page was last edited on 15 May 2019, at 09:54
```

```
      <span class="anonymous-show">
       (UTC)
      </span>
      .
     </li>
     <li id="footer-info-copyright">
      Text is available under the
      <a href="//en.wikipedia.org/wiki/Wikipedia:Text_of_Creative_Commons_Attribution-ShareAlik
       Creative Commons Attribution-ShareAlike License
      </a>
      <a href="//creativecommons.org/licenses/by-sa/3.0/" rel="license" style="display:none;">
      </a>
      ;
additional terms may apply.  By using this site, you agree to the
      <a href="//foundation.wikimedia.org/wiki/Terms_of_Use">
       Terms of Use
      </a>
      and
      <a href="//foundation.wikimedia.org/wiki/Privacy_policy">
       Privacy Policy
      </a>
      . Wikipediaő is a registered trademark of the
      <a href="//www.wikimediafoundation.org/">
       Wikimedia Foundation, Inc.
      </a>
      , a non-profit organization.
     </li>
    </ul>
    <ul id="footer-places">
     <li id="footer-places-privacy">
      <a class="extiw" href="https://foundation.wikimedia.org/wiki/Privacy_policy" title="wmf:Pr
       Privacy policy
      </a>
     </li>
     <li id="footer-places-about">
      <a href="/wiki/Wikipedia:About" title="Wikipedia:About">
       About Wikipedia
      </a>
     </li>
     <li id="footer-places-disclaimer">
      <a href="/wiki/Wikipedia:General_disclaimer" title="Wikipedia:General disclaimer">
       Disclaimers
      </a>
     </li>
     <li id="footer-places-contact">
      <a href="//en.wikipedia.org/wiki/Wikipedia:Contact_us">
       Contact Wikipedia
      </a>
```

```
    </li>
    <li id="footer-places-developers">
     <a href="https://www.mediawiki.org/wiki/Special:MyLanguage/How_to_contribute">
      Developers
     </a>
    </li>
    <li id="footer-places-cookiestatement">
     <a href="https://foundation.wikimedia.org/wiki/Cookie_statement">
      Cookie statement
     </a>
    </li>
    <li id="footer-places-mobileview">
     <a class="noprint stopMobileRedirectToggle" href="//en.m.wikipedia.org/w/index.php?title=W
      Mobile view
     </a>
    </li>
   </ul>
   <ul class="noprint" id="footer-icons">
    <li id="footer-copyrightico">
     <a href="https://wikimediafoundation.org/">
      <img alt="Wikimedia Foundation" height="31" src="/static/images/wikimedia-button.png" src
     </a>
    </li>
    <li id="footer-poweredbyico">
     <a href="//www.mediawiki.org/">
      <img alt="Powered by MediaWiki" height="31" src="/static/images/poweredby_mediawiki_88x31
     </a>
    </li>
   </ul>
   <div style="clear: both;">
   </div>
  </div>
  <script>
   (window.RLQ=window.RLQ||[]).push(function(){mw.config.set({"wgPageParseReport":{"limitreport
  </script>
  <script type="application/ld+json">
   {"@context":"https:\/\/schema.org","@type":"Article","name":"Web scraping","url":"https:\/\/
  </script>
  <script>
   (window.RLQ=window.RLQ||[]).push(function(){mw.config.set({"wgBackendResponseTime":807,"wgHo
  </script>
 </body>
</html>
```

Podemos utilizar um seletor CSS para extrair somente o corpo do HTML, veja que agora já está mais fácil encontrar o primeiro parágrafo, basta localizar o < p> a esquerda:

In [68]:
```
#print(soup.find('body').prettify()) # é uma outra opção via find
soup.select("html body")
```

Out[68]:
```
[<body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject mw-editable
 <div class="noprint" id="mw-page-base"></div>
 <div class="noprint" id="mw-head-base"></div>
 <div class="mw-body" id="content" role="main">
 <a id="top"></a>
 <div class="mw-body-content" id="siteNotice"><!-- CentralNotice --></div>
 <div class="mw-indicators mw-body-content">
 </div>
 <h1 class="firstHeading" id="firstHeading" lang="en">Web scraping</h1>
 <div class="mw-body-content" id="bodyContent">
 <div class="noprint" id="siteSub">From Wikipedia, the free encyclopedia</div>
 <div id="contentSub"></div>
 <div id="jump-to-nav"></div>
 <a class="mw-jump-link" href="#mw-head">Jump to navigation</a>
 <a class="mw-jump-link" href="#p-search">Jump to search</a>
 <div class="mw-content-ltr" dir="ltr" id="mw-content-text" lang="en"><div class="mw-p
 <div class="hatnote navigation-not-searchable" role="note">For broader coverage of th
 <p><b>Web scraping</b>, <b>web harvesting</b>, or <b>web data extraction</b> is <a hr
 </p><p>Web scraping a web page involves fetching it and extracting from it.<sup class
 </p><p>Web scraping is used for <a href="/wiki/Contact_scraping" title="Contact scrap
 </p><p><a href="/wiki/Web_page" title="Web page">Web pages</a> are built using text-b
 </p><p>Newer forms of web scraping involve listening to data feeds from web servers.
 </p><p>There are methods that some websites use to prevent web scraping, such as dete
 </p>
 <div class="toc" id="toc"><input class="toctogglecheckbox" id="toctogglecheckbox" rol
 <ul>
 <li class="toclevel-1 tocsection-1"><a href="#History"><span class="tocnumber">1</spa
 <li class="toclevel-1 tocsection-2"><a href="#Techniques"><span class="tocnumber">2</
 <ul>
 <li class="toclevel-2 tocsection-3"><a href="#Human_copy-and-paste"><span class="tocn
 <li class="toclevel-2 tocsection-4"><a href="#Text_pattern_matching"><span class="toc
 <li class="toclevel-2 tocsection-5"><a href="#HTTP_programming"><span class="tocnumbe
 <li class="toclevel-2 tocsection-6"><a href="#HTML_parsing"><span class="tocnumber">2
 <li class="toclevel-2 tocsection-7"><a href="#DOM_parsing"><span class="tocnumber">2.
 <li class="toclevel-2 tocsection-8"><a href="#Vertical_aggregation"><span class="tocn
 <li class="toclevel-2 tocsection-9"><a href="#Semantic_annotation_recognizing"><span
 <li class="toclevel-2 tocsection-10"><a href="#Computer_vision_web-page_analysis"><sp
 </ul>
 </li>
 <li class="toclevel-1 tocsection-11"><a href="#Software"><span class="tocnumber">3</s
 <ul>
 <li class="toclevel-2 tocsection-12"><a href="#Example_tools"><span class="tocnumber"
 <ul>
 <li class="toclevel-3 tocsection-13"><a href="#Javascript_tools"><span class="tocnumb
 <li class="toclevel-3 tocsection-14"><a href="#Web_crawling_frameworks"><span class="
```

```
</ul>
</li>
</ul>
</li>
<li class="toclevel-1 tocsection-15"><a href="#Legal_issues"><span class="tocnumber">
<ul>
<li class="toclevel-2 tocsection-16"><a href="#United_States"><span class="tocnumber"
<li class="toclevel-2 tocsection-17"><a href="#The_EU"><span class="tocnumber">4.2</s
<li class="toclevel-2 tocsection-18"><a href="#Australia"><span class="tocnumber">4.3
</ul>
</li>
<li class="toclevel-1 tocsection-19"><a href="#Methods_to_prevent_web_scraping"><spar
<li class="toclevel-1 tocsection-20"><a href="#See_also"><span class="tocnumber">6</s
<li class="toclevel-1 tocsection-21"><a href="#References"><span class="tocnumber">7<
</ul>
</div>
<h2><span class="mw-headline" id="History">History</span><span class="mw-editsection"
<table class="box-Unreferenced_section plainlinks metadata ambox ambox-content ambox-
<p>The history of the web scraping is actually much longer, dating back significantly
</p>
<ul><li>After the birth of <a href="/wiki/History_of_the_World_Wide_Web" title="Histo
<li>In 1993,December, the First <b>crawler-based web search engine</b> - <a href="/wi
<li>In 2000, the <b>first Web API and API crawler</b> came. <a href="/wiki/Applicatio
<li>In 2004, <a href="/wiki/Beautiful_Soup_(HTML_parser)" title="Beautiful Soup (HTML
<h2><span class="mw-headline" id="Techniques">Techniques</span><span class="mw-editse
<p>Web scraping is the process of automatically mining data or collecting information
</p>
<h3><span class="mw-headline" id="Human_copy-and-paste">Human copy-and-paste</span><s
<p>Sometimes even the best web-scraping technology cannot replace a humans manual exa
</p>
<h3><span class="mw-headline" id="Text_pattern_matching">Text pattern matching</span>
<p>A simple yet powerful approach to extract information from web pages can be based
</p>
<h3><span class="mw-headline" id="HTTP_programming">HTTP programming</span><span clas
<p><a href="/wiki/Static_web_page" title="Static web page">Static</a> and <a href="/w
</p>
<h3><span class="mw-headline" id="HTML_parsing">HTML parsing</span><span class="mw-ed
<p>Many websites have large collections of pages generated dynamically from an underl
</p>
<h3><span class="mw-headline" id="DOM_parsing">DOM parsing</span><span class="mw-edit
<div class="hatnote navigation-not-searchable" role="note">Further information: <a hr
<p>By embedding a full-fledged web browser, such as the <a href="/wiki/Internet_Explo
</p>
<h3><span class="mw-headline" id="Vertical_aggregation">Vertical aggregation</span><s
<p>There are several companies that have developed vertical specific harvesting plati
</p>
<h3><span class="mw-headline" id="Semantic_annotation_recognizing">Semantic annotatio
<p>The pages being scraped may embrace <a href="/wiki/Metadata" title="Metadata">meta
```

```
</p>
<h3><span class="mw-headline" id="Computer_vision_web-page_analysis">Computer vision
<p>There are efforts using <a href="/wiki/Machine_learning" title="Machine learning">
</p>
<h2><span class="mw-headline" id="Software">Software</span><span class="mw-editsectio
<p>There are many software tools available that can be used to customize web-scraping
</p>
<h3><span class="mw-headline" id="Example_tools">Example tools</span><span class="mw-
<ul><li><a href="/wiki/CURL" title="CURL">cURL</a>  command line tool and library for
<li><a href="/wiki/Data_Toolbar" title="Data Toolbar">Data Toolbar</a>  web scraping
<li><a href="/wiki/Diffbot" title="Diffbot">Diffbot</a>  uses computer vision and ma
<li><a href="/wiki/Heritrix" title="Heritrix">Heritrix</a>  gets pages (lots of them)
<li><a href="/wiki/HtmlUnit" title="HtmlUnit">HtmlUnit</a>  headless browser that can
<li><a href="/wiki/HTTrack" title="HTTrack">HTTrack</a>  free and open source Web cra
<li><a href="/wiki/IMacros" title="IMacros">iMacros</a>  a browser extension to recor
<li><a href="/wiki/Selenium_(software)" title="Selenium (software)">Selenium (softwar
<li><a href="/wiki/Aptana#Aptana_Jaxer" title="Aptana">Jaxer</a></li>
<li><a href="/wiki/Mozenda" title="Mozenda">Mozenda</a>  is a WYSIWYG software that
<li><a href="/wiki/Nokogiri_(software)" title="Nokogiri (software)">nokogiri</a></li>
<li><a href="/wiki/OutWit_Hub" title="OutWit Hub">OutWit Hub</a>  Web scraping applic
<li><a href="/wiki/Watir" title="Watir">watir</a> </li>
<li><a href="/wiki/Wget" title="Wget">Wget</a>  computer program that retrieves conte
<li><a href="/wiki/WSO2_Mashup_Server" title="WSO2 Mashup Server">WSO2 Mashup Server<
<li><a href="/wiki/Yahoo!_Query_Language" title="Yahoo! Query Language">Yahoo! Query
<h4><span class="mw-headline" id="Javascript_tools">Javascript tools</span><span clas
<ul><li><a href="/wiki/Greasemonkey" title="Greasemonkey">Greasemonkey</a></li>
<li><a href="/wiki/Node.js" title="Node.js">Node.js</a></li>
<li><a href="/wiki/PhantomJS" title="PhantomJS">PhantomJS</a>  scripted, <a href="/wi
<li><a href="/wiki/JQuery" title="JQuery">jQuery</a></li></ul>
<h4><span class="mw-headline" id="Web_crawling_frameworks">Web crawling frameworks</s
<p>These can be used to build web scrapers.
</p>
<ul><li><a href="/wiki/Scrapy" title="Scrapy">Scrapy</a></li></ul>
<h2><span class="mw-headline" id="Legal_issues">Legal issues</span><span class="mw-ed
<table class="box-Globalize plainlinks metadata ambox ambox-content ambox-globalize"
<p>The legality of web scraping varies across the world. In general, web scraping may
</p>
<h3><span class="mw-headline" id="United_States">United States</span><span class="mw-
<table class="plainlinks metadata ambox ambox-move" role="presentation"><tbody><tr><t
<p>In the United States, website owners can use three major <a href="/wiki/Cause_of_a
</p><p>U.S. courts have acknowledged that users of "scrapers" or "robots" may be held
</p><p>One of the first major tests of <a class="mw-redirect" href="/wiki/Screen_scra
</p><p><a href="/wiki/Southwest_Airlines" title="Southwest Airlines">Southwest Airlin
In 2012, a startup called 3Taps scraped classified housing ads from Craigslist. Craig
</p><p>Although these are early scraping decisions, and the theories of liability are
</p><p>While the law in this area becomes more settled, entities contemplating using
</p><p>In the plaintiff's web site during the period of this trial the terms of use l
</p><p>In <i><a href="/wiki/Facebook,_Inc._v._Power_Ventures,_Inc." title="Facebook,
```

```
</p><p>Internet Archive collects and distributes significant number of publicly avail
</p>
<h3><span class="mw-headline" id="The_EU">The EU</span><span class="mw-editsection">
<p>In February 2006, the Danish Maritime and Commercial Court (Copenhagen) ruled that
</p><p>In a February 2010 case complicated by matters of jurisdiction, Ireland's High
<sup class="reference" id="cite_ref-22"><a href="#cite_note-22">[22]</a></sup> The de
</p>
<h3><span class="mw-headline" id="Australia">Australia</span><span class="mw-editsect
<p>In Australia, the <a href="/wiki/Spam_Act_2003" title="Spam Act 2003">Spam Act 200
</p>
<h2><span class="mw-headline" id="Methods_to_prevent_web_scraping">Methods to prevent
<p>The administrator of a website can use various measures to stop or slow a bot. Som
</p>
<ul><li>Blocking an <a href="/wiki/IP_address" title="IP address">IP address</a> eith
<li>Disabling any <a href="/wiki/Web_service" title="Web service">web service</a> <a
<li>Bots sometimes declare who they are (using <a href="/wiki/User_agent" title="User
<li>Bots can be blocked by monitoring excess traffic</li>
<li>Bots can sometimes be blocked with tools to verify that it is a real person acces
<li>Commercial anti-bot services: Companies offer anti-bot and anti-scraping services
<ul><li>Locating bots with a <a href="/wiki/Honeypot_(computing)" title="Honeypot (co
<li><a href="/wiki/Obfuscation" title="Obfuscation">Obfuscation</a> using <a class="m
<li>Because bots rely on consistency in the front-end code of a target website, addin
<li>Websites can declare if crawling is allowed or not in the <a href="/wiki/Robots_e
<h2><span class="mw-headline" id="See_also">See also</span><span class="mw-editsectio
<div class="div-col columns column-width" style="-moz-column-width: 22em; -webkit-col
<ul><li><a class="mw-redirect" href="/wiki/Archive.is" title="Archive.is">Archive.is<
<li><a href="/wiki/Comparison_of_feed_aggregators" title="Comparison of feed aggregat
<li><a href="/wiki/Data_scraping" title="Data scraping">Data scraping</a></li>
<li><a href="/wiki/Data_wrangling" title="Data wrangling">Data wrangling</a></li>
<li><a href="/wiki/Importer_(computing)" title="Importer (computing)">Importer</a></l
<li><a href="/wiki/Job_wrapping" title="Job wrapping">Job wrapping</a></li>
<li><a href="/wiki/Knowledge_extraction" title="Knowledge extraction">Knowledge extra
<li><a href="/wiki/OpenSocial" title="OpenSocial">OpenSocial</a></li>
<li><a href="/wiki/Scraper_site" title="Scraper site">Scraper site</a></li>
<li><a href="/wiki/Fake_news_website" title="Fake news website">Fake news website</a>
<li><a href="/wiki/Blog_scraping" title="Blog scraping">Blog scraping</a></li>
<li><a href="/wiki/Spamdexing" title="Spamdexing">Spamdexing</a></li>
<li><a href="/wiki/Domain_name_drop_list" title="Domain name drop list">Domain name d
<li><a href="/wiki/Text_corpus" title="Text corpus">Text corpus</a></li>
<li><a href="/wiki/Web_archiving" title="Web archiving">Web archiving</a></li>
<li><a class="mw-redirect" href="/wiki/Blog_network" title="Blog network">Blog networ
<li><a class="mw-redirect" href="/wiki/Search_Engine_Scraping" title="Search Engine S
<li><a href="/wiki/Category:Web_crawlers" title="Category:Web crawlers">Web crawlers<
</div>
<h2><span class="mw-headline" id="References">References</span><span class="mw-editse
<div class="reflist columns references-column-width" style="-moz-column-width: 20em;
<ol class="references">
<li id="cite_note-Boeing2016JPER-1"><span class="mw-cite-backlink">^ <a href="#cite_r
```

```
</li>
<li id="cite_note-2"><span class="mw-cite-backlink"><b><a href="#cite_ref-2">^</a></b
</li>
<li id="cite_note-3"><span class="mw-cite-backlink"><b><a href="#cite_ref-3">^</a></b
</li>
<li id="cite_note-4"><span class="mw-cite-backlink"><b><a href="#cite_ref-4">^</a></b
</li>
<li id="cite_note-5"><span class="mw-cite-backlink"><b><a href="#cite_ref-5">^</a></b
</li>
<li id="cite_note-6"><span class="mw-cite-backlink"><b><a href="#cite_ref-6">^</a></b
</li>
<li id="cite_note-7"><span class="mw-cite-backlink"><b><a href="#cite_ref-7">^</a></b
</li>
<li id="cite_note-8"><span class="mw-cite-backlink"><b><a href="#cite_ref-8">^</a></b
</li>
<li id="cite_note-9"><span class="mw-cite-backlink"><b><a href="#cite_ref-9">^</a></b
</li>
<li id="cite_note-10"><span class="mw-cite-backlink"><b><a href="#cite_ref-10">^</a><
</li>
<li id="cite_note-11"><span class="mw-cite-backlink"><b><a href="#cite_ref-11">^</a><
</li>
<li id="cite_note-12"><span class="mw-cite-backlink"><b><a href="#cite_ref-12">^</a><
</li>
<li id="cite_note-impervawp2011-13"><span class="mw-cite-backlink"><b><a href="#cite_
</li>
<li id="cite_note-14"><span class="mw-cite-backlink"><b><a href="#cite_ref-14">^</a><
</li>
<li id="cite_note-15"><span class="mw-cite-backlink"><b><a href="#cite_ref-15">^</a><
</li>
<li id="cite_note-16"><span class="mw-cite-backlink"><b><a href="#cite_ref-16">^</a><
</li>
<li id="cite_note-17"><span class="mw-cite-backlink"><b><a href="#cite_ref-17">^</a><
</li>
<li id="cite_note-18"><span class="mw-cite-backlink"><b><a href="#cite_ref-18">^</a><
</li>
<li id="cite_note-19"><span class="mw-cite-backlink"><b><a href="#cite_ref-19">^</a><
</li>
<li id="cite_note-20"><span class="mw-cite-backlink"><b><a href="#cite_ref-20">^</a><
</li>
<li id="cite_note-21"><span class="mw-cite-backlink"><b><a href="#cite_ref-21">^</a><
</li>
<li id="cite_note-22"><span class="mw-cite-backlink"><b><a href="#cite_ref-22">^</a><
</li>
<li id="cite_note-23"><span class="mw-cite-backlink"><b><a href="#cite_ref-23">^</a><
</li>
<li id="cite_note-24"><span class="mw-cite-backlink"><b><a href="#cite_ref-24">^</a><
</li>
<li id="cite_note-25"><span class="mw-cite-backlink"><b><a href="#cite_ref-25">^</a><
```

```
</li>
<li id="cite_note-26"><span class="mw-cite-backlink"><b><a href="#cite_ref-26">^</a>
</li>
</ol></div>
<!--
NewPP limit report
Parsed by mw1264
Cached time: 20190516030922
Cache expiry: 2592000
Dynamic content: false
CPU time usage: 0.492 seconds
Real time usage: 0.660 seconds
Preprocessor visited node count: 1805/1000000
Preprocessor generated node count: 0/1500000
Postexpand include size: 65153/2097152 bytes
Template argument size: 772/2097152 bytes
Highest expansion depth: 15/40
Expensive parser function count: 8/500
Unstrip recursion depth: 1/20
Unstrip postexpand size: 70826/5000000 bytes
Number of Wikibase entities loaded: 3/400
Lua time usage: 0.279/10.000 seconds
Lua memory usage: 6.63 MB/50 MB
-->
<!--
Transclusion expansion time report (%,ms,calls,template)
100.00%  548.078      1 -total
 59.50%  326.090      1 Template:Reflist
 38.08%  208.730      5 Template:Cite_journal
 19.94%  109.301      1 Template:More_citations_needed
 18.46%  101.199      4 Template:Ambox
 16.32%   89.451     18 Template:Cite_web
  6.60%   36.163      1 Template:Find_sources_mainspace
  5.52%   30.244      1 Template:Split_section
  4.82%   26.437      1 Template:Split_portions
  2.82%   15.468      1 Template:US-centric
-->
<!-- Saved in parser cache with key enwiki:pcache:idhash:2696619-0!canonical and time
 -->
</div><noscript><img alt="" height="1" src="//en.wikipedia.org/wiki/Special:CentralAu
<div class="printfooter">Retrieved from "<a dir="ltr" href="https://en.wikipedia.org/
<div class="catlinks" data-mw="interface" id="catlinks"><div class="mw-normal-catlink
<div class="visualClear"></div>
</div>
</div>
<div id="mw-navigation">
<h2>Navigation menu</h2>
<div id="mw-head">
```

```
<div aria-labelledby="p-personal-label" id="p-personal" role="navigation">
<h3 id="p-personal-label">Personal tools</h3>
<ul>
<li id="pt-anonuserpage">Not logged in</li><li id="pt-anontalk"><a accesskey="n" hre
</div>
<div id="left-navigation">
<div aria-labelledby="p-namespaces-label" class="vectorTabs" id="p-namespaces" role=
<h3 id="p-namespaces-label">Namespaces</h3>
<ul>
<li class="selected" id="ca-nstab-main"><span><a accesskey="c" href="/wiki/Web_scrap
</div>
<div aria-labelledby="p-variants-label" class="vectorMenu emptyPortlet" id="p-varian
<input aria-labelledby="p-variants-label" class="vectorMenuCheckbox" type="checkbox"
<h3 id="p-variants-label">
<span>Variants</span>
</h3>
<ul class="menu">
</ul>
</div>
</div>
<div id="right-navigation">
<div aria-labelledby="p-views-label" class="vectorTabs" id="p-views" role="navigatio
<h3 id="p-views-label">Views</h3>
<ul>
<li class="collapsible selected" id="ca-view"><span><a href="/wiki/Web_scraping">Rea
</div>
<div aria-labelledby="p-cactions-label" class="vectorMenu emptyPortlet" id="p-cactio
<input aria-labelledby="p-cactions-label" class="vectorMenuCheckbox" type="checkbox"
<h3 id="p-cactions-label"><span>More</span></h3>
<ul class="menu">
</ul>
</div>
<div id="p-search" role="search">
<h3>
<label for="searchInput">Search</label>
</h3>
<form action="/w/index.php" id="searchform">
<div id="simpleSearch">
<input accesskey="f" id="searchInput" name="search" placeholder="Search Wikipedia" t
</form>
</div>
</div>
</div>
<div id="mw-panel">
<div id="p-logo" role="banner"><a class="mw-wiki-logo" href="/wiki/Main_Page" title=
<div aria-labelledby="p-navigation-label" class="portal" id="p-navigation" role="nav
<h3 id="p-navigation-label">Navigation</h3>
<div class="body">
```

```
<ul>
<li id="n-mainpage-description"><a accesskey="z" href="/wiki/Main_Page" title="Visit
</div>
</div>
<div aria-labelledby="p-interaction-label" class="portal" id="p-interaction" role="na
<h3 id="p-interaction-label">Interaction</h3>
<div class="body">
<ul>
<li id="n-help"><a href="/wiki/Help:Contents" title="Guidance on how to use and edit
</div>
</div>
<div aria-labelledby="p-tb-label" class="portal" id="p-tb" role="navigation">
<h3 id="p-tb-label">Tools</h3>
<div class="body">
<ul>
<li id="t-whatlinkshere"><a accesskey="j" href="/wiki/Special:WhatLinksHere/Web_scrap
</div>
</div>
<div aria-labelledby="p-coll-print_export-label" class="portal" id="p-coll-print_expo
<h3 id="p-coll-print_export-label">Print/export</h3>
<div class="body">
<ul>
<li id="coll-create_a_book"><a href="/w/index.php?title=Special:Book&amp;bookcmd=book
</div>
</div>
<div aria-labelledby="p-lang-label" class="portal" id="p-lang" role="navigation">
<h3 id="p-lang-label">Languages</h3>
<div class="body">
<ul>
<li class="interlanguage-link interwiki-ar"><a class="interlanguage-link-target" hre
<div class="after-portlet after-portlet-lang"><span class="wb-langlinks-edit wb-langl
</div>
</div>
</div>
<div id="footer" role="contentinfo">
<ul id="footer-info">
<li id="footer-info-lastmod"> This page was last edited on 15 May 2019, at 09:54<spa
<li id="footer-info-copyright">Text is available under the <a href="//en.wikipedia.o
additional terms may apply.  By using this site, you agree to the <a href="//foundati
</ul>
<ul id="footer-places">
<li id="footer-places-privacy"><a class="extiw" href="https://foundation.wikimedia.o
<li id="footer-places-about"><a href="/wiki/Wikipedia:About" title="Wikipedia:About"
<li id="footer-places-disclaimer"><a href="/wiki/Wikipedia:General_disclaimer" title=
<li id="footer-places-contact"><a href="//en.wikipedia.org/wiki/Wikipedia:Contact_us"
<li id="footer-places-developers"><a href="https://www.mediawiki.org/wiki/Special:My
<li id="footer-places-cookiestatement"><a href="https://foundation.wikimedia.org/wik
<li id="footer-places-mobileview"><a class="noprint stopMobileRedirectToggle" href="/
```

```
</ul>
<ul class="noprint" id="footer-icons">
<li id="footer-copyrightico">
<a href="https://wikimediafoundation.org/"><img alt="Wikimedia Foundation" height="3
<li id="footer-poweredbyico">
<a href="//www.mediawiki.org/"><img alt="Powered by MediaWiki" height="31" src="/sta
</ul>
<div style="clear: both;"></div>
</div>
<script>(window.RLQ=window.RLQ||[]).push(function(){mw.config.set({"wgPageParseReport
<script type="application/ld+json">{"@context":"https:\/\/schema.org","@type":"Artic
<script>(window.RLQ=window.RLQ||[]).push(function(){mw.config.set({"wgBackendResponse
</body>]
```

Podemos selecionar o primeiro parágrafo utilizando o método `find` para localizar uma *tag* **HTML** e o `get_text` para extrair somente o texto:

```
In [71]: print(soup.find('p').get_text())

Web scraping, web harvesting, or web data extraction is data scraping used for extracting data
```

Veja que estamos próximos do nosso objetivo final, basta usar o método `find_all` que conseguiremos econtrat todas as tags **p**'s. Atente que os resultados serão retornados em uma lista:

```
In [73]: for item in soup.find_all('p')[:3]: print(item.get_text())

Web scraping, web harvesting, or web data extraction is data scraping used for extracting data

Web scraping a web page involves fetching it and extracting from it.[1][2] Fetching is the dow

Web scraping is used for contact scraping, and as a component of applications used for web ind
```

Resumindo, poderíamos ter utilizado somente os seguintes comandos:

```
In [74]: import requests
         from bs4 import BeautifulSoup

         page = requests.get("https://en.wikipedia.org/wiki/Web_scraping")
         soup = BeautifulSoup(page.content, 'html.parser')
         for item in soup.find_all('p')[:3]: print(item.get_text())

Web scraping, web harvesting, or web data extraction is data scraping used for extracting data

Web scraping a web page involves fetching it and extracting from it.[1][2] Fetching is the dow

Web scraping is used for contact scraping, and as a component of applications used for web ind
```

Para fixar as ideias veja como seria simples retornar os dois primeiros parágrafo na wikipédia sobre HTML, CSS e JS em três passos: + Lista de *url*'s + Função para retornar o parágrafo + Loop para aplicar a função a cada item na lista

Passo 1:

```
In [105]: paginas = [
              'https://pt.wikipedia.org/wiki/HTML',
              'https://pt.wikipedia.org/wiki/Cascading_Style_Sheets',
              'https://pt.wikipedia.org/wiki/JavaScript'
          ]

          paginas
```

```
Out[105]: ['https://pt.wikipedia.org/wiki/HTML',
           'https://pt.wikipedia.org/wiki/Cascading_Style_Sheets',
           'https://pt.wikipedia.org/wiki/JavaScript']
```

Passo 2:

```
In [106]: def retorna_n_paragrafos(url, n = 1):
              '''
              Essa função retorna os n primeiros parágrafos da url selecionada.
              Possui os seguintes argumentos:
              @url: a url desejada
              @n: número de parágrafos para retornar, default = 1
              '''

              page = requests.get(url)
              soup = BeautifulSoup(page.content, 'html.parser')
              paragrafos = soup.find_all('p')[:n]
              paragrafos_texto = []
              for item in paragrafos: paragrafos_texto.append(item.get_text())

              return '/n'.join(paragrafos_texto) #concatenar elementos e separar por nova linh


          print(retorna_n_paragrafos('https://pt.wikipedia.org/wiki/HTML', n = 2))
```

```
HTML (abreviação para a expressão inglesa HyperText Markup Language, que significa Linguagem d
/nHyTime é um padrão para a representação estruturada de hipermídia e conteúdo baseado em tempo
```

Vemos que a função funciona, o primeiro /n ocorre porque o primeiro parágrafo não tinha conteúdo.

Passo 3 (criando o loop):

```
In [107]: for pagina in paginas:
              out = retorna_n_paragrafos(pagina, n = 2)
              print(pagina + '\n' + out)
```

```
https://pt.wikipedia.org/wiki/HTML
HTML (abreviação para a expressão inglesa HyperText Markup Language, que significa Linguagem de
/nHyTime é um padrão para a representação estruturada de hipermídia e conteúdo baseado em tempo

https://pt.wikipedia.org/wiki/Cascading_Style_Sheets
Cascading Style Sheets (CSS) é um mecanismo para adicionar estilo (cores, fontes, espaçamento,
/nO código CSS pode ser aplicado diretamente nas tags ou ficar contido dentro das tags <style>

https://pt.wikipedia.org/wiki/JavaScript
JavaScript , frequentemente abreviado como JS, é uma linguagem de programação interpretada de a
/nÉ atualmente a principal linguagem para programação client-side em navegadores web. É também
```

```
In [ ]: Fim.
```