UNIVERSITÀ
DEGLI STUDI
DI PALERMO

UNIVERSITÀ
DEGLI STUDI
DI BRESCIA

SCRAPING DATA FROM WEB: EXAMPLES AND APPLICATIONS

Paola Zola

Palermo, May, 20-21, 2019

TODAY CONTENTS:

INTRODUCTION TO WEB SCRAPING AND R APPLICATION.

- What is Web scraping?
- The Web languages.
- Web limitations and common errors.
- R applications.

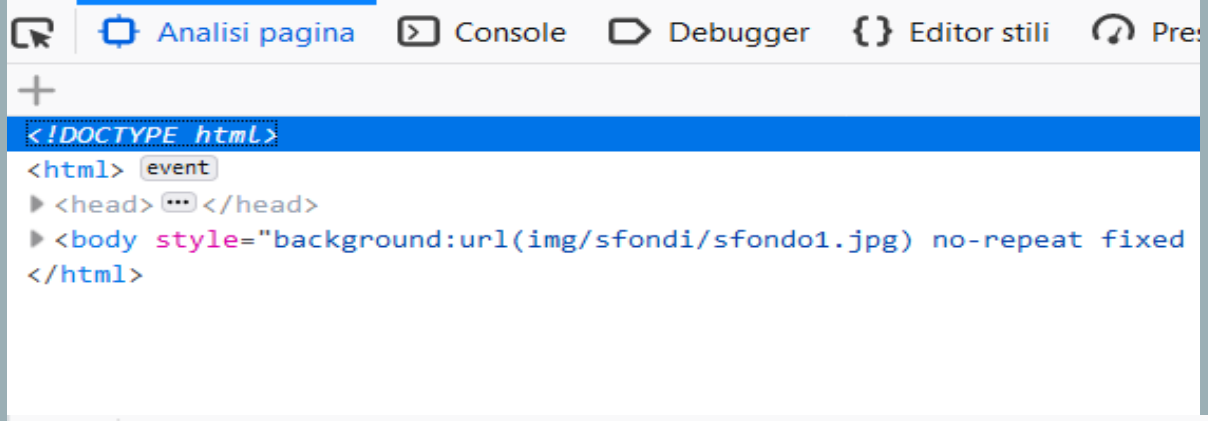
WHAT IS WEB SCRAPING? IS IT NECESSARY?

THE WEB LANGUAGES

1. HTML
2. XML
3. JSON

1. HYPERTEXT MARKUP LANGUAGE (HTML)

- HTML is a language for presenting content on the Web that was first proposed by Tim Berners-Lee (1989).
- It is simple plain text, but its power is its marked up structure. HTML markup allows defining the parts of a document that need to be displayed as headlines, the parts that contain links, etc.
- HTML syntax rules: tags, elements and attributes.


Analisi pagina
Console
Debugger
Editor stili
<!DOCTYPE html>
<html> event
<head> ··· </head>
<body style="background:url(img/sfondi/sfondo1.jpg) no-repeat fixed
</html>


HEAD


Analisi pagina
Console
Debugger
Editor stili
Prestazioni
Memoria
Rete
Archiviazion
<!DOCTYPE html>
<html> event
<head>
<script type="text/javascript" src="//m.addthis.com/live/red_lojson/300lo.json?si=5caa20f72d6fde…fe72d000&skipb
<script async="" src="//cse.google.com/adsense/search/async-ads.js"></script>
<script async="" src="//cse.google.com/adsense/search/async-ads.js"></script>
<script async="" src="//cse.google.com/adsense/search/async-ads.js"></script>
<script type="text/javascript" async="" src="https://www.google.it/cse/cse.js?cx=004601980036908349058:qfba1ybb
<link rel="icon" type="image/png" href="img/favicon.png">
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-15">
<meta name="viewport" content="width=device-width, initial-scale=1">
<!--<meta name="google" content="7056e77bf0fabe1-52ac0d299874837f-g8077de969e087499-15">-->
<title>Comune di Palermo - Sito Istituzionale - Home Page</title>
<link rel="stylesheet" media="screen and (min-width: 320px) and (max-width : 667px)" href="css/max-480px.css">
<link href="css/main.css" rel="stylesheet" type="text/css">
<link href="js/flexslider/flexslider.css" rel="stylesheet" type="text/css">
<link href="css/nivo-slider.css" rel="stylesheet" type="text/css">
<link href="css/default.css" rel="stylesheet" type="text/css">
<link href="js/jquery.cookiebar/jquery.cookiebar.css" rel="stylesheet" type="text/css">
<link rel="stylesheet" href="js/slicknav/slicknav.css">
<link rel="stylesheet" href="js/bxslider/jquery.bxslider.css">
<script async="" src="//www.google-analytics.com/analytics.js"></script>
<script src="js/jquery-1.10.2.min.js"></script>
<script async="" src="https://embed.twitcker.com/ticker/PALERMOPM.js"></script>
<style type="text/css"> ··· </style>

BODY

```
▼<li class>
  ▶<a class="messagesContent" role="button" href="https://www.facebook.com/messages/t/
  100002054230481">…</a>
  </li>
```

```
▶<script>…</script>
▼<article id="article-1-3360037" class="article-base ng-scope" ng-controller="ArticleController" ng-init="init({index: '1-
3360037',
    title:     '\u00ABHa nevicato pi\u00F9 in maggio che durante l\u2019inverno\u00BB',
    occhiello: 'MALTEMPO',
    data:      '5 May, 2019'
  })">
  ▶<div class="container adv-1-3360037 adv-top article-adv pushbar">…</div>
  ▼<div class="container">
    ::before
    ▼<div class="row">
      ::before
      ▼<div class="col-xs-12 col-sm-10 col-sm-offset-1">
        ▼<div class="article-base-head">
          ▼<h4 class="article-base-category"> == $0
            <span class="article-base-half-title text-uppercase">MALTEMPO</span>
          </h4>
          <h1>«Ha nevicato più in maggio che durante l’inverno»</h1>
          <hr>
        </div>
      </div>
```

2. EXTENSIBLE MARKUP LANGUAGE (XML)

- XML, is one of the most popular formats for exchanging data over the Web. Differently from HTML, it is born to store data.
- As HTML format, XML is also a plain text. Thus, it is intuitive to read and compatible with different browser and operating system.
- One limitation of XML is that the plain text XML format is often redundant. In a standard XML, the starting and closing tags are repeated for every entry. This can consume more space in the document than the actual data.

EXAMPLE OF XML:

```xml
<?xml version="1.0" encoding="ISO-8859-1"?>
<bond_movies>
  <movie id="1">
    <name>Dr. No</name>
    <year>1962</year>
    <actors bond="Sean Connery" villain="Joseph Wiseman"/>
    <budget>1.1M</budget>
    <boxoffice>59.5M</boxoffice>
  </movie>
  <movie id="2">
    <name>Live and Let Die</name>
    <year>1973</year>
    <actors bond="Roger Moore" villain="Yaphet Kotto"/>
    <budget>7M</budget>
    <boxoffice>126.4M</boxoffice>
  </movie>
  <movie id="3">
    <name>Skyfall</name>
    <year>2012</year>
    <actors bond="Daniel Craig" villain="Javier Bardem"/>
    <budget>175M</budget>
    <boxoffice>1108.6M</boxoffice>
  </movie>
</bond_movies>
```

3. JAVA SCRIPT OBJECT NOTATION (JSON)

- Similar to XML, JSON has been introduced for the storage and exchange of human readable data. Many APIs by popular web applications provide data in the JSON format.
- Despite it is originally developed in Java language (as the name suggest), it is compatible with many programming languages as R and Python.
- JSON syntax rules: brackets.

EXAMPLE OF JSON


```
ntext': 'http://schema.org',
ype': 'Review',
thor': 'un utente di TripAdvisor',
atePublished': '27 marzo 2019',
 image': 'https://media-cdn.tripadvisor.com/media/photo-s/02/1e/8a/4e/dolci-al-pistacchio.jpg',
 'itemReviewed': {'@type': 'FoodEstablishment',
  'address': {'@type': 'PostalAddress',
   'addressCountry': {'@type': 'Country', 'name': 'Italia'},
   'addressLocality': '',
   'addressRegion': 'Provincia di Palermo',
   'postalCode': '90146',
   'streetAddress': "Viale Alcide De' Gasperi 237 Via Lincoln"},
  'image': 'https://media-cdn.tripadvisor.com/media/photo-s/02/1e/8a/4e/dolci-al-pistacchio.jpg',
  'name': 'Bar Touring'},
 'name': "E per pranzo l\\'arancina bomba",
 'reviewBody': "Con 1,8 \\u20ac risolto il pranzo di lavoro, un bell\\'arancino al prosciutto e via. D\
\'altronde prendere altro sarebbe difficile viste le dimensioni del dispositivo. Locale da frequentare
prima dell\\'arrivo della torma di studenti locali che lo occupano militarmente. ",
 'reviewRating': {'@type': 'Rating', 'ratingValue': '4'},
 'url': '/Restaurant_Review-g187890-d1115284-Reviews-Bar_Touring-
Palermo_Province_of_Palermo_Sicily.html'}
```

XPath	CSS Selector
//span[@class = ‘article-base-half-title text-uppercase’]	html.find_element_by_css_selector(“span.class").get_attribute("article-base-half-title text-uppercase")

TAGS SELECTORS: XPATH VS CSS SELECTOR

XPATH

- It is a query language useful in extracting HTML/XML items.
- It is based on the tree structure present in a HTML/XML document.
- It is based on four major concepts which are: root nodes vs not-root nodes (`/div` vs `//div`), attribute selection (`/@href` vs `//a[@href='http://google.com']`), selection nodes by position (`//a[3]`, `//table[last()]`)

CSS SELECTOR

- The CSS (Cascading Style Sheets) selector is an alternative of the XPath selector.
- Not every packages both in R and Python support the CSS selector.
- CSS selector is based on own sintax that might leads it to be preferable than XPath. For example:
  - ✓ Sub-string matching by the symbols `^`, `$` or `*`:

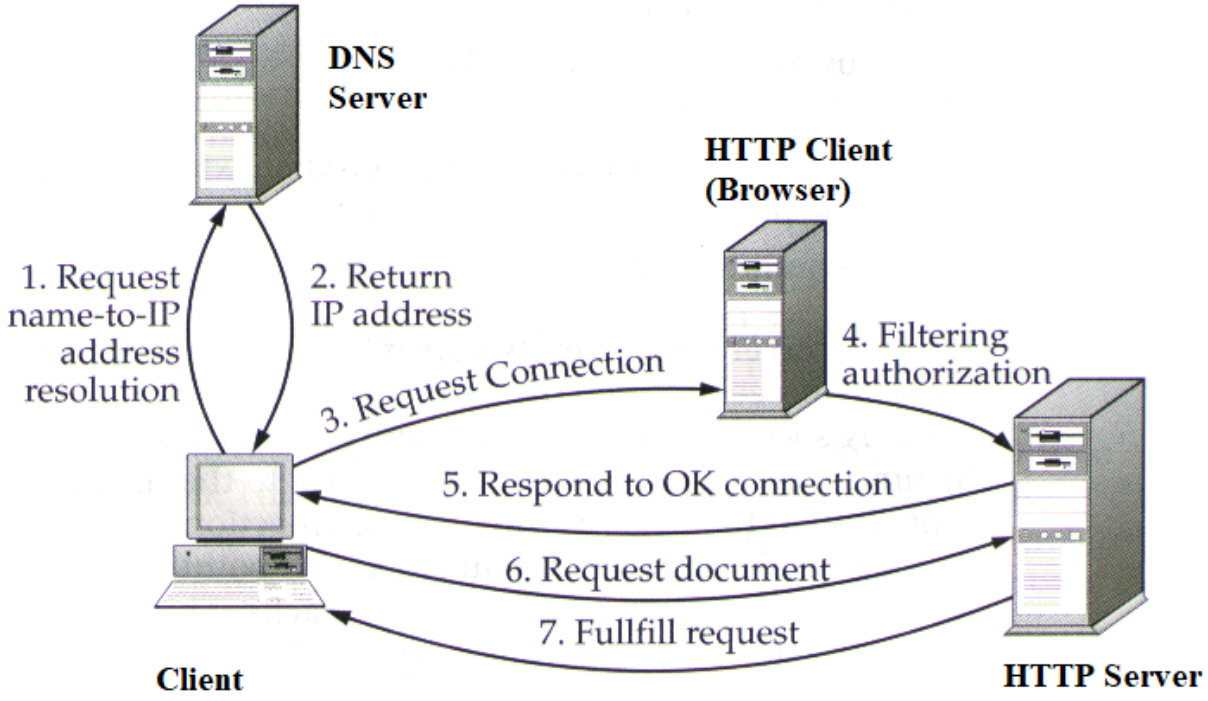
```
a[id^='match_str_starts_with_text'],
a[id$='match_str_ending_with_text'],
a[id*='march_str_containing_text']
```

  - ✓ Matching by inner text: (`a:contains('text')`) will match the item in the HTML that match the desire text no matter where it's located.

HTTP

- Hypertext Tranfer Protocol (HTTP) is the most common protocol for communication between web clients and servers, that is, computers that respond to requests from the network.
- Virtually every HTML page we open, every image we view in a browser, every video we watch is delivered by HTTP.
- The techniques, standards, and protocols that allow to communicate with the Web are called Internet Protocol Suite(IPS). Two of the most prominent players of IPS are TCP (Transmission Control Protocol) and IP (Internet Protocol).

HTTP CONNECTION



ERRORS AND LIMITATIONS

HTTP ERRORS

404
Not Found

504
Gateway Timeout

429
Too Many Requests

403
Forbidden

500
Internal Server Error

408
Request Timeout

SOLUTIONS

- Timeout
- Handle Exceptions
- Use VPN (virtual private network)
- Use HTTP proxies

R APPLICATIONS

DEVTOOLS

- Devtools is a package build by Hadley Wickham and Winston Chang with the aim to simplify programmers life.
- have provided the handy CRAN package devtools (Wickham and Chang 2013) which makes it easy to install R software that is not published on CRAN but on GitHub using the install_github() function.

RCURL

- The `Rcurl` package provides bindings to the libcurl C library for R. It is composed by several functions to help in HTTPs queries. In this seminar we are using the following functions:

1. `getURL()`: it is the basic function to the GET request to retrieve a resource from a web server. `getURL()` is similar to `getBinaryURL()` while
2. `getURLContent()` is a more sophisticated function that tries to identify the type of content in advance by inspecting the Content-Type field in the response header.
3. `getURLhandle()` it handle recursive connections establishing the so-called curl handles. It is possible to specify further information as the `user-agent` and cookies management

XML

- `Htmlparser()`: Parses an XML or HTML file or string containing XML/HTML content, and generates an R structure representing the XML/HTML tree. Use `htmlTreeParse()` when the content is known to be(potentially malformed) HTML.
- `xpathSApply ()`: this functions provide a way to find XML nodes that match a particular criterion. It uses the Xpath syntax and allows very powerful expressions to identify nodes of interest within a document both clearly and efficiently.

RVEST

- `html_table()`: it parses an html table into a data frame,
- `html_nodes()`: easy function to extract pieces of HTML documents using XPath and CSS selectors.

TWITTER

- The package allows to download data from Twitter passing thought the REST API service.

- To access to Twitter data it is necessary to have a Twitter Account and to sign in in the Twitter developer dashboard (https://twitter.com/login?redirect_after_login=https%3A%2F%2Fdeveloper.twitter.com%2Fapps). You need to create your own App to get the credentials : consumer key (API Key), consumer secret (API Secret), Access Token and Access Token Secret. Having the four keys it is possible to create the Oauth for the user authentication and download the data (`setup_twitter_oauth()`).
- `searchTwitter()`: this function collect tweets following the keywords required. The function include several parameters that might be fixed by the users to bound the time period, the geographic range and the language.

REFERENCES:

- Gentry, J., Gentry, M. J., RSQLite, S., & Artistic, R. L. (2016). Package 'twitteR'.
- Lang, D. T., & Lang, M. D. T. (2019). Package 'XML'.
- Lang, D. T., & Lang, M. D. T. (2019). Package 'RCurl'.
- Munzert, S., Rubba, C., Meißner, P., & Nyhuis, D. (2014). Automated data collection with R: A practical guide to web scraping and text mining. John Wiley & Sons.
- Wickham, H., & Wickham, M. H. (2016). Package 'rvest'.