



كيفية تجريف مواقع الانترنت العربية لتجميع البيانات

مع سامر حجازي



ما معنى تجريف مواقع الانترنت؟

استخراج المعلومات من صفحات الانترنت بشكل برمجي

Also known as Web Scraping

ما هي الاسباب لتجريف الانترنت؟

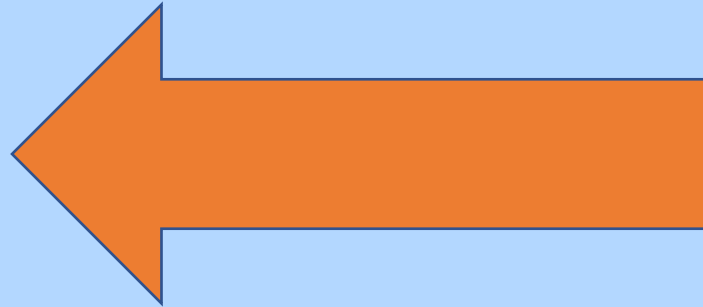
1. الحصول على بيانات كثيرة في وقت قصير

2. تحويل البيانات من شكل غير منظم الى اعمدة وصفوف منظمة

3. سهولة تطبيقه على مواقع مختلفة

كيف تقوم بالتجريف؟

HTML



مفتاح التجريف

جميع الصفحات على الانترنت مكونة من html

The screenshot shows a web browser displaying a real estate website. The URL in the address bar is `om.opensooq.com/ar/عقارات-للإيجار/شقق-للإيجار?page=114`. The page is in Arabic and features two real estate listings. The first listing is for a "Brand new 4rooms flat for rent East of Sultan Qaboos Mosque in Ruwi" priced at 200 ريال. The second listing is for "الإيجار شقق في المعيلة" priced at 150 ريال. Both listings include a profile picture, a date (12-08-2022), and a location (مسقط). The browser's developer tools are open on the right, showing the "Elements" panel with the HTML structure of the page. The HTML includes a DOCTYPE declaration, an `<html>` tag with various attributes, and a `<body>` tag with a class of "post-listing om". The `<body>` tag contains several nested divs, including a "fb-root" div, a "header" div, a "nav" div, and a "page" div. The "page" div contains a "breadcrumbs" ul and a "mainContent" div. The "mainContent" div contains several flash messages and a "div.rectLiDetails.tableCell.vMiddle.p8" element. The "Styles" panel shows the default style for the "element.style" and a specific style for ".rectLiDetails>h2>a" from "os-Listing_64216.css:1".

om.opensooq.com/ar/عقارات-للإيجار/شقق-للإيجار?page=114

الوسيط الظاهرة

ع سعر مميعة إعلانات المناجر فقط إعلانات العضوية المميزة

التاريخ

عرض الإعلانات على

كن أول من يعلم عن الإعلان شقق للإيجار أعلمني

200 ريال Brand new 4rooms flat for rent East of Sultan Qaboos Mosque in Ruwi

مسقط | روي | 12-08-2022

شقق للإيجار | 4 غرف نوم | حمامين | غير مفروشة

متصل

درش اتصل أصف الى المفضلة

150 ريال الإيجار شقق في المعيلة

مسقط | المعيلة | 12-08-2022

شقق للإيجار | غرفة نوم | 3 حمامات | غير مفروشة

متصل

درش اتصل أصف الى المفضلة

```
<!DOCTYPE html>
<html class="htmlTopBrd gradient" xmlns="http://www.w3.org/1999/xhtml" xml:lang="ar"
xmlns:fb="http://ogp.me/ns/fb#" prefix="og: http://ogp.me/ns#" lang="ar" dir="rtl">
  <head>...</head>
  <body id="body" class="post-listing om" style>
    <div id="fb-root" class="fb_reset">...</div>
    <!--promotion-->
    <!--end promotion-->
    <!--header-->
    <div id="header" class>...</div>
    <!--end header-->
    <!--nav-bar-->
    <div id="nav">...</div>
    <!--end nav-bar-->
    <div id="page">
      <ul class="breadcrumbs">...</ul>
      <!--main content-->
      <div id="mainContent">
        <div class="mb15 flashMsg warning hide chat-disconnected-flash" style="z-index: 9">...
        </div>
        <div class="mb15 flashMsg warning hide chat-not-ready" style="z-index: 9">...</div>
        <div class="mb15 flashMsg deleteChat success hide" style="z-index: 9">...</div>
      </div>
    </div>
  </body>
</html>
```

div.rectLiDetails.tableCell.vMiddle.p8 h2.fRight.mb15 a.block.postLink.notEg.postSpanTitle.noEmojiText

Styles Computed Layout Event Listeners DOM Breakpoints Properties Accessibility

Filter :hov .cls +

```
element.style {
}

.rectLiDetails>h2>a {
  os-Listing_64216.css:1
```

تعرف على علامة الhtml

العلامة

HTML tag

{ }

نوع العلامة

HTML tag type

صفة العلامة

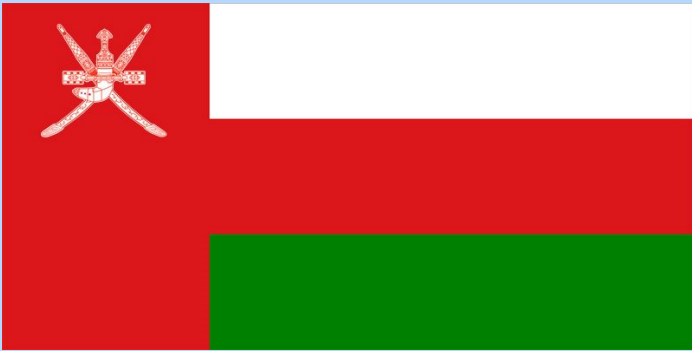
HTML attribute type

قيمة الصفة

HTML attribute value

موضوع الورشة

استخراج معلومات من موقع السوق المفتوح عن الشقق
المتوفرة للايجار في سلطنة عمان



السوق المفتوح
opensooq.com

