

Realestate.com.au sold spider

Deliverable:

The spider should run weekly on the same day each week and create a file with the following structure:

Output: A CSV file with the following structure:

DATE_SOLD, PRICE, URL

File name: re_sold_yyyymmdd.csv (where yyyymmdd is a date when the spider finished)

Reliability: The website might implement anti-bot defence counter-measures, meaning that one of the requirements is to have an infrastructure with multiple IP addresses (to avoid banning).

Operation: Schedule is weekly with an option to run it manually and to stop it.

Task:

Starting point <https://www.realestate.com.au/sold/>

In the search field input a suburb from the seed list (a CSV file provided by us)

the “Surrounding suburbs” check box should be unchecked

Clicking the search button will open the page looking like this for each suburb

<https://www.realestate.com.au/sold/in-carlton,+vic+3053%3b/list-1?activeSort=solddate&includeSurrounding=false>

On the top of the page there's information about suburb name and number of search results.

“Sold Real Estate & Property in Carlton, VIC 3053

1-25 of 2706 results”

We need to extract information from the first page only from all 25 items (or less)

The required information is

- 1) Sold date: in this example 18 Jun 2019
- 2) Price: \$710000
- 3) URL: A url that will open the detailed information about the sold property, in this example /sold/property-apartment-vic-carlton-131283490. (It is not necessary to actually open it, we just need the URL)

Sold Real Estate & Property in Carlton, VIC 3053

1-25 of 2706 results

List

Map

Sort by:

Date (Newest - Oldest)



JellisCraig

Charles Atkins



\$710,000



29/100 Queensberry Street, Carlton

2 2 1 | Apartment

Sold on 18 Jun 2019

international
equities

Agent: Daniel Lin

Seed list file sample:

State,Pcode,Locality

NSW,2540,JERVIS BAY

ACT,2600,BARTON

WA,6728,ST GEORGE RANGES

Please note that the website expects search field in the following format

<locality>, <state> <postcode>;