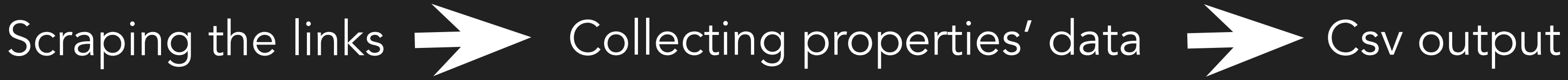
Challenge Collecting Data
-
Web Scraping

Minh Hien Vo, Michel Ombessa, Jose Roldan and Logan Vendrix

How we did it

1. Inspection
2. Using Selenium, BeautifulSoup and threading to get the links of properties
3. Using requests and PyQuery to collect the properties' data
4. Creating a csv file with the results
5. Issues



Scraping the links

Collecting properties’ data

Csv output

Inspection

- Captcha on Zimmo ⇒ too hard to scrape
- Cookie on Immoweb ⇒ manageable
- Inspection of the HTML to understand the structure of the site

⇒ LET’S GO!

Getting the links

- Selenium to bypass the cookie click button (interaction)
- BeautifulSoup to find the ‘a’ and ‘href’ of all the properties’ links
- “For loop” to go through all the pages (replace() func)
- Thread class to speed up the process of data collection
- Write results to a .txt file (+/- 33.000 links)

Getting the properties’ data

- PyQuery to get the desired features from the HTML
- Check the differences between pages of different types of properties

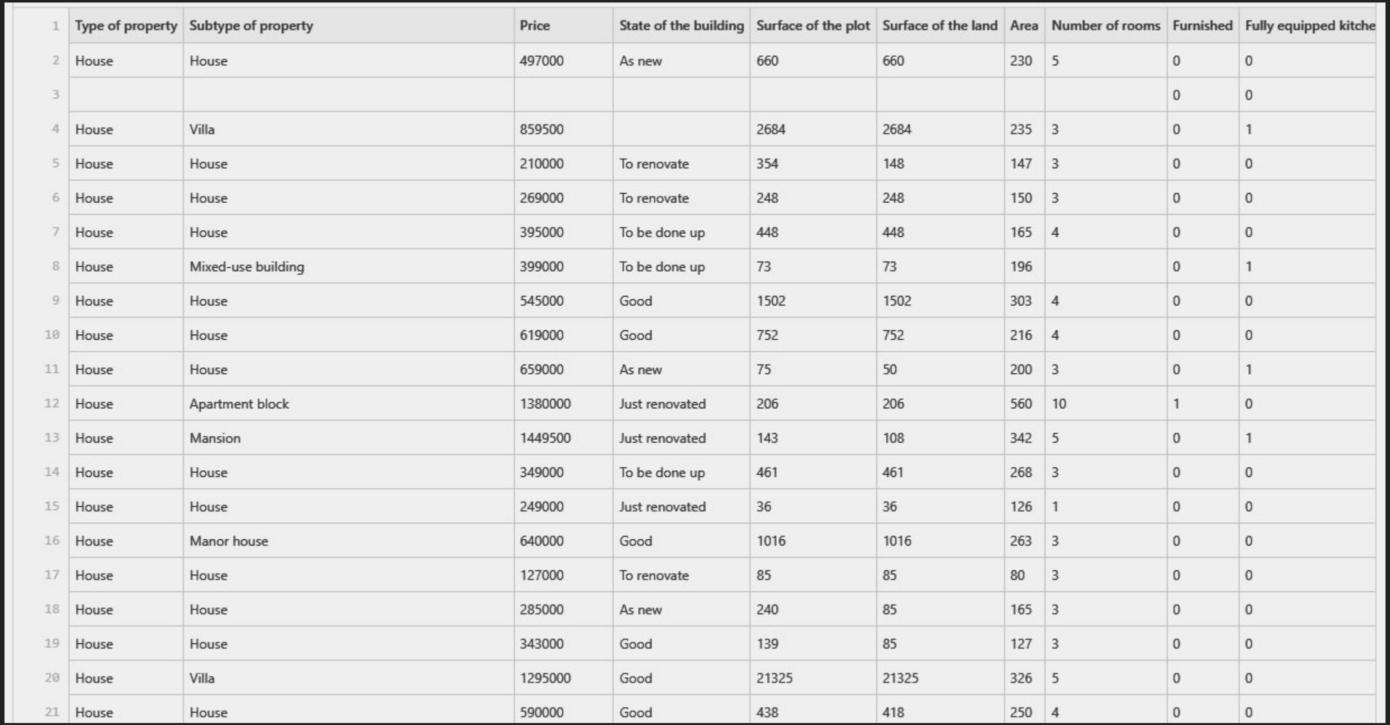
  ⇒ Price, swimming-pool, ...

- Formatting the data (‘Yes’ = 1 , ‘No’ = 2, ...)
- Cleaning the data (names, …)

Result csv

- “For loop” to go through all the links of the txt file
- Dict.writer to create a row for each property
- Sample of 5.001 results (not enough time)

Result csv



Type of property	Subtype of property	Price	State of the building	Surface of the plot	Surface of the land	Area	Number of rooms	Furnished	Fully equipped kitche
House	House	497000	As new	660	660	230	5	0	0
								0	0
House	Villa	859500		2684	2684	235	3	0	1
House	House	210000	To renovate	354	148	147	3	0	0
House	House	269000	To renovate	248	248	150	3	0	0
House	House	395000	To be done up	448	448	165	4	0	0
House	Mixed-use building	399000	To be done up	73	73	196		0	1
House	House	545000	Good	1502	1502	303	4	0	0
House	House	619000	Good	752	752	216	4	0	0
House	House	659000	As new	75	50	200	3	0	1
House	Apartment block	1380000	Just renovated	206	206	560	10	1	0
House	Mansion	1449500	Just renovated	143	108	342	5	0	1
House	House	349000	To be done up	461	461	268	3	0	0
House	House	249000	Just renovated	36	36	126	1	0	0
House	Manor house	640000	Good	1016	1016	263	3	0	0
House	House	127000	To renovate	85	85	80	3	0	0
House	House	285000	As new	240	85	165	3	0	0
House	House	343000	Good	139	85	127	3	0	0
House	Villa	1295000	Good	21325	21325	326	5	0	0
House	House	590000	Good	438	418	250	4	0	0

Issues

- Who does what?
- Problem with requests so Selenium/BeautifulSoup instead
- Problem with timeout ⇒ time.sleep()
- Code integration
- Hard to locate all desired features
- Multi-threading not used ⇒ time lost
- Empty lines in csv
- “None” values ⇒ not showing in csv
- Multi-contribution in git repo