

# Machine Learning For a House pricing Prediction Web Application

ALEX SOUDANT

Ynov Ingésup M1,  
20 Boulevard Général de Gaulle, 44200 Nantes

Correspondence: Alex Soudant. E-mail: [alex.soudant@ynov.com](mailto:alex.soudant@ynov.com)

## **report on the 30th of January 2017**

Our planned objectives for Week 1 were "Tutorial to scrapy - finding the websites to scrape".

In that purpose, I first followed the "first steps" and "basic concepts" sections of the scrapy website. When finished, I wanted to apply this knowledge to a target website containing information about real estates like prices, size and frontside pictures. Therefore, I looked for adequate websites that led to a selection of five real estate letting agencies :

1. Century21.fr ;
2. pap.fr (particulier a particulier) ;
3. seloger.com ;
4. paruvendu.fr ;
5. explorimmo.com.

I selected these websites for their similarity in terms of webpage organisation which could permit to only tweak a little the scrapy code developed to scrape the information from these different websites.

However, by trying basic procedures on the seloger.com website I noticed that some of these companies have disallowed scrapy in the robot.txt file. After, looking how to adress this inconveniency, I found out that some options in the setup.py file of scrapy can change my signature as a user agent and remove the automatic reading of robots.txt when attempting to scrape a website. Still, this procedure can be seen as bad practice as I am intentionally trying to use a disallowed scraper. Another alternative could be to use a javascript scraper that will log to websites directly through the navigator and therefore is not referenced in the robots.txt file. I found out that PhantomJS could be used as such and could provide me the wanted scraping tools I need. However, after trying a few times the tutorial code to achieve data scrapping I failed to make it work properly. There could be some versioning problems with recent updates that prevent me from learning directly

with web available pre-generated example code. I also had the time to take a glimpse to selenium/webdriverio which looks like a promising framework to scrape from website whith only one dependency to Node.js.

As I could still train our scraping skills with scrapy, we then used it on the pap.fr website that does not disallow scrapy in robots.txt. I found out that with only a few lines of code I could extract the wanted information from this website main page. However, I still have to implement navigation between pages and ultimatly to be able to visit individual housing sale page to obtain the full description of the associated real estate. Finally I have to find a way to save not only the reference link to the picture but the picture itself when scraping.

To conclude, I feel that I explored quite a few directions to take for next week. I can advance on scrapy code if I still can put to work the phantomJS or webdriverio. I do not see any delays that could prevent me from starting the collection of data. My objective is therefore to collect information and pictures of at least a hundred to a thousand real estates which can then be used to run computer vision algorithms after data cleansing and image reformatting.