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

report on the 30th of January 2017

Our planned objectives for Week 1 were "Tutorial to scrappy - finding the websites to scrap".

In that purpose, we first followed the "first steps" and "basic concepts" sections of the scrapy website. When finished, we wanted to apply this knowledge to a target website containing information about real estates like prices, size and frontside pictures. Therefore, we 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.

We selected these websites for their similarity in terms of webpage organisation which could permit to only tweak a little the scrapy code developed to scrap the information from these different websites.
However, by trying basic procedures on the seloger.com website we noticed that some of these compagnies have disallowed scrapy in the robot.txt file. After, looking how to adress this inconveniency, we found out that some options in the setup.py file of scrapy can change our signature as a user agent and remove the automatic reading of robots.txt when attempting to scrap a website. Still, this procedure can be seen as bad practice as we are intentionally trying to use a disallowed scrapper. Another alternative could be to use a javascript scrapper that will log to websites directly through the navigator and therefore is not referenced in the robots.txt file. We found out that PhantomJS could be used as such and could provide us the wanted scapping tools we need. However, after trying a few times the tutorial code to achieve data scrapping we failed to make it work properly. There could be some versioning problems with recent updates that prevent prevent us

from learning directly with web available pre-generated example code. We also had the time to take a glimpse to selenium/webdriverio which looks like a promising framework to scrap from website whith only one dependency to Node.js.
As we 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. We found out that with only a few lines of code we could extract the wanted information from this website main page. However, we 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 we have to find a way to save not only the reference link to the picture but the picture itself when scraping.
To conclude, we feel that we explored quite a few directions to take for next week. We can advance on scrapy code if we still can put to work the phantomJS or webdriverio. We do not see any delays that could prevent us from starting the collection of data. Our 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.