Requirements for Data mining/scraping software

Author : Juan Agustín Tibaldo
UPWORK profile : https://www.upwork.com/freelancers/~01d2be0f4fbb85d044
MAIL : juan.agustin.tibaldo@gmail.com
SKYPE: agustintibaldo

Requested by: Simon Osborne

version : 1

Software which will scrape multiple data from website listings and export into a database where search listings based on specific criteria can be made.

requirement given:

1. From every UK property listing (updated daily to get new daily listings) in Rightmove.co.uk and Zoopla.co.uk, I need:

a. Type of property (Detached, terraced, flat/apartment etc)
b. No. Bedrooms
c. Postcode (this is accessible via source code)
d. Estate Agent (Name and postcode)
e. Price
f. Date of listing

2. This data needs to be stored in a database/software (please recommend) to do the following analysis:

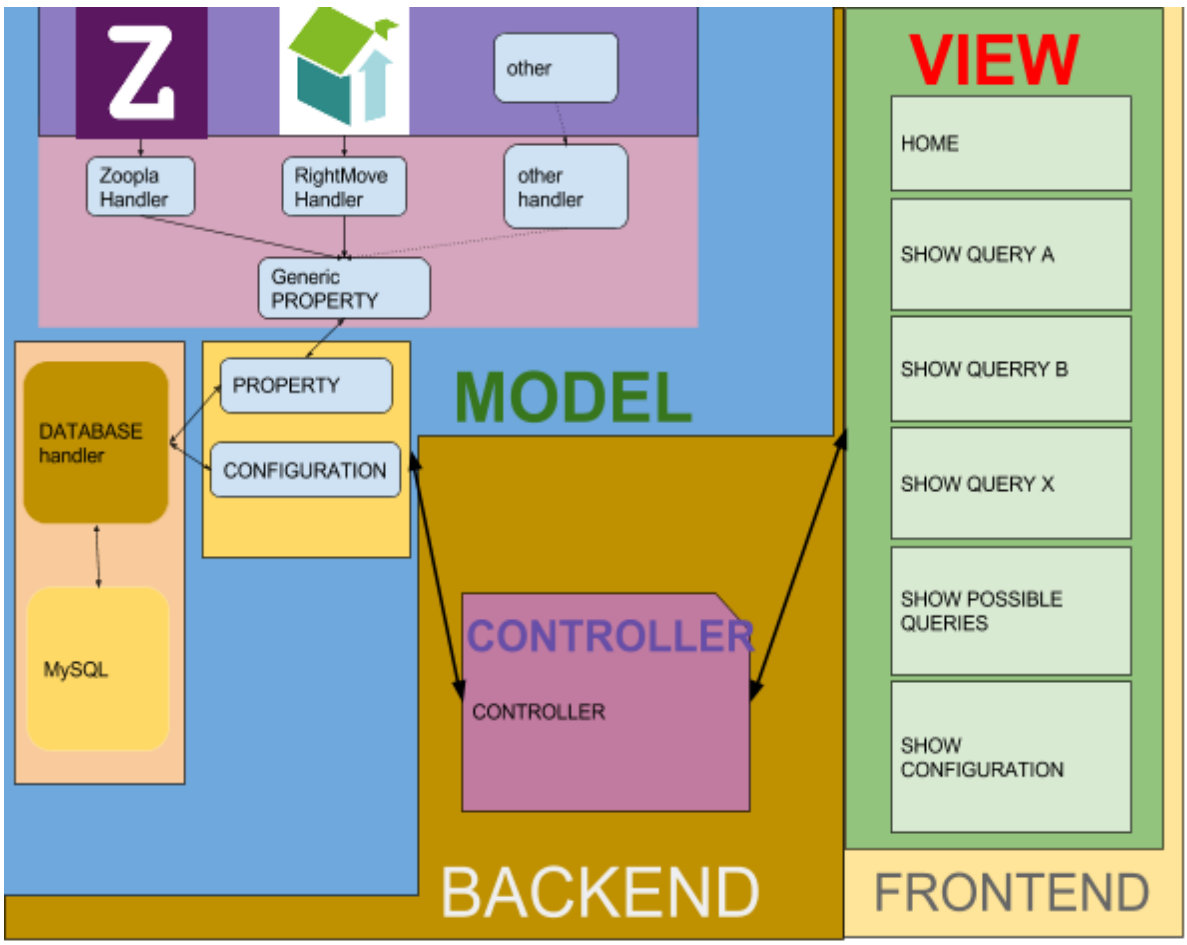
a. Using any postcode, search for all properties listed within 'x' time (eg. 6 months) and within 'y' distance (eg. 1 mile) from that postcode. Each result must also contain all associated data such as type of property, bedrooms, price, estate agent etc.
b. Then I need to be able to run a filter – Example: 'detached' home, '£500,000-£750,000' price, '4' bedrooms.

extras:
-strip out duplicates based on location/type/agent from the database.
-keep duplicates by agents in the database.

I propose the following implementation:

Overview :

The application will be a backend-frontend application. Backend will run locally on background whereas frontend will be accessible through browser. Backend will handle the application logic whereas the frontend it will be everything you will be able to see and interact with (like in any other web application). At time I am using a design pattern called MVC. MVC stands for MODEL, VIEW, CONTROLLER. This MVC scheme is quite useful to organize se production and to allow the software to be modified and extended easily.



BACKEND : model + controller.

FRONTEND : view.

Specification of each MVC layers

Model:

Property

-Type of property ID (to identify the type into the other table)
-No. Bedrooms
-Postcode
-Estate Agent ID (to identify the agent into the other table)
-Price
-Date of listing

Estate Agent

-ID
-Name
-Postcode

Type of property

-ID
-TYPE NAME

Configuration:

-key (for example what time during the day there is need to look up for new information on the websites)
-value (the specific time)

Controller:

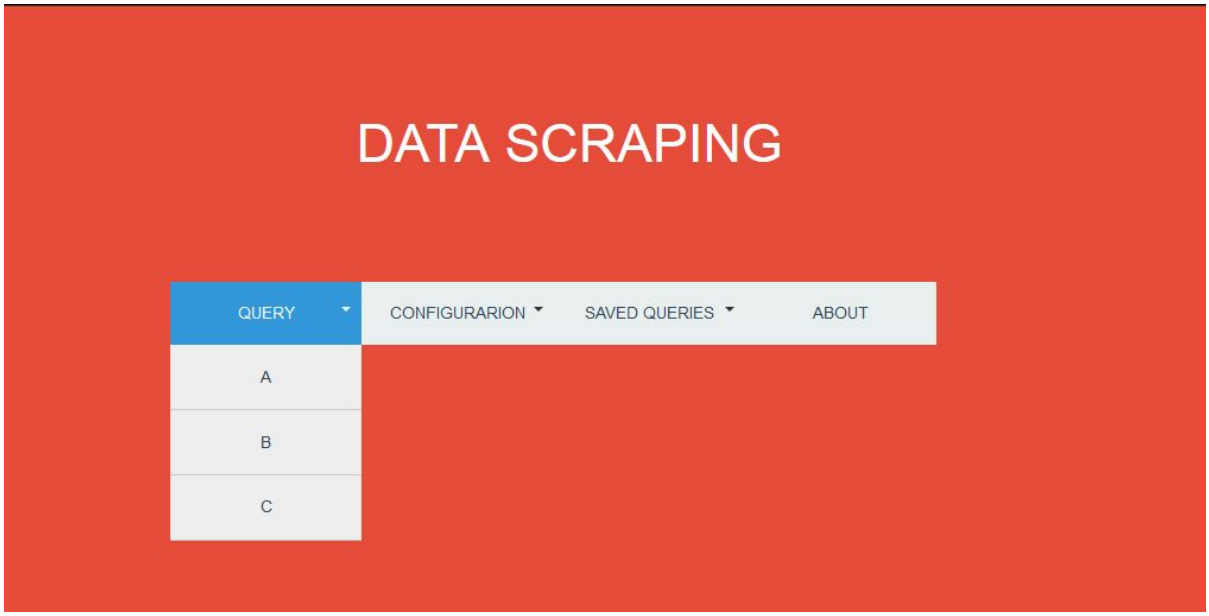
-Change configuration(key,value)
-Add configuration (key,value) (basicly there will be the chance to add new times to look for information in the web , for example “searchZoopla, 3:00”, “searchAll:2:00” and so on.
-Delete configuration (key)
-Get query Result A (postcode, xTime, yDistance)
-Get query Result B (list of types, list of number of bedrooms, price range)

View:

-Main windows

-Show current configuration

-Show available queries which their parameters. (queryA and queryB)



Implementation :

The implementation will be written in javascript both backend and frontend. Front end will be a HTML + CSS + JS page. Communication between the server and the frontend will be carried out by common HTTP request (GET, POST). MySQL database will be used.
The application will be highly modular. Each functionality will be written into their own .js file.

Release:

A prototype with full functionality will be out a week after the contracts is accepted by the two sides. Modification can be done in that stage.

Release:

Full price will be 450 USD.

Future implementation/extension:

This are not included in this release, but I think the software could be extended with on of the followings:

-Saved queries : you will be able to save previous made queries under a names you want. Later you will be able to use this query over the actual data.

-Add new websites to scap.

-Add new databases.

-Add new queries.

-Statistics over the data obtained

-Print sheets, queries results, charts, graphics.

-more.