The Introduction of the Coding Part

In this document, I want to give an introduction to my coding work including the list of the dependent libraries and tools, the content of each file as well as the instructions to run the code.

list of the dependent libraries and tools

Python3 : a programming language I used for most of the coding work.

    + Scrapy : an open source and collaborative framework for extracting the data you need from websites based on Python.

    + Numpy : a fundamental package for scientific computing with Python.

    + Pandas : a data structures and data analysis tools for the Python.

    + matplotlib : a 2D plotting library for Python.

    + scikit-learn : a machine learning packages for Python.

Jupyter Notebook : a web application allows people to create and share documents that contain live code, equations, visualizations and narrative text.

Node.js : a JavaScript runtime built on Chrome's V8 JavaScript engine.

    + Expree : a web application framework for Node.js

Directory Structure

Root Directory

+ crawler (dir) : all the related file of the crawler.

    - scrapy.cfg : the overall configure file of Scrapy

    - spider (dir)

        __init__.py : /

        middlewares.py: /

        pipeline.py: /

        settings.py : configure the crawling speeding, cookie, user_agent and request_headers.

        items/CrazyItem.py : the definition of the data which defines all the features contained in the data item.

        spiders/CrazySpider.py : the spider used to crawl the whole Nottingham house data.

        spiders/SingleSpider.py : the spider used to crawl specific house (providing house id)

+ web (dir) : all related files of web application.

    - node_modules : the modules used during the project

    - app.js : the main program of the web application

    - static (dir) : static HTML, CSS and Javascript

        assets/result/meta_model : the model trained by the data, it will use for online prediction

        assets/result/single_predict.py : the Python script for the online prediction

        assets/result/result.json : the prediction result for the test data, it will be used for the visualization.

+ data (dir) : all the data used in this project

    - house.jl : the intrinsic, area and description data

    - images : all the images that will be used for training. They are named following the rule : house_id.price.0.jpg

Instructions to run the code

Before you start running the code below, please make sure all the above tools have been successfully installed.
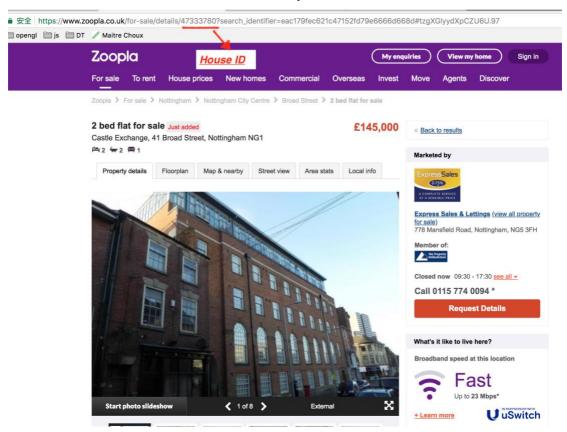
Crawler:

To start the crawler for the all the houses in Nottingham:

1. Change directory to crawler
2. Run the following commond:

   scrapy crawl crazy_spider -o *the_dirctory_name/file_name_you_want.json*

To start the crawler for the one house

1. Change directory to crawler
2. Run the following command:

   scrapy crawl single_spider -a house_id=*the_house_id_of_the_house*

   -o *the_dirctory_name/file_name_you_want.json*

Where **house_id** can be fetched in this way:

Framework:

1. In the root directory
2. Run the following command:

   Jupyter notebook
3. Or you can open the file *HousePrediction_Part1.html,
   and HousePrediction_Part2.html in your browser to check
   the static file of my work.*

Web Server:

1. Change directory web
2. Run the following command:

   node app.js
3. Open a web browser and try to visit *http://localhost:3000/*