Очень важно для проведения такого рода анализов осуществить доскональное изучение доступных источников данных, провести полноценный анализ сферы коммерческой недвижимости и принципов оценки объектов. []

Для анализа рынка и построения в дальнейшем модели – необходимо собрать основную базу данных недвижимости. В организации, как правило, используется собственная база данных, однако в рамках конфиденциальности собственности компании возникает необходимость использовать данные с открытых источников, тем не менее, в самой компании регулярно практикуется сбор и использование открытых источников, в особенности для обновления базы данных и понимания функционирования рынка.

На нынешнем рынке есть несколько крупных открытых баз недвижимости. В первую очередь это такие сайты как «Авито», «Циан», «Яндекс Недвижимость». Однако есть и другие, достаточно подробные информационные источники, однако они принадлежат определенным агентствам недвижимости, то есть подразумевают узконаправленность имеющихся данных и не предоставляют необходимый для данного рода работы объем и спектр объектов.

Из указанных источников необходимо выбрать один основной, соответствующий основным параметрам:

* Полнота данных – наличие максимума информации об объекте;
* Целостность и корректность данных;
* Отсутствие ложных данных;
* Наличие критически важных параметров: адрес, квадратура, стоимость.

Стоит рассмотреть конкретные примеры объектов на данных площадках. В таком случае будет возможность изучить возможный набор данных, который можно будет собрать с каждого объекта.

На данном рисунке видно пример объекта с сайта «Яндекс Недвижимость». Если изучить код страницы, а также визуальную составляющую – со страницы объявления можно извлечь следующую информацию:

* URL;
* Стоимость;
* Площадь;
* Стоимость за квадратный метр;
* Адрес;
* Этаж;
* Назначение (офис, склад, торговое);
* Фиктивные переменные, т.е. подробная информация;
* Агенство / Собственник;
* Расстояние до метро в минутах.

Далее рассмотрим пример страницы объявления Авито:

* Стоимость;
* Площадь;
* Стоимость за квадратный метр;
* Фиктивные переменные, т.е. подробная информация;
* Адрес;
* Этаж;
* Агенство / Собственник;

Как последний подробно-изученный источник для будущей базы данных – рассмотрим Циан.

* URL;
* Стоимость;
* Площадь;
* Стоимость за кв метр;
* Дом;
* Этаж;
* Назначение (офис, склад, торговое);
* Тип здания;
* Фиктивные переменные;
* Агенство;
* Расстояние до метро в минутах;

С учетом проведенного анализа, включая обзор корректности данных – было решено, что основным источником для данных стоит использовать Циан. Яндекс и Авито не так подробно изучают объявления и оставляют большое количество ложных объявлений, которые размещают некоторые агентства недвижимости для привлечения клиентов.

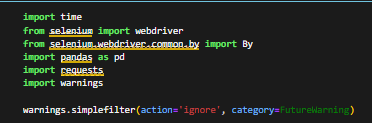
ААААААААААААААА

Для сбора данной информации придется обратиться к написанию программы-парсера. Парсер — это программа, сервис или скрипт, который собирает данные с указанных веб-ресурсов, анализирует их и выдает в нужном формате. Использование парсеров очень распространено в современном мире, во многом это один из основных способов получения информации из открытого доступа. Более того, в сфере недвижимости это фундамент для изучения рынка. Все современные агентства стараются постоянно следить за новыми данными и собирать максимум из них. Такой задачей и занимается один из отделов компании «Актив-Инвест».

После изучения основных источников на тему парсеров, их типов, принципов использования и основных библиотек на языке программирования python возникла необходимость также и в изучении основ написания сайтов – html. Проконсультировавшись со специалистами – была проделана работа по написанию парсера специально для данного сайта, с целью собрать максимум возможной информации.

Для данной задачи были изучены основные библиотеки для парсинга на языке программирования python – “Beautifull4Soup” и “Selenium”. Используя основные API этих библиотек были изучены возможности для сбора данных. Самым оптимальным и эффективным способом стоит считать использование библиотеки “Selenium”, во многом благодаря возможностям данной библиотеки взаимодействовать с сайтами, используя имитацию нажатию кнопок, открытия ссылок и т.д.

Для того, чтобы собрать все объекты, первоначально необходимо собрать все ссылки на них с предварительно настроенной фильтрацией сайта. Настроим среду для кода.



Вот итоговый код для парсинга ссылок всех объектов в сфере коммерческой недвижимости в Санкт-Петербурге и Ленинградской области.

driver **=** webdriver.Chrome()

hrefs **=** []

**try**:

**for** page **in** range(1,50):

        driver.get(url**=f**'https://spb.cian.ru/cat.php?deal\_type=sale&engine\_version=2&offer\_type=offices&office\_type%5B0%5D=2&office\_type%5B1%5D=4&office\_type%5B2%5D=5&office\_type%5B3%5D=7&office\_type%5B4%5D=9&office\_type%5B5%5D=12&p={page}&region=2&sort=creation\_date\_desc')

        time.sleep(10)

        print('main base loaded')

*#   Постепенная прогрузка страницы*

        skrol **=** 0

**for** i **in** range(15):

            skrol **+=** 1000

            driver.execute\_script(**f**"window.scrollTo(0, {skrol});")

            time.sleep(1)

        time.sleep(10)

        print(**f**'page {page} loaded')

*#   Взятие href с каждой страницы*

**for** i **in** range(1, 34):

**try**:

                        item **=** driver.find\_element(By.XPATH, **f**'/html/body/div[1]/div/div[5]/div[{i}]/div/div[2]/div/div[1]/div[1]/div[1]/h3/div/a').get\_attribute('href')

                        hrefs.append(item)

                        time.sleep(1)

**except** Exception **as** inner\_ex:

                        print(**f**'Error in inner loop: {inner\_ex}')

**continue**

**except** Exception **as** ex:

    print(**f**'Error in outer loop: {ex}')

driver.close()

driver.quit()

Данный код включает в себя цикл с искусственными задержками, для прохождения по каждому из объявлений на всех 50 страницах сайта с определенным нами заранее условиями фильтрации. В цикле предусмотрено взятие href ссылок, т.е. url-ссылок опубликованных объявлений, учитывая структура кода сайта «Циан». Также было предусмотрено возникновение ошибок, почему после возникновения ошибки сам процесс парсинга не прекращается. Каждая ссылка объявления, при этом, записывается в список и затем сохраняется.