Задание на ПР-3

Цель: выработка умений, необходимых для самостоятельного извлечения данных из веб-ресурсов.

Для успешного выполнения работы необходимы первичные знания requests, BeautifulSoup, возможно, pandas.

Последовательность выполнения:

1. Выберите на каком-нибудь сайте объявлений или в интернет-магазине интересующую вас рубрику самого нижнего уровня (например, Для дома и дачи – Мебель и интерьер – Кровати, диваны и кресла – Диваны).

2. Убедитесь в том, что сайт учел в выдаче ваше местоположение (если нет – укажите его).

3. Обратите внимание на адрес полученной вами страницы и на то, как он изменяется при переходе к следующей странице выдачи (где и как указывается номер страницы), зафиксируйте это.

4. Откройте код страницы выдачи (F12 или правой кнопкой мыши).

Выберите ключевые элементы блока записи об одном объявлении так, чтобы вы могли выделить этот блок из всего текста: например, это может быть определенный тег, с определенным набором атрибутов.

5. Напишите скрипт, который перебирает все страницы выдачи, и собирает в список следующие параметры: id объявления, название, цена, срок публикации.

6. Проверьте полученный список на корректность – в нем не должно быть посторонних записей и выбросов (слишком малых, слишком больших и отсутствующих цен). При необходимости, устраните ненужные записи программно.

7. Отфильтруйте полученный список по заданному вами критерию: например, оставьте только цены между заданными вами значениями, или оставьте только объявления, находящие в просмотре не больше суток.

8. Сохраните результат вашей работы так, чтобы он был пригоден для машинного анализа.

9. Сделайте вывод о работе, указав встретившиеся сложности и то, как они были преодолены. Оформите отчет в jupyter notebooks.