# Домашнее задание 7

1. Выберите веб-сайт, который содержит информацию, представляющую интерес для извлечения данных. Это может быть новостной сайт, платформа для электронной коммерции или любой другой сайт, который позволяет осуществлять скрейпинг (убедитесь в соблюдении условий обслуживания сайта).

Попытки выбрать маркетплейсы не увенчалась успехом – сервисы блокирут скрейпинг.

Выберем сайт ТВ-программы <https://tv.starhit.ru/chelyabinsk>.

1. Используя Selenium, напишите сценарий для автоматизации процесса перехода на нужную страницу сайта.

Точкой входа выбираем <https://tv.starhit.ru/chelyabinsk>. Как оказалось, для наших целей достаточно одной первой страницы. Надо только подтвердить во всплывающем окне свой город, и организовать проявление всех каналов на странице, а также раскрытие всего списка передач канала.

1. Определите элементы HTML, содержащие информацию, которую вы хотите извлечь (например, заголовки статей, названия продуктов, цены и т.д.).

Будем извлекать названия каналов, время начала передачи, название передачи.

1. Используйте BeautifulSoup для парсинга содержимого HTML и извлечения нужной информации из идентифицированных элементов.

Ок. Так и поступил! ☺

1. Обработайте любые ошибки или исключения, которые могут возникнуть в процессе скрейпинга.

Для проявления всех каналов пришлось организовать прокрутку вниз. Для раскрытия всего списка передач первых двух каналов пришлось организовать прокрутку сайта вверх. У некоторых каналов кнопки раскрытия всего списка нет – пришлось использовать конструкцию «try: - except:». Для записи «русского» текста в файл JSON пришлось использовать ключи «encoding='utf-8'» и «ensure\_ascii=False».

1. Протестируйте свой скрипт на различных сценариях, чтобы убедиться, что он точно извлекает нужные данные.

Проверял скрипт на «Челябинск, сейчас, всё», «Челябинск, сейчас, фильмы» и «Челябинск, 01 мая, спорт».

Не стал пока добавлять автоматическое формирование названия выходного файла в зависимости от точки входа.

1. Предоставьте ваш Python-скрипт вместе с кратким отчетом (не более 1 страницы), который включает следующее: URL сайта. Укажите URL сайта, который вы выбрали для анализа. Описание. Предоставьте краткое описание информации, которую вы хотели извлечь из сайта. Подход. Объясните подход, который вы использовали для навигации по сайту, определения соответствующих элементов и извлечения нужных данных. Трудности. Опишите все проблемы и препятствия, с которыми вы столкнулись в ходе реализации проекта, и как вы их преодолели. Результаты. Включите образец извлеченных данных в выбранном вами структурированном формате (например, CSV или JSON). Примечание: Обязательно соблюдайте условия обслуживания сайта и избегайте чрезмерного скрейпинга, который может нарушить нормальную работу сайта.

Из неуказанного ранее: даже выбранный сайт «ругался» на отсутствие у меня сертификатов для парсинга – этот сайт позволил игнорировать выпадающие ошибки.

Кроме непосредственно файла скрипта DZ-07.py будут прикреплены файлы результатов tv\_programm.json, kino\_programm.json и sport\_programm\_01\_05.json, а также этот файл-отчёт ДЗ-07.docx. Все указанные файлы в одном архиве DZ-07.zip.