Вступление

Здравствуйте. Я ученик 11 «Т» класса, Аврамов Вячеслав Сергеевич. Темой моего проекта является разработка парсера для сбора данных с одного из крупных маркетплейсов на Python.

Актуальность

Последние годы маркетплейсы демонстрируют бурный рост. Крупные и мелкие торговые компании пользуются ими для реализации своих товаров конечному потребителю. Все заинтересованы в продаже максимального количества своей продукции. Для обеспечения этого необходимо определить конкурентную цену и оптимальное качество товара. Разработка ценового предложения требует актуальной информации по всем товарам в своей нише, т.е. мониторинг текущей ситуации. Именно для этого и необходима автоматическая система сбора информации со страниц маркетплейса, т.к. ручной сбор является очень затратным по времени.

Цель

Таким образом, целью данного проекта являлось создание парсера на python, который собирает информацию с одного из популярных в России маркетплейсов, а именно с Ozon.

Задачи

1. Определение маркетплейса для сбора информации
2. Определение набора данных для сбора
3. Разработка алгоритма сбора информации
4. Алгоритм обработки полученных данных, формат и экспорт в xlsx-файл
5. Определение оптимальных библиотек для реализации алгоритмов
6. Создание программы

Источники информации

1. Habr – крупное интернет-сообщество людей разных сфер деятельности, в том числе и программистов
2. <https://pandas.pydata.org/> - документация библиотеки Pandas
3. <https://www.selenium.dev/> - документация библиотеки Selenium

Результатом проведенной мною работы является программа, которая за 45 секунд собирает 360 наименований товаров заданной категории с 10 первых страниц сайта Ozon.

Процесс

Первый этап – написание программы, которая собирает нужную информацию. Много полезной информации по созданию парсера оказалось на сайте habr. Оттуда стало понятно, что для реализации цели не получится использовать простое решение, как дуэт библиотек BS4 и requests. Ozon создает страницу с помощью JS, следовательно необходимо ее страницу прогружать. Для решения этой задачи подходит библиотека Selenium, которая используется в первую очередь для автоматизированного тестирования сайтов, но также подходит и для парсинга. Чтобы использовать библиотеку, я воспользовался ее документацией. Этого хватило, чтобы написать алгоритм для сбора данных.

Следующий этап – экспорт данных в xlsx-файл. Для этого была использована библиотека Pandas, используемая для обработки и анализа данных. Для ее использования, я изучил документацию и несколько статей с сайта habr.

В процессе работы над проектом я научился работать с html файлами, изучил библиотеки Pandas, Selenium, BeautifulSoup и научился пользоваться веб-сервисом GitHub.

Я буду продолжать заниматься этим проектом, потому что существует множество способов улучшить программу, как например добавление многопоточности.

Благодарю за внимание.