Университет ИТМО, факультет программной инженерии и компьютерной техники

Двухнедельная отчётная работа по «Информатике»: аннотация к статье

Дата прошедшей лекции: 11.10.2022 Номер прошедшей лекции: 3 Дата сдачи: 20.12.2022

Выполнил(а) Пономарёв М.И. , № группы P3117 , оценка

Фамилия И.О. студента не заполнять

|  |  |  |
| --- | --- | --- |
| **Название статьи/главы книги/видеолекции**   1. Скрапинг современных веб-сайтов без headless-браузеров | | |
| **ФИО автора статьи (или e-mail)** [**host\_m**](https://habr.com/ru/users/host_m/) | **Дата публикации**  **(не старше 2019 года)**  "14" Января 2021 г. | **Размер статьи**  **(от 400 слов)**  600+ слов |
| **Прямая полная ссылка на источник или сокращённая ссылка (bit.ly, tr.im и т.п.)**  <https://habr.com/ru/company/vdsina/blog/537174/> | | |
| **Теги, ключевые слова или словосочетания**  Python, Скрапинг, Парсинг, Headless CMS | | |
| **Перечень фактов, упомянутых в статье**   1. Самой тяжёлой частью скрапинга являются HTTP-запросы, обработка не должна занимать много времени, однако необходимо постараться минимизировать количество запросов. 2. Безголовые браузеры превосходно справляются с извлечением информации с веб-сайтов, но они очень тяжелы. Они более надёжны и с меньшей вероятностью сломаются, чем доступ к недокументированным API 3. Цель веб-скрапинга — доступ к информации и её анализ, чтобы создать на её основе что-то полезное, а не вызвать проблемы и торможение серверов. | | |
| **Позитивные следствия и/или достоинства описанной в статье технологии (минимум три пункта)**   1. Способ скрапинга описанный в статье удобный и простой в использовании, потому что использует простые python запросы без использования сложных библиотек. 2. Данный способ скрапинга использует только публичное api, которые потребляет на фронтэнде большинство современных веб-сайтов, следовательно подходит для большиества современных сайтов. 3. Этот способ более уважителен к серверу, поскольку требует меньшего количества запросов и не обязывает загружать статические ресурсы. | | |
| **Негативные следствия и/или недостатки описанной в статье технологии (минимум три пункта)**   1. Данный способ парсинга подойдёт только если:  * API меняется не слишком часто * Вам нужны данные только один раз * Скорость очень важна * API предоставляет больше информации, чем сама страница  1. Скрапинг достаточно неустойчивый процесс, который запросто в любой момент может прерваться, поэтому важно часто сохранять данные и обеспечить удобный перезапуск скраперов с места, где произошёл сбой. 2. Скрапинг подходит исключительно для небольших объёмов данных | | |
| **Ваши замечания, пожелания преподавателю *или* анекдот о программистах[[1]](#footnote-1)** | | |

1. Наличие этой графы не влияет на оценку [↑](#footnote-ref-1)