ПАРСИНГ САЙТОВ НА PYTHON

Захаренко А. Д.

Белорусский государственный университет, Минск, Беларусь,
e-mail: arsen.zaharenko@gmail.com

Сегодня, во времена активного развития информационных технологий, новой информации становится всё больше с каждым днём.

Благодаря сети Интернет возможность узнать что-нибудь новое есть у каждого. Однако выбрать, проанализировать и структурировать нужные данные вручную порой бывает совсем непросто. В современном мире с этой задачей поможет справиться компьютер, в частности технология парсинга.

Парсинг – это процесс сбора данных с последующей их обработкой и анализом. К этому способу прибегают, когда нужно обработать большой объём информации и автоматизировать процесс работы с информацией.

Программа, которая производит сбор и синтактический анализ, называется парсер.

Для написания парсера отлично подойдёт язык программирования Python. Данный язык имеет понятный и простой синтаксис, адаптирован для работы с большим массивом данных.

Разработанный парсер был протестирован на задаче нахождения букмекерских вилок.

Букмекерская вилка – это ситуация, когда разница коэффициентов в двух или более конторах позволяет сделать по ставке на каждый взаимоисключающий исход у разных букмекеров и остаться в прибыли при любом результате.

Чтобы справиться с этой задачей разобьём её на небольшие этапы:

1. Выгрузить и сохранить HTML-страницы букмерских контор A и B.
2. Распарсить HTML-документ для конторы A в удобный для дальнейшего анализа формат.
3. Распарсить HTML-документ для конторы B в удобный для дальнейшего анализа формат.
4. Проанализировать полученные данные и отобрать события, для которых есть вилка.

Условие наличия вилки:

$$\frac{1}{k_1} + \frac{1}{k_2} < 1$$

где k_1 – коэффициент на первый исход в конторе A, k_2 – коэффициент на второй исход в конторе B.

Таким образом, парсинг помог решить достаточно сложную задачу, если бы она решалась вручную.

Существует очень много практических сфер, где требуется доступ к данным практически неограниченного объёма.

Рыночное прогнозирование рынка, машинный перевод и даже медицинская диагностика уже извлекли огромную пользу, воспользовавшись возможностью собрать и проанализировать данные новостных сайтов, переведённый контент и сообщения на медицинских форумах.

Даже в мире искусства парсинг уже открыл новые горизонты для творчества. В рамках проекта 2006 года «We Fell Fine» Джонатан Харрис и Сеп Камвар провели парсинг англоязычных блогов для поиска фраз, начинающихся с «I fell» или «I am feeling». Это позволило построить визуализацию данных, описать, как люди в мире чувствуют себя изо дня в день, с минуты на минуту.

Независимо от вашей предметной области, почти всегда есть способ, благодаря которому парсинг может повысить эффективность бизнес-практик, улучшить производительность или даже открыть совершенно новое направление в бизнесе.

Литература

1. Mitchell, R. Web Scraping with Python. Collecting Data from the Modern Web / R. Mitchell. – Published by O'Reily Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.