Сбор и подготовка данных

При анализе поставленных задач были выделены три основные сущности, о которых будет собираться информация: игрок, команда и турнир.

Об игроке необходима следующая информация: ник, ФИО, страна, дата рождения, годы активности, общий денежный выигрыш, турниры, в которых принимал участие, и занятое место.

Требующиеся данные о турнирах: название, тип турнира (онлайн или офлайн), страна, призовой фонд, дата проведения.

Необходимые данные о команде: название, страна, общий денежный выигрыш, турниры, в которых принимала участие, и занятое место.

Было принято решение брать данные с сайта liquipedia.net по нескольким причинам. Во-первых, данный сайт содержит информацию о большом количестве киберспортивных игр. В отличии от liquipedia большинство сайтов, ведущих статистику киберспорта, специализируется на какой-то одной конкретной игре. Так как предполагается, что анализироваться будут несколько игр, то целесообразно выбрать универсальный источник информации, так как это поможет снизить трудозатраты на создание парсера.

Во-вторых, liquipedia предоставляет все необходимые данные для дальнейшего анализа за весь промежуток времени существования киберспорта.

Таким образом, нашими изначальными данными являются полуструктурированные данные в лице HTML-страниц. Содержимое этих страниц анализируется и из них извлекается вся полезная информация и записывается в структурированном виде в файлы, для дальнейшего анализа.

На данный момент есть три основных бесплатных инструмента для анализа HTML-страниц: Scrapy, Selenium, Beautiful Soup.

Scrapy – это фреймворк предназначенный для получения данных с сайтов. Он предоставляет большое количество функции для анализа интернет страниц. Является одним из самых быстрых средств веб-скрапинга, и при этом достаточно экономичен в плане производительности. Легко расширяем и позволяет создавать гибкие и надёжные приложения.

Одной из самых больших проблем Scrapy это плохая документация продукта, что создаёт большие трудности при обучении работе с данным фреймворком.

Selenium – это инструмент для автоматизации действий браузера. Изначально он предназначен для создания тесто для веб-приложений, но также его можно использовать для анализа содержимого страниц.

Он имеет достаточно простой интерфейс, позволяет легко работать с Javascript, AJAX и PJAX. Но в тоже время работает он достаточно медленно.

Beautiful Soup – это инструмент для веб-скрапинга для языка Python. Он имеет хорошую документацию и простой интерфейс, что делает его легким в обучении. Предоставляет большой набор функций для быстрого извлечения данных с конкретных страниц. Но данный инструмент не имеет встроенных средств для запросов к серверу и анализа HTML-страниц. Таким образом, для использования Beautiful Soup требуется устанавливать их отдельно.

Принимая во внимания вышеперечисленное был выбран Beautiful Soup, так как он лучше всего подходит новичкам для создания небольших проектов, в которых не требуется работа с javascript.

В качестве средства для обращения к серверу и анализа HTML были выбраны библиотека Requests и анализатор html.parser, как одни из самых популярных представителей своих классов.

Для сохранения данных в настоящее время используется три основных формата: CSV, XML, JSON.

Формат CSV представляет собой список элементов, разделённых запятыми. Преимуществом данного формата является его компактный размер. Меньший из всех рассматриваемых. Главная проблема данного формата — это то, что он не позволяет сохранять данные в виде иерархии.

Формат XML полностью поддерживает хранение данных в структурированном виде. Но при этом требует большого количества места, превышающее примерно в три раза чем те же данные в формате CSV.

Формат JSON так же поддерживает представление данных в иерархическом виде. Сохранённые данные занимают примерно в два раза места больше чем в формате CSV.

Так как нам требуется хранить структурированные данные, то формат CSV нам не подходит. При выборе между XML и JSON предпочтение было отдано последнему из-за меньшего размера.