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, 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.