УДК 004

АВТОМАТИЗАЦИЯ СБОРА ИНФОРМАЦИИ ДЛЯ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

В. А. ЛИВИНСКАЯ
Белорусско-Российский университет
Могилев, Беларусь

Интеллектуальный анализ данных предполагает извлечение знаний из информации, хранящейся в структурированном виде. Однако получить доступ к такому представлению информации не всегда представляется возможным. Основной целью исследования явилась автоматизация сбора информации о вакансиях в IT-области, размещенной на сайте jobs.dev.by.

Для получения всего массива вакансий с целью выявления наиболее востребованных компетенций, уровня предлагаемой заработной платы в аналогичных позициях в разных компаниях возможно либо ручное переписывание данных с веб-страницы, либо разработка программы (так называемого парсера) для автоматизации данного процесса.

На сайте jobs.dev.by отсутствуют официальный открытый API, возможность организации XHR-запросов, а также JSON-а в конце HTML-файла. Поэтому единственным доступным способом получения информации явилось написание парсера этого HTML-кода.

Для осуществления этого процесса был использован свободно распространяемый язык программирования Python, т. к. он предоставляет огромное количество открытых библиотек, в том числе и для парсинга. В качестве фреймворка был выбран Scrapy. С помощью данного приложения был извлечен и структурирован массив данных, представленный в виде датасета со следующими полями:

– название компании;
– название вакансии;
– количество вакансий компании на текущий момент;
– специализация, уровень и все остальные ключевые моменты, описанные в блоке информации;
– основные тэги, находящиеся под блоком информации;
– информация о компании;
– информация о рекрутере данной вакансии.

Непосредственное извлечение знаний из этого массива осуществлялось с помощью библиотек языка анализа данных *R*, позволяющих анализировать как количественные, так и качественные признаки единиц совокупности и визуализировать результаты анализа.