Анализ текстов, содержащих почтовые адреса

В. Р. Фёдорова

В статье [1] описана реализация информационной системы сбора адресов. В рамках системы был реализован поисковый бот для скачивания WEB-страниц, реализован модуль поиска и определения адреса в тексте с последующим геокодированием. Таким образом, были собраны WEB-страницы, содержащие в своем тексте почтовые адреса. После всех этапов поиска адреса, необходимо произвести анализ текстов, а именно, выделить наиболее ценную информацию, связанную с адресами. Анализ текстов проводится в следующих направлениях: классификация текстов и установление семантических отношений между текстами. Классификация текстов производится на основе применения методов обучения с учителем. На текущий момент выполняется поиск размеченных текстов. Установление семантических отношений реализовано следующими методами:

1) связывание по заголовку WEB-страницы: если в тексте одной WEB-страницы находится множество различных почтовых адресов, то с помощью заголовка можно определить тематику, объединяющие эти адреса;
2) анализ текста и заголовка WEB-страницы с помощью ключевых слов: для выделения основной информации о событиях, происходящих по адресам, необходимо составить список ключевых слов, обозначающих возможные события: ремонт, мероприятие, выставка и т. д. Затем осуществить поиск данных ключевых слов в тексте WEB-страницы;
3) анализ пространственного положения событий или объектов, которые описывают тексты;
4) связывание данных по дате и времени событий.

С помощью описанных выше методов были сгруппированы адреса по тематикам. Например, 56 адресов на странице https://irkobl.ru/authorities/mirovye-sudi/ сгруппированы по теме «Судебные участки».

Литература

1) Valentina R. Fedorova, Roman K. Fedorov - «Collecting data with postal address on the Internet» [Электронный ресурс] / CEUR – Режим доступа: http://ceur-ws.org/Vol-2677/paper12.pdf