Унификация адреса на основе методов NLP

В.Р. Фёдорова,
В.В. Парамонов

Приведение адреса к единому формату является актуальной задачей для многих сфер. Основная проблема – это сложность автоматической обработки адресов. В большинстве систем адрес заполняется человеком вручную, поэтому не исключены различные ошибки, такие, как отсутствие ключевых слов, неправильное наименование элемента планировочной структуры, неоднозначность и т.д. В свою очередь, унификация адреса решает множество задач, например, проверка соответствия адресов, получение геопозиции объекта и т.д. Рассмотрим основные шаги алгоритма.

Шаг 1: Подготовительные действия.

Производится загрузка в оперативную память всех населенных пунктов Российской Федерации. Информация об адресных объектах и их связях берется из классификатора адресов Российской Федерации [1]. Хранение населенных пунктов в оперативной памяти требуется для уменьшения количества запросов к базе данных.

Шаг 2: Получение и предобработка данных.

Выполняется выгрузка из базы данных обрабатываемых адресов в виде массива строк. Для каждого адреса производится токенизация адреса с помощью библиотеки для обработки естественного языка Natasha [2]. Полученные токены приводятся к нижнему регистру, удаляются специфичные стоп-слова и символы.

Шаг 3: Выделение субъекта Российской Федерации.

Так как предполагается обрабатывать большой массив данных, то изначально для сужения поиска последующих структурных элементов адреса, необходимо найти субъект Российской Федерации. Для этого каждый токен предложения проверяется на соответствие ключевым словам субъектов Российской Федерации или их сокращениям, например край, республика и т.д. После того, как совпадение было найдено, токены, близлежащие к токену с ключевым словом, проверяются на наличие в массиве с субъектами Российской Федерации. При совпадении токена с субъектом запоминается код региона. Обработанные токены удаляются из последующего анализа.

Шаг 4: Поиск населенного пункта.

В оставшихся необработанных токенах производится поиск населенного пункта по такому же алгоритму, т.е. поиск ключевого слова, проверка близлежащих токенов в массиве с населенными пунктами региона, который был выявлен при поиске субъекта Российской Федерации. Если среди возможных вариантов населенных пунктов встречаются составные названия, например Ростов-на-Дону или Русская Аларь, то алгоритм проверяет наличие всех составных частей среди токенов предложения, а также придает составным названиям большую значимость.

Шаг 5: Запись полученных результатов в базу данных.

Шаг 6: Поиск улицы.

Хранить все улицы Российской Федерации в памяти невозможно на персональных компьютерах ввиду большого объема требуемой оперативной памяти. Поэтому их выгрузка производится частями по регионам. Предобработанные адреса сортируются по выделенным регионам, что позволяет загрузить улицы для текущего региона и обработать все адреса этого региона. Поиск улицы проводится по той же логике, что и поиск населенного пункта.

Шаг 7: Поиск номера дома и квартиры.

Поиск номера дома и квартиры осуществляется с помощью регулярных выражений. Все возможные варианты номера дома проверяются в классификаторе адресов.

В результате работы данного алгоритма удалось обработать около 3,4 млн адресов с точностью 85%. В ходе оптимизации кода скорость обработки одного адреса снизилась с 1 минуты до 18 миллисекунд в среднем. Тестирование алгоритма на реальных данных показывает практическую его значимость и применимость.

Литература

1) Федеральная информационная адресная система в формате КЛАДР 4.0 [Электронный ресурс] – Режим доступа: https://fias.nalog.ru/Updates (дата обращения: 29.04.2021).
2) Natasha [Электронный ресурс] – Режим доступа: https://github.com/natasha/natasha (дата обращения: 29.04.2021).