РАЗРАБОТКА МЕТОДА ИЗВЛЕЧЕНИЯ НАИМЕНОВАНИЙ ГОРОДСКИХ СЕРВИСОВ ИЗ НЕФОРМАЛИЗОВАННЫХ ТЕКСТОВ СОЦИАЛЬНЫХ СЕТЕЙ

Г. Ю. Худяков
Университет ИТМО
Санкт-Петербург

Работа посвящена проблеме обработки неформализованных сообщений граждан о городской инфраструктуре. Обращения граждан — один из немногих источников данных, необходимых для организации мероприятий по улучшению городской среды, управления и проведения различных исследований, связанных с городом. Разработанный метод, на основе NLP (обработка естественного языка), призван обеспечить возможность извлечения обращений и сообщений о городской среде, упоминаемых в них городских сервисов. В центре метода находится модель машинного обучения, которая была обучена на 10 тыс. комментариях граждан в социальной сети «ВКонтакте». Данные для обучения модели были размечены при помощи большой языковой модели GPT-4. Этот подход позволит получить актуальную информацию о сервисах, которые, при сочетании с другими методами и подходами, могут быть использованы как для проведения различных научных исследований, так и для трансформации городской среды.

Ключевые слова: обработка естественного языка, тексты социальных сетей, цифровая урбанистика, граф знаний, машинное обучение

DEVELOPMENT OF A METHOD FOR EXTRACTING URBAN SERVICES FROM UNSTRUCTURED TEXTS OF SOCIAL NETWORKS

G. Yu. Khudyakov
ITMO University
St. Petersburg

The work is dedicated to the problem of processing unstructured citizen comments/appeals about urban infrastructure. Citizen comments/appeals are one of the few data sources necessary for organizing activities to improve the urban environment, management, and conducting various research related to the city. The developed method based on NLP is designed to provide the ability to extract appeals and messages about the urban environment, the urban services mentioned in them. At the heart of the method is a machine learning model that was trained on 10,000 citizen appeals on social network «VKontakte». The data for training the model were processed using the large language model GPT-4. This approach will allow obtaining up-to-date information about services, which, when combined with other methods and approaches, can be used both for conducting various scientific research and for transforming the urban environment.

Keywords: NLP, social network texts, digital urbanism, knowledge graph, machine learning

Введение

Городской сервис — понятие, не имеющее единой коннотации. В обывательском ключе первое, что приходит в голову, когда слышим слово «сервис» — это люди, оказывающие какую-либо услугу. В контексте городских сервисов данное определение не будет корректным. Под городским сервисом, в данном случае, подразумевается городской объект, физически существующий в городской среде и представляющий определенную возможность или услугу населению [1]. Сервисы разделяют на разные типы: базовая инфраструктура, инфраструктура досуга, инфраструктура обслуживания населения, социальная инфраструктура, транспортная инфраструктура, туристическая инфраструктура. На практике каждый сталкивается с различными городскими сервисами ежедневно: посещая продуктовый магазин, прогуливаясь по парку, скрываясь от дождя под козырьком остановки городского транспорта. Большое количество, плотность, качество и разнообразие городских сервисов являются важными характеристиками городской среды, позволяющими качественно и количественно отличить её от территории более низкой организации [1].

Таким образом, исследования городских сервисов являются важным направлением в изучении города. Помимо этого, изучение городских сервисов имеет вполне прикладной характер. В контексте различных ресурсов, позволяющих сообщить о проблеме, таких как «Наш Санкт-Петербург», изучение городских сервисов дает возможность различным представителям органов власти Санкт-Петербурга обеспечивать

грамотную политику по обслуживанию и улучшению городской среды. Упомянутые причины послужили поводом для создания метода извлечения городских сервисов из неформализованных обращений горожан.

Цель разработки метода — создание нового инструмента для анализа неформализованных текстовых данных из социальных сетей для научных и прикладных исследований, связанных с городской средой. Разработка метода проходила в рамках научно-исследовательского проекта «Библиотека алгоритмов машинного обучения и обработки естественных языков для обогащения пространственных городских моделей и моделирования вернакулярной оценки качества городской среды» в 2024 г.

Особенности создание и примеры применения метода

Техническая новизна данного проекта в том, что для его создания активно использовалась большая языковая модель GPT-4 от компании OpenAI [2]. Вначале при помощи модуля парсинга комментариев в социальной сети ВКонтакте, были собраны массивы комментариев из групп, посвященных районам Санкт-Петербурга, за несколько лет. Далее был сформирован предположительный перечень городских сервисов (порядка 130 наименований), который в дальнейшем загружался в чат GPT вместе с массивом комментариев и запросом о том, чтобы языковая модель обнаружила сервисы, относящиеся к перечню, в представленном массиве комментариев. На выходе получился массив из 10 тыс. комментариев, в которых были обнаружены те или иные городские сервисы. Далее происходил процесс предобработки: предложения разбивались на отдельные слова и транспонировались в формат BIO, а затем полученный массив разбивался на train, test и dev массивы. Полученные массивы в формате BIO использовались для обучения модели классификатора при помощи библиотеки Flair. В качестве модели для embedding использовалась модель FastText, которая была обучена на массиве статей из Википедии [3]. Она одна из немногих легких моделей, которая дает возможность работать с текстами на русском языке [4].

Разработанный метод извлечения наименований городских сервисов предоставляет большой простор для исследований. В рамках научно-исследовательского проекта использовался данный метод в сочетании с модулем геокодирования, который извлекал из сообщений горожан адрес и определял его географическую координату. В дальнейшем сочетание этих двух методов позволяет построить пространственный семантический граф, который позволяет анализировать упоминания тех или иных городских сервисов с привязкой к географическому происхождению данного сообщения/комментария.

Другим примером использования метода извлечения наименования сервисов может служить комбинация с моделью определения тональности предложений. Используя подобную модель, можно оценить тональность — определить эмоциональное отношение автора к объекту высказывания, в нашем случае, к городскому сервису. Таким образом можно провести оценку качества городских сервисов на разных уровнях городской структуры. При комбинации этих методов можно достичь субъективной оценки городской среды [5].

Таким образом, данные методы обработки неформализованных текстов комментариев в социальных сетях и обращений граждан открывают новые горизонты для систематического анализа с целью организации системы управления и обслуживания городских сервисов.

Заключение

Развитость сервисной инфраструктуры имеет большое значение, поскольку не только обеспечивает удобство района, но и создает потенциал для развития социальных, экологических и экономических ценностей среды. Наличие городских сервисов в пешеходной доступности сказывается как на стоимости недвижимости, так и на формировании экономического благополучия и психологического комфорта жителей. Этим обусловлена важность изучения городских сервисов как одной из категорий городского анализа.

Разработанный метод извлечения наименований городских сервисов из неформализованных текстов может стать важным подспорьем при проведении различных социологических, экономических и урбанистических исследований, а также одним из инструментов, используемых при организации различных мероприятий по обслуживанию городской среды.

Увеличение и усложнение структуры города создает вызовы для представителей городских властей и исследователей, занимающихся проблемами городов. В дальнейшем необходимо обратить внимание на другие важные объекты городской инфраструктуры, анализ которых позволит обогатить цифровую модель города.

ЛИТЕРАТУРА

1. Ненько А. Е., Недосека Е. В., Курилова М. С. «Соседскость» городских сервисов как измерение пространственной сегрегации // Laboratorium. 2022. Т. 14, № 3. С. 34–58. DOI: 10.25285/2078-1938-2022-14-3-34-58.
2. ChatGPT by OpenAI // ChatGPT. URL: https://chatgpt.com/ (дата обращения: 20.05.2024).
3. Fasttext: Library for efficient text classification and representation learning // Fasttext. URL: https://fasttext.cc/.

4. Низомутдинов Б. А. Тестирование методов обработки комментариев из Telegram-каналов и пабликов ВКонтакте для анализа социальных медиа // International Journal of Open Information Technologies. 2023. Т. 5. С. 137–145.
5. Nizomutdinov B., Uglova A. Application of data from social networks for value-based management of city development programs // Computational Science and its Applications (ICCSA 2023). Lecture Notes in Computer Science (LNCS. Vol. 13957). Springer, 2023. P. 369–382. DOI: 10.1007/978-3-031-36808-0_24.
6. Antonov A., Vidiasova L., Chugunov A. Detecting Public Spaces and Possibilities of Risk Situations in Them via Social Media Data // Social Computing and Social Media (HCII 2023). Lecture Notes in Computer Science (LNCS. Vol. 14025). Springer, 2023. P. 3–13 DOI: 10.1007/978-3-031-35915-6_1.