УДК 004.62

РАЗРАБОТКА МЕТОДА ВЫЯВЛЕНИЯ ТОЧЕК ОБЩЕСТВЕННОГО ИНТЕРЕСА НА ОСНОВЕ ОТКРЫТЫХ ДАННЫХ

Катынсус А.В. (ИТМО),
Научный руководитель – кандидат технических наук, директор Института дизайна и урбанистики Митягин С. А.
(ИТМО)

Введение. Сегодня у понятия точки общественного интереса нет ни четкого определения, ни нормативной базы, для этого в представленной работе составлено определение для данного термина, которое понадобится дальше для интерпретации результатов. Точка общественного интереса — это объект на карте, касающийся инфраструктуры общества и содержащий информацию о реальных общественно значимых местах, которые имеют координаты, название и характеристики, а также привлекающий интерес людей в большей степени, чем окружающее его пространство.

На данный момент представлено небольшое количество методов, с помощью которых можно найти и выделить такие точки [1-4]. Они концентрируются в основном на сборе закрытых данных опросов и анкет, что делает их невоспроизводимыми по определению. Методов, которые работают с открытыми данными из множества источников, вовсе нет. Исследователи в основном выбирают одну социальную сеть или один сервис, с которого берут всю информацию. Таким образом, формируется узкая целевая аудитория.

Поиск информации по точкам общественного интереса через поисковые сервисы часто не предоставляет нужный результат, в особенности если вводить запрос на русском языке. В России на данный момент формулировка «точка общественного интереса» не является распространенной, поэтому и результаты объединяются с общественными пространствами и достопримечательностями. Метод, разрабатываемый в данной исследовательской работе призван решить основные проблемы с нехваткой данных, невоспроизводимостью и автоматизацией, а также обогащением данных. С его помощью аналитики городских данных смогут определять, какие новые места в городе являются популярными, а какие наоборот, не пользуются спросом. Также он даст возможность определять эмоциональную составляющую точки и пространства в целом, получить ее ключевые характеристики.

Основная часть. Разрабатываемый в представленной работе метод по выявлению точек общественного интереса будет базироваться на открытых данных из социальных сетей и картографических сервисов, что позволит не собирать данные с помощью опросов и анкет у жителей города.

Разработанный метод разбит на несколько частей, в каждой из которых применятся разные технологии. В первой части, посвященной сбору данных, используютс основные методы сбора данных: скрапинг и парсинг. Скрапинг позволит автоматизировать процесс сбора данных с выбранных сайтов и приложений. Для первой части был написан DAG Airflow со всеми задачами для сбора данных из Яндекс карт, групп, постов и фотографий с описанием из социальной сети ВК.

Вторая часть подразумевает анализ полученных данных и выделение точек общественного интереса с их характеристиками (ключевыми словами) и эмоциями. На данном этапе используются методы обработки естественного языка: токенизация, лемматизация, мешок слов, tf-idf и удаление графической информации из текстов.

Третья часть состоит из выделения именованных сущностей с использованием заранее обученной модели Stanza, дообученной на собранных текстах [5]. В качестве пайплайна для обработки данных здесь используется токенизация. Далее происходит процесс выделения именованных сущностей с фильтрацией по типам.

На четвертом этапе определяются привлекательность каждой точки и ключевые слова, описывающие данную точку. Здесь используются методы машинного обучения с заранее

обученными моделями на распознавание эмоций в тексте и модели для определения ключевых слов, основанные на модели BERT. Выделение эмоций из текста происходит с помощью трансформера blanchefort/rubert-base-cased-sentiment, а для определения ключевых слов используется сущность KeyBERT() из библиотеки keybert.
Разработанный метод имеет множество сценариев для применения. В качестве исходного запроса можно отправить и полигон в черте города, в котором нужно найти все точки общественного интереса, и область города, и район, и улицу. Главная сложность – выделить из ВК нужные посты в группах и на стенах, которые относятся к выбранному полигону. В данном случае решением является указание названия района или улицы, а также координаты поиска.
Этапы со второго по четвертый выделены в отдельный Airflow DAG для обработки данных и выделения точек общественного интереса.
Данный метод может подойти и для данных, собранных другим образом. Тогда не будет требоваться включать даг с выгрузкой данных. Нужно запустить лишь второй DAG, заранее загрузив в базу все данные, и далее запускать метод на них.

Выводы. Данная работа сосредоточена на создании нового метода по выявлению точек общественного интереса на основе открытых данных, используя все плюсы и минусы ранее изученных аналогов и похожих алгоритмов.
В ходе разработки метода были использованы основные библиотеки для работы с NLP, выделены модели для выделения именованных сущностей и анализа текстов на тональности. Был разработан метод сбора данных из трех разных источников. Данные были проведены через несколько этапов: предобработка, очистка, обработка, приведение к виду, понятному для машины. В результате работы метода был собран 1441 текст из разных источников. Из них локации были найдены в 39 процентах текстов, куда были включены отзывы из картографических сервисов, которые уже имеют адреса. Коэффициент геолоцирования в данных из социальной сети ВК 88 процентов. Наличие точек в текстах отзывов (0.09% геолоцирования) подтверждает теорию о том, что в отзывах люди упоминают и новые места.

Список использованных источников:

1. Binak Beqaj. Public Space, public interest and Challenges of Urban Transformation // ScienceDirect. IFAC-PapersOnLine – 2016 - С. 320–324.
2. Л. В. Козлова. Методика исследования общественных пространств центра города как основа их совершенствования // Строительство и архитектура – 2015 – С. 1–6.
3. Uriwan Angkhawey, Veera Muangsin. Detecting Points of Interest in a City from Taxi GPS with Adaptive DBSCAN // Conference: 2018 Seventh ICT International Student Project Conference (ICT-ISPC) – 2018 – С. 1 – 6.
4. Nai Chun Chen, Yan Zhang, Marrisa Stephens, Takehiko Nagakura, Kent Larson. Urban Data Mining with Natural Language Processing: Social Media as Complementary Tool for Urban Decision Making // - CAADFutures – 17 – 2017 – С. 101 – 109
5. Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, Christopher D. Manning. Stanza : A Python Natural Language Processing Toolkit for Many Human Languages // Conference: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations – 2020 – С. 1 – 8.