УДК 004

РАЗРАБОТКА СИСТЕМЫ ДЛЯ МОНИТОРИНГА ПРОИСШЕСТВИЙ НА ОСНОВЕ ДАННЫХ ИЗ СОЦИАЛЬНЫХ СЕТЕЙ. КЕЙС САНКТ-ПЕТЕРБУРГА

Низомутдинов Б.А. (Университет ИТМО), **Беген П.Н.** (Университет ИТМО)
Научный руководитель – к.т.н., директор Института дизайна и урбанистики Митягин С.А. (Университет ИТМО)

Рассмотрены результаты работы разработанной системы по выделению адреса и типа происшествия из текстовых данных социальных сетей. В качестве тем для распознавания были выделены следующие: Угон, Авария, Пожар, Ограбление, Нападение. Определена точность распознавания.

Из-за отсутствия потенциала по ограничению воздействия опасных факторов многие города по-прежнему сталкиваются с высоким уровнем угроз. По мере того, как угроз в городах становится все больше, повышение устойчивости городов становится главной задачей. Для повышения устойчивости городов растет потребность в информации, которая имеет значение для всех этапов развития городов. Таким образом, лучшее понимание пространственно-временных закономерностей общественного реагирования является ключевым шагом на пути к снижению ущерба и повышению устойчивости городов.

Так, группа исследователей из Нью-Йоркского университета проанализировала данные об инцидентах из двух различных источников: от традиционного поставщика данных, который собирает отчеты об инцидентах от нескольких агентств, и сообщения пользователей из Twitter во время урагана Сэнди, затопившего многие районы Нью-Йорка в 2012 году. Результат показал, что Twitter может предоставить подробную информацию о местоположении конкретного инцидента, а также его интенсивности, продолжительности и т. д.

Оперативные данные о состоянии города представляют большую ценность для многих отраслей, ими пользуются органы власти, силовые структуры и аварийно-спасательные службы, такую информацию могут использовать исследователи в своих проектах. Но, как правило, исследователи получают сжатый вариант официальной информации, например о ДТП или угонах, кроме того, такие отчеты публикуются раз в квартал или год. Также, сами оперативные службы, часто агрегируют информацию с запозданием, в виду сложности коммуникации. Альтернативный источник информации о происшествиях - результат общения людей в социальных сетях. Люди ставят отметки на карте, пишут в онлайн сообществах о происшествиях, практически, в режиме онлайн.

В данном исследовании мы рассматриваем информацию о происшествиях в городе, полученную из социальных сетей, на примере Санкт-Петербурга. В качестве источника информации выбрано сообщество Вконтакте - ДТП и ЧП | Санкт-Петербург | Питер Онлайн.

В 2017 компания Яндекс представила аналитический инструмент, который отражает концентрацию аварийных участков на карте. В исследовании использованы данные Яндекс.Навигатора и Яндекс.Карт где пользователи предупреждают друг друга об авариях, камерах, дорожных работах и эвакуаторах, просто общаются. Сотрудники компании проанализировали распределение отметок о ДТП по городу и нашли самые опасные и самые аварийные места. Однако, данный инструмент снова не отражает актуальных данных.

В своем исследовании, мы предлагаем использовать данные из социальных сетей, для поиска и отражения на карте различных происшествий - Угон, Авария, Пожар, Ограбление, Нападение.

Разработка сервиса для автоматической обработки и анализа набора данных постов из социальных сетей являлась одной из задач исследования. Сервис подразумевает

автоматическое определение темы в тексте поста, а также распознавание адреса или его частей, например улица, дом, район и т.д., на основе методов машинного обучения.
Для разработки сервиса автоматического выделения сущностей в русскоязычном тексте была использована библиотека natasha. Из данной библиотеки был использован стандартный извлекатель адресов. Для распознавания тематик постов был настроен парсер с помощью встроенного модуля yargy. Для этого в yargy были заведены специальные правила и отношения с помощью контекстно-свободных грамматик, созданы отдельные сущности под каждую из тематик. В качестве тем для постов были выделены следующие: Угон, Авария, Пожар, Ограбление, Нападение. Количество постов по этим пяти тематикам составляет примерно 1/3 от общего набора. В рамках исследовательской работы были добавлены правила с готовыми предикатами yargy-парсера, распознающие «облако тегов» для выделения темы. Например, для тематики Угон были использованы слова: угон, угнали, кража автомобиля, похищение ТС.
На основе собранной выборки были получены результаты по автоматическому выделению тематик постов и частей адресов, а также получена средняя точность (accuracy) распознавания, рассчитанная как процентное соотношение распознанных сущностей библиотекой natasha с сущностями, выделенными вручную.
В результате сервис определил тематику «Угон» с точностью 82,13%, тему «Авария» – 86,25%, «Пожар» – 94,91%, «Ограбление» – 100%, «Нападение» – 96,77%. Стоит отметить, что количество постов для последних двух тематик крайне низкое (<70), что не позволяет говорить о достаточности полученных данных точности. Также был отмечена длительная продолжительность времени работы парсера, на основе чего можно сделать предположение о медленной работе алгоритма, и с увеличением выборки данных (более 100 тыс. записей) алгоритм парсера будет работать еще медленнее. На дальнейших этапах исследования планируется решить данную проблему и повысить точность распознавания тематик.
Для распознавания адресов в тексте была использована встроенная функция AddrExtractor из библиотеки natasha. Распознавание проводилось на всей выборке данных, а также на отдельных частях выборки, разделенных по тематикам. Для подсчета средней точности распознавания было задано условие, что если в тексте поста распознана хотя бы одна часть адреса (например, улица, название, номер дома и т.д.), то адрес считается распознанным. В результате средняя точность распознавания адресов на всей выборке составила 58,76%. Для тематики «Угон» средняя точность составила 78%, для «Авария» – 60%, «Пожар» – 81%, «Ограбление» – 45%, «Нападение» – 57%. При распознавании адреса было отмечено, что наибольший процент точности достигается при определении части адреса если есть слово-маркер, например «ул.», «пр.» в формате «Московская ул.» или «Ленинский пр.». Однако, если убрать слово-маркер, то точность распознавания существенно снижается.

Выводы.

Метод показал свою перспективность, полученные данные могут быть использованы, как исследователям, так и представителями государственных ведомств. В данный момент ведется разработка картографического сервиса, для визуальной аналитики. Данное исследование проведено в рамках НИР Университета ИТМО № 620179 «Разработка картографического сервиса мониторинга потребностей жителей в развитии инфраструктуры городской среды с применением автоматизированных систем обработки данных из социальных сетей».

Низомутдинов Б.А. (автор) Подпись

Беген П.Н. (автор) Подпись

Митягин С.А. (научный руководитель) Подпись