**УДК** 004.912

# ОПРЕДЕЛЕНИЕ ТЕМАТИКИ ОБЪЕКТОВ ГОРОДСКОЙ СРЕДЫ НА ОСНОВЕ ДАННЫХ СОЦИАЛЬНЫХ СЕТЕЙ

**Дрожжин А.И.** (НИУ ИТМО)
**Научный руководитель – к.э.н., Репкин А.И.** (НИУ ИТМО)

Доклад предлагает взгляд на использование данных социальных сетей для определения значимости и тематик различных объектов городской среды. Работа посвящена описанию исследований методов определения конкретных объектов, упоминаемых в текстах пользователей социальных сетей.

**Введение.** Многие объекты культурного наследия в современном мире находятся под угрозой. Это происходит по двум причинам: во-первых, выявленные объекты зачастую переводятся из списка «представляющих ценность» и сносятся ради постройки жилого комплекса или бизнес-центра; во-вторых, невыявленные объекты в принципе не состоят на охране государства и могут быть уничтожены для тех же самых целей.
Однако в современном мире почти стерлась грань между общепризнанными объектами искусства и предметами или местами, важными для относительно небольшой группы людей. Чаще всего последнее связано не маргинальностью интересантов, а с неизвестностью самого места как такового.
Была поставлена задача определить возможность идентификации объектов по открытым данных социальных сетей, их тематику и значимость.

**Основная часть.**
До начала работ был произведен выбор наиболее удачного инструмента для работы с постами в социальных сетях.
В основе этой задачи лежал набор данных, собранный на основе постов в социальных сетях с эталонной разметкой именованных сущностей. На начальный момент в наборе находилось около 500 размеченных записей.
В качестве инструментов для выявления именованных сущностей были выбраны SpaCy (с предобученной языковой моделью SpaCy-ru) и Natasha (Slovnet BERT NER). После оценки результатов выделения именованных сущностей этими инструментами, был сделан вывод о недостаточном качестве оригинального набора данных.
Было принято решение дообучить Entity Recognizer выбранных инструментов. Для этого был создан второй набор в 2000 размеченных записей. По итогам второго обучения был проведен анализ результатов, которые доказали значительную эффективность инструмента SpaCy. Кроме того, можно отметить также простоту и удобство работы с ним.
После обучения SpaCy был применен ко всему имеющемуся объему обработанных данных социальных сетей. В итоге был получен предварительный, сырой список городских объектов.
Дальнейшая работа связана с темпоральным майнингом текстов: необходимо из массива имеющихся данных извлечь упоминания собственно объектов городской среды: время, действие, связь с какими-либо объектами для выявления отношения автора текста к объекту. Это необходимо для понимания связи между объектом и локации сообщения. Например, сообщения об Эрмитаже могут иметь геометку, указывающую на сам музей, а могут содержать упоминание о музее в контексте поста, имеющую иные геометки. Такие сообщения надо отсортировать и выбраковывать.
В контексте поставленной задачи имеет смысл вести статистику упоминания тех или иных объектов и в подобных сообщениях для определения значимости объекта. Однако эти сообщения сложно (практически невозможно) корректно перепривязать к связанным с объектом настоящим геометкам.

После проведения предварительной выбраковки полученный список городских объектов требуется дополнительно обработать: удалить мусор, отфильтровать по типам сущностей, сагрегировать синонимы.
По итогам работ был получен рабочий набор, на основе которого предполагается дальнейшая работа по уточнению тематик и уровня значимости тех или иных объектов городской среды.

**Выводы.** Рассмотренный алгоритм позволил утвердительно ответить на поставленную задачу возможности идентификации объектов по открытым данным социальных сетей. В ходе работы был вычислен наиболее удобный и корректный инструмент работы с данными, были проведены работы по очистке и фильтрации выбранным инструментом. В результате работ был получен рабочий набор городских сущностей, готовый для дальнейшей работы по установлению тематики и значимости.

Дрожжин А.И. (автор) Подпись

Репкин А.И. (научный руководитель) Подпись