УДК 004.89

ИССЛЕДОВАНИЕ ПРИМЕНЕНИЯ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОПРЕДЕЛЕНИЯ ГОРОДСКИХ ЛОКАЦИЙ НА ОСНОВЕ ДАННЫХ В СОЦИАЛЬНЫХ СЕТЯХ

Кремень Т.А. (федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

Научный руководитель – кандидат технических наук, директор ИДУ ИТМО Митягин С.А. (федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»**)**

В данном докладе представляются результаты работы по исследованию применения машинного обучения для определения географии городских объектов и локаций на основе публикаций и комментариев в социальных сетях. Рассматриваются проблемы анализа текстовых данных в социальных сетях при необходимости отображения упоминаемых в этих данных событий. Предлагается система, состоящая из компонентов сбора и обработки публикаций и комментариев в сообществах социальной сети VK, классификации текстовых адресов моделью на основе искусственного интеллекта, геокодирования текстовых адресов и создания файла формата geojson с координатами определенных событий.

Введение.

Стремительный рост использования социальных сетей в совокупности с легким доступом к большому объему данных вызывает интерес исследователей к анализу этих данных с помощью искусственного интеллекта для решения прикладных задач, таких как:

- Определение спама;
- Классификация тем сообщений пользователей;
- Классификация поведения пользователей;
- Создание системы рекомендаций;
- Создание систем распространения информации и продвижения товаров или контента.

Однако, при внушительном списке применения систем для анализа текста в социальных сетях на основе искусственного интеллекта, задачи определения географии упоминаемых в этих текстах городских объектов и локаций выходят за пределы возможностей одного лишь искусственного интеллекта. Зарубежные и отечественные исследователи и специалисты использую инструменты геокодирования для определения координат топонима по его адресу. Основными инструментами геокодирования за рубежом являются Google Geocoding Api и Nominatim (OSM Geocoding Api). В России к данным инструментам добавляется Yandex Geocoder Api.

Основная часть.

Для решения проблемы определения географии городских объектов и локаций из текстов в социальных сетях предлагается система, состоящая из следующих компонентов:

- Сбор и обработка записей и комментариев из социальной сети ВКонтакте;
- Модель на основе искусственного интеллекта для определения текстового адреса городских объектов и локаций;
- Геокодирование полученных адресов и преобразование координат в файл формата geojson.

Компонент сбора и обработки записей и комментариев из социальной сети VK использует VK API для запроса соответствующих данных при предоставлении аутентификационного токена. Из полученного ответа достаются поля с текстом записей и комментариев и объединяются в один массив.

Предложения из полученного массива обрабатываются моделью определения географии городских объектов и локаций, и классифицированные адреса добавляются в

результирующий массив для геокодирования. На основе анализа литературных источников в качестве метода классификации адресов предлагается использовать модель DeepPavlov, дообученную на подготовленных данных. Определение адреса осуществляется методов распознавания именованных сущностей (Named Entity Recognition — NER) при принадлежности к классу LOC с использованием тегов BIO (Beginning, Inside, Outside) для обозначения начала и середины именованной сущности класса LOC.

В качестве инструмента геокодирования предлагается использовать Google Geocoding Api из-за наличия исчерпывающей пользовательской документации и возможности получить координаты адресов в формате geojson.

Выводы. В рамках данного исследования предлагается использовать данную систему в работе сервиса Мониторинга потребностей жителей города в развитии инфраструктуры, разработанного Университетом ИТМО для определения координат адресов чрезвычайных происшествий в Санкт-Петербурге из соответствующего сообщества социальной сети ВКонтакте. Будет проведено сравнение результатов определения географии городских объектов и локаций с используемым в данный момент методом библиотеки Natasha, использующим парсер на основе правил.