УДК 004.048

РАЗРАБОТКА АЛГОРИТМОВ ДЛЯ АНАЛИЗА РАСПРЕДЕЛЕННОГО НАСТРОЕНИЯ ГОРОДСКОЙ СРЕДЫ НА ОСНОВЕ ОТКРЫТЫХ ДАННЫХ СОЦИАЛЬНЫХ СЕТЕЙ

Филатова А.А. (Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)
Научный руководитель – к.т.н, старший научный сотрудник Насонов Д.А.
(Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

Доклад посвящен описанию возможностей использования общедоступных открытых данных социальных сетей для анализа распределенного настроения городской среды. В докладе описываются результаты анализа текстового содержимого публикаций для экстремально популярных и непопулярных городских районов, описываются разработанные алгоритмы автоматического и полуавтоматического выявления фоновых событий, которые могут охарактеризовать общее настроение городской среды и событий, которые имеют непосредственное отношение к точкам притяжения, расположенным в популярных районах.

Введение. В современном мире общение посредством социальных сетей является неотъемлемой частью жизни большинства людей. Их привлекает возможность оперативно транслировать свои мысли, делиться мнением по важным для них вопросам и всегда быть на связи со своими близкими. В последнее время социальные сети стали полноценной заменой традиционным СМИ. Люди высказывают свое отношение к важным событиям, которые ежедневно происходят вокруг них, поэтому умение выделять такие полезные публикации среди большого количества информационного шума и их последующий качественный анализ позволят лучше понимать, что происходит в городской среде и, при необходимости, увеличить скорость реакции на такие события.

Основными задачами исследования являются: анализ открытых данных социальных сетей и последующее построение алгоритмов, которые позволят в автоматическом или полуавтоматическом режиме выделять события, которые формируют городской фон и события, которые происходят в популярных районах и связаны с известными точками притяжения.

Основная часть. В рамках исследования использовались открытые данные из социальной сети Instagram для г. Санкт-Петербург, собранные за 2019 год. Набор данных включал в себя порядка 7’760 тысяч публикаций с географическими координатами мест, к которым они относятся.

В начале был проведен пространственный и частотный анализ данных. В рамках него для Санкт-Петербурга был построена прямоугольная сетка, поделенная на квадраты 1 км на 1 км, и для каждого такого полигона было подсчитано количество публикаций за 2019 год, геометки которых относятся к этому полигону. После такого анализа стало понятно, что все городские районы можно поделить на три большие группы: районы, в которых количество публикаций за год очень мало (меньше 150), районы, количество публикаций в которых за год очень велико (больше 80 тыс.) и районы со средним количеством публикаций. Дальнейший анализ было решено проводить отдельно для каждой из полученных групп.

Перед тем, как начать анализ районов с малым количеством публикаций, был разработан алгоритм, который объединял районы, в которых количество публикаций было меньше 10, с соседними районами, в которым количество публикаций было наименьшим. После его применения образовалось 35 полигонов, количество публикаций за год в которых было от 10 до 150. Далее для каждого из таких полигонов был проведен анализ текстов публикаций с целью выявления основных тем, которые обсуждаются в каждом таком районе.

Анализ состоял из следующих шагов: предобработка текстов публикаций, которая включала в себя как стандартные шаги, так и дополнительные шаги, связанные со спецификой обработки текстов социальных сетей. Далее все тексты были векторизованы методом TF-IDF, и для каждого полигона был определен список наиболее значимых слов, из которого выбирались 10 самых значимых слов, которые и считались ключевыми для этого полигона. Такой несложный подход показал хорошие результаты и позволил сделать выводы о том, что все публикации в непопулярных районах можно условно поделить на 6 групп: географическое описание места; описание некоторого экономического объекта, расположенного в этой зоне (завод, порт, торговый центр); реклама, имеющая привязку к конкретной географической точке; реклама, не имеющая такую привязку и неповторяющиеся события с периодичностью порядка 1 года. Также был сделан вывод о том, что публикации в таких районах хорошо описывают городской фон и могут использоваться для его выявления.

Последующий анализ проводился для районов с экстремально большим количеством публикаций (более 80 тыс.). Каждый такой район был дополнительно поделен на меньшие области (25 квадратов 250 м на 250 м) и проанализирован вручную. Анализ показал, что практически все популярные районы имеют некоторый объект повышенного внимания пользователей, который является своеобразной точкой притяжения. Для дальнейшего анализа был выбран один из таких полигонов, имеющий известную точку притяжения – Ледовый дворец. Его анализ показал, что все публикации в этой области можно поделить на следующие группы: публикации, посвященные событиям, которые происходят непосредственно в точке притяжения; публикации, относящиеся к точке притяжения, но не описывающие происходящие там события (например, описание прогулки около нее) и публикации, которые не имеют отношения к точке притяжения (в основном, это реклама).

После анализа был разработан алгоритм, позволяющие отделить рекламные публикации от не рекламных. В его основу легла модель LDA. После этого был разработан алгоритм, который позволил научиться отделять события, которые происходят непосредственно в точке притяжения, от прочих. Он базируется на идее о том, что такие событийные публикации, в отличие от других публикаций, растянуты во времени. Поэтому все не рекламные публикации за 2019 год, относящиеся к исследуемому полигону, были поделены на 8760 часовых групп, после чего на таких группах была обучена модель LDA. Это позволило для каждых соседних часовых групп проверять, насколько они семантически похожи, и в случае, если их сходство велико, объединять их в одну группу. Границы процента схожести, необходимого для объединения групп были подобраны практическим путем. Также было совершено несколько итераций объединения групп с последовательным повышением этой границы. Это позволило добиться выделения протяженных во времени групп, содержащих ключевые слова, которые описывали практически все события, которые происходили на площадке Ледового дворца в 2019 году (результаты работы алгоритма были сравнены с реальным календарем мероприятий).

Выводы. В результате текущего исследования были выявлены городские районы с малым, средним и большим количеством постов, созданы и внедрены алгоритмы выделения основных тем публикаций в районах с небольшим количеством постов, разработаны и внедрены алгоритмы анализа районов с чрезвычайно большим количеством постов, включающие разделение публикаций в таких районах на событийные, рекламные и формирующие городской фон. В план ближайших работ входит разработка алгоритмов для определения комплексных событий, состоящих из атомарных событий и анализ районов со средним числом публикаций.

Филатова А.А. (автор) Подпись

Насонов Д.А. (научный руководитель) Подпись