УНИВЕРСИТЕТ ИТМО

НАЦИОНАЛЬНЫЙ ЦЕНТР
КОГНИТИВНЫХ РАЗРАБОТОК

НЦКР

ДУМАТЬ ЧТОБЫ ДЕЙСТВОВАТЬ

Профиль городских районов через призму событий

Анастасия Филатова
anastasia.a.filatova@gmail.com

Что нам может дать анализ социальных сетей

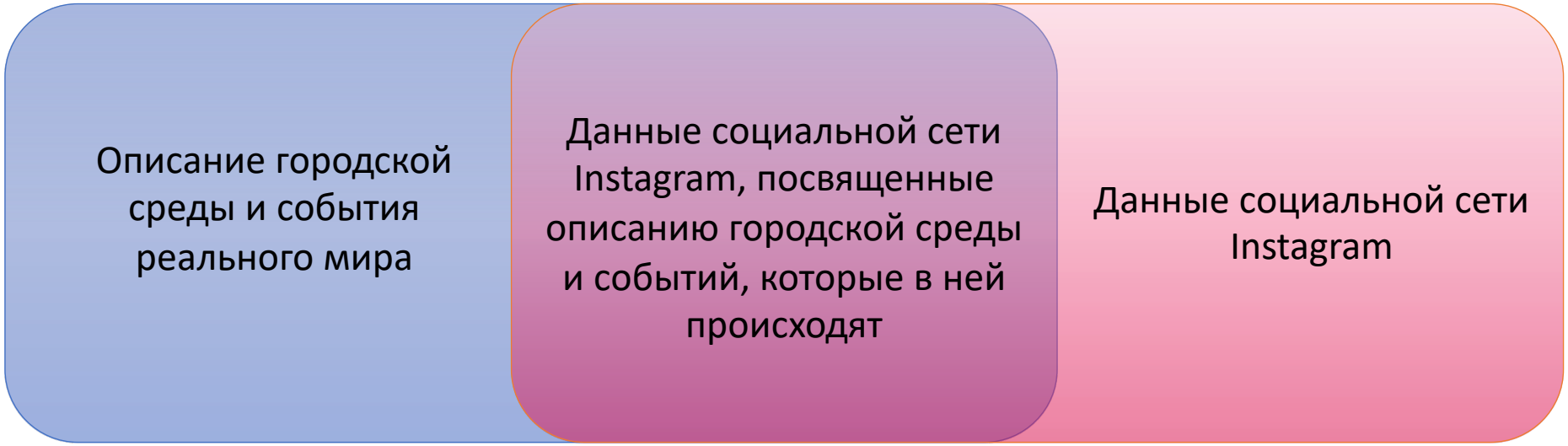
- Описания событий, происходящих в реальном мире и отношение людей к ним.
- Информацию о том, какие районы и места в городе популярны и непопулярны, и что определяет уровень популярности.
- Понимание того, как можно улучшить городские процессы, чтобы повысить уровень удовлетворенности жителей и гостей города.

Опять что-то строят, надоел этот шум, сил уже нет… 😤😡

Отличную площадку построили у нас во дворе! #лето #прогулка

Поучаствовали в художественном мастер-классе! 👧👧 #1июня #деньзащитыдетей

На какие глобальные вопросы хотим получить ответ



- Какую часть человеческой жизни покрывают данные социальной сети Instagram?
- Как из большого количества данных, публикуемых в Instagram получить релевантное описание городской среды?
- Можно ли получить комплексное описание городской среды по таким данным и насколько применимой окажется такая модель? Для решения каких практических задач ее можно будет применить?

Подход к анализу городской среды через построение профилей городских районов

НЦКР

Выявление релевантной информации, относящейся к анализу городской среды и характеристик, которые можно использовать в анализе

→

Выявление параметров, по которым можно разделить городские районы на группы

→

Классификация городских районов по выявленным параметрам

Использование особенностей каждого из выделенных классов для получения профилей городских районов и выявления событий, которые в них происходят

Особенности работы с данными социальной сети Instagram

- Большое количество информации для анализа.
- Большое количество характеристик, которые можно использовать в анализе (включая метаинформацию).
- Данные о событиях реального мира попадают в соцсети с минимальной задержкой.

- Далеко не все используют Instagram для обмена информацией о том, что происходит вокруг.
- Часть публикаций не содержит текстовую или мета информацию (хештеги, геолокационные метки).
- Очень сильная зашумленность данных.

НЦКР

Данные для анализа

Источник данных	Социальная сеть Instagram
Количество публикаций	7 760 тысяч
Язык	Русский (преимущественно)
Временной период	2019 год (Январь - Декабрь)
Географическая область	Санкт-Петербург, Россия

Текст публикаций + метаинформация

Дата и время

Географические координаты

НЦКР

Шаг 1 – Предобработка данных и выявление рекламных публикаций

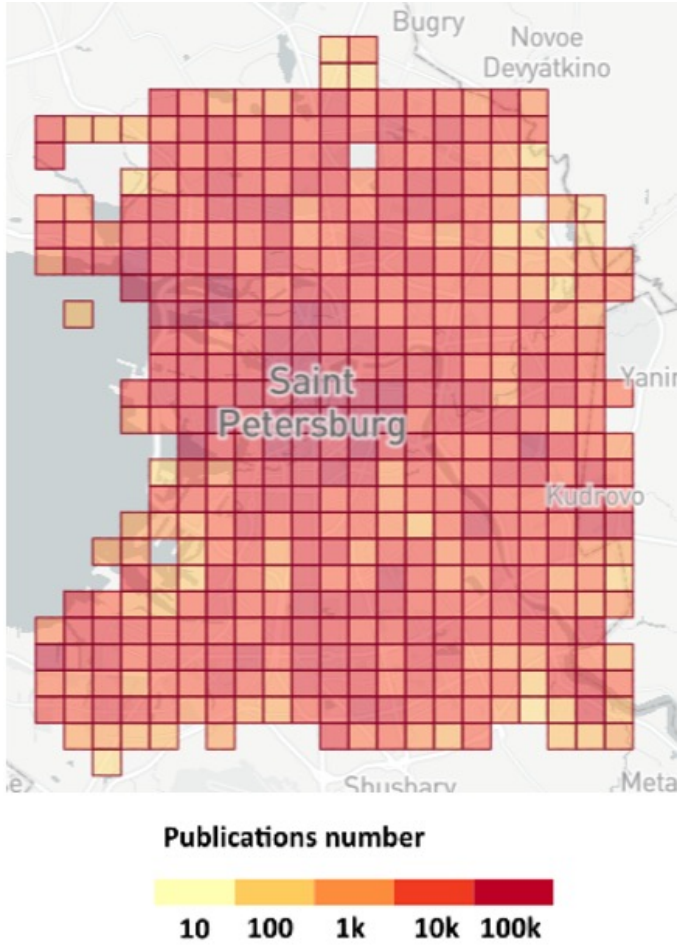
Предобработка данных

- Классическая предобработка текстовых данных: токенизация, удаление стоп-слов, лемматизация.
- Выделение мета-информации из текстов публикаций: хештегов, упоминаний и ссылок.

Выявление рекламных публикаций и пользователей-ботов

- Базу для выявления рекламных публикаций составила модель BigARTM с разделением на 2 темы, обученная на нормализованных текстах публикаций.
- Базовая модель была улучшена на основе анализа характеристик рекламных и нерекламных публикаций.
- Для выявления ботов анализировалась активность пользователей на протяжении года.

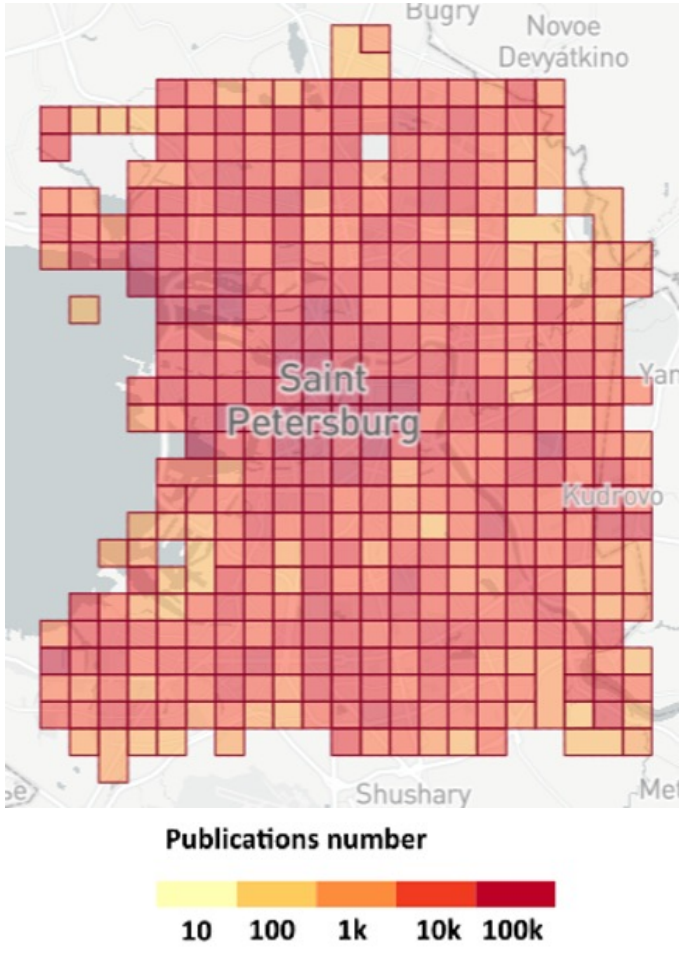
Шаг 2 – Геочастотный анализ публикаций



Непопулярные городские районы
(меньше 700 публикаций в год)

Районы среднего уровня популярности
(от 2.5 тыс до 65 тыс публикаций в год)

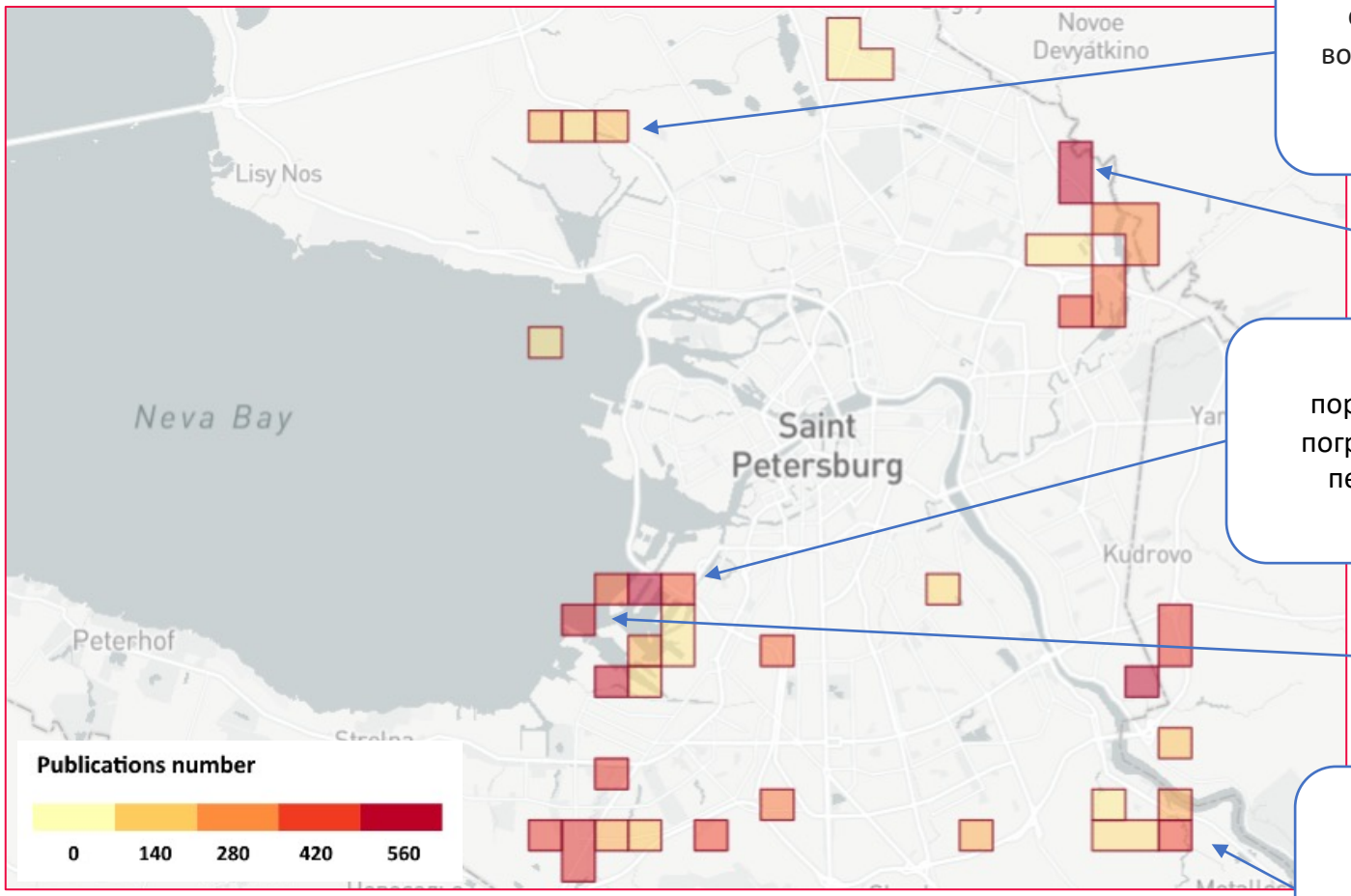
Популярные городские районы
(более 78 тыс публикаций в год)



Распределение количества публикаций по городским полигонам.
Слева: карта полигонов до объединения полигонов с малым количеством публикаций.
Справа: карта полигонов после объединения.

Шаг 3 – Анализ городских районов с разным уровнем популярности. Непопулярные районы

TF-IDF + выделение ключевых слов



Карта непопулярных городских районов (менее чем с 700 публикациями)

студия, ноготь, воплотить, мастер, юнтолово

производить, мебель, дверь, фабрика, шкаф

порт, партия, погрузка, груз, перевозка

тоннель, край, дикий, остров, канонерский

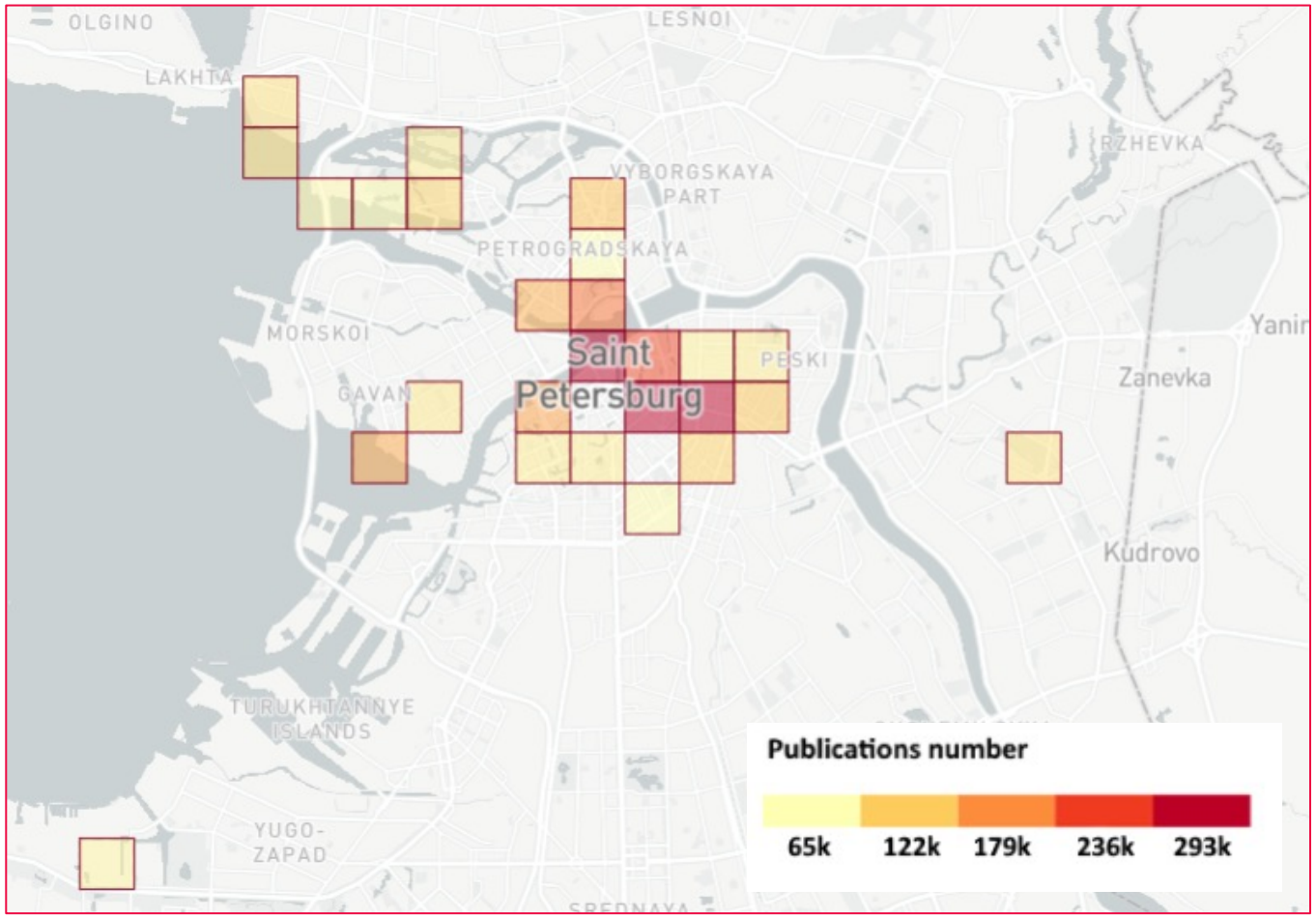
гдзс, команда, спасательный, соревнование, пожарный

Шаг 3 – Анализ городских районов с разным уровнем популярности. Популярные районы

Анализ популярных городских районов показал, что почти во всех таких областях присутствует объект, который является точкой притяжения

Группы публикаций

- Относящиеся к событиям в точке притяжения
- Относящиеся к точке притяжения, но не описывающие события
- Не относящиеся к точке притяжения, рекламные



Карта популярных городских районов (более чем с 78 тысячами публикаций)

Шаг 3 – Анализ городских районов с разным уровнем популярности. Районы со средним уровнем популярности

Районы со средним уровнем популярности – это, преимущественно, спальные районы.

Основные характеристики таких районов

- Нет явно выделяющихся точек притяжения, зато есть несколько локальных объектов, которые характерны конкретно для этого района.
- Большинство событий, происходящих в таких районах – это локальные события, связанные с такими локальными объектами.

НЦКР

Итог анализа профилей городских районов

Городские районы можно поделить на следующие глобальные группы

- Непопулярные районы – районы без собственной семантики и локальных точек притяжения. Характеризуются тем, что происходящие в них события – внешние.

- Районы со средним уровнем популярности – районы с очень сильной локальной семантикой. В них присутствует большое количество локальных объектов, которые задают основной контент публикаций и локальный сентимент, характерный для каждого конкретного района.

- Районы с высоким уровнем популярности – районы, характеризующиеся наличием глобальной (общегородской) точки притяжения, которая практически полностью задает семантику публикаций в районе и формирует сентимент, который также может распространяться на окрестные области.

НЦКР

Спасибо за внимание!