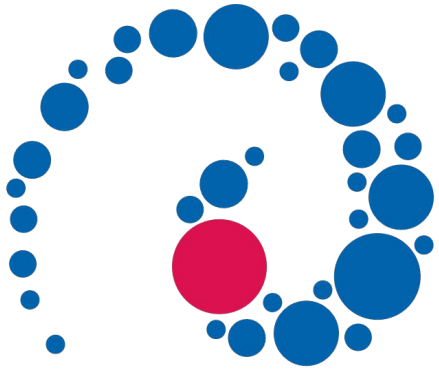
УНИВЕРСИТЕТ ИТМО



НАЦИОНАЛЬНЫЙ ЦЕНТР
КОГНИТИВНЫХ РАЗРАБОТОК

НЦКР

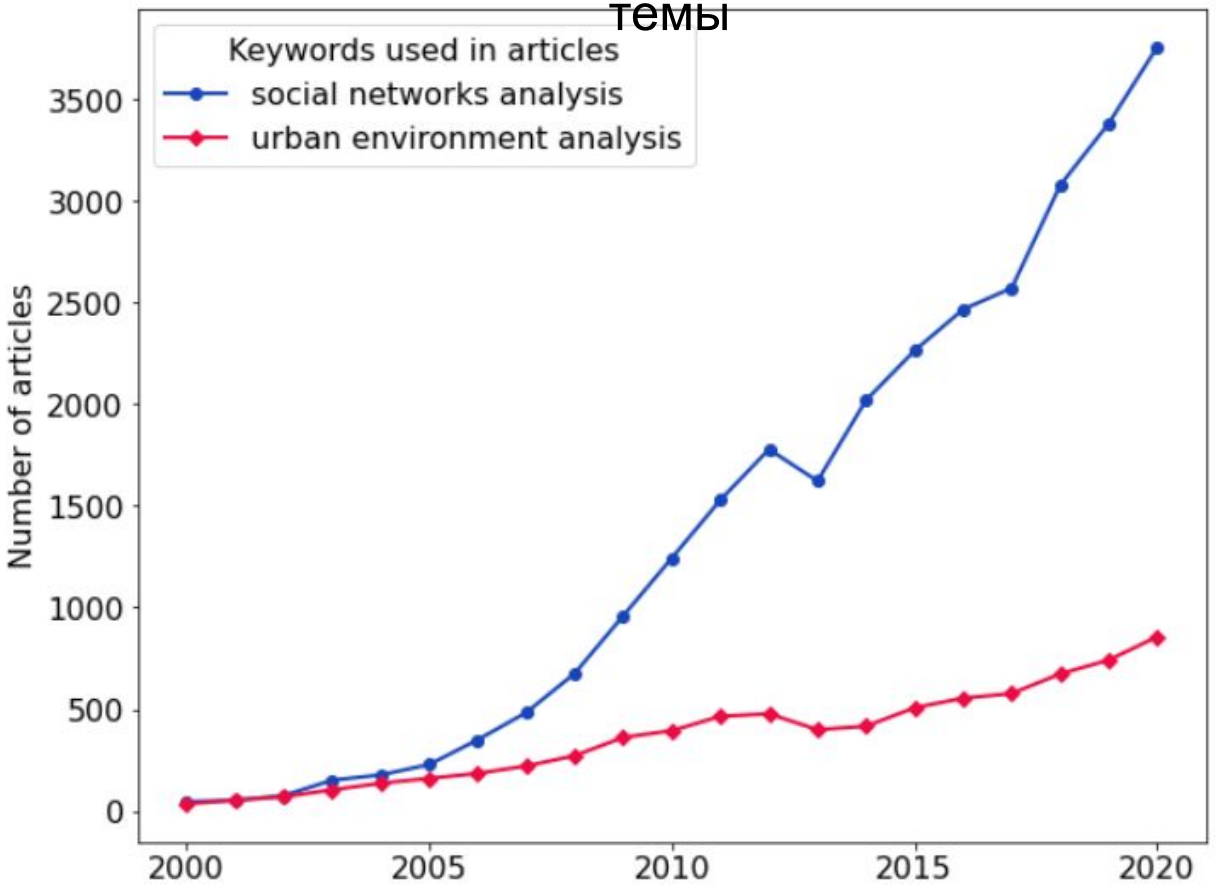
ДУМАТЬ ЧТОБЫ ДЕЙСТВОВАТЬ

Улучшенная кластеризация городских событий или при чем здесь Хештеги?

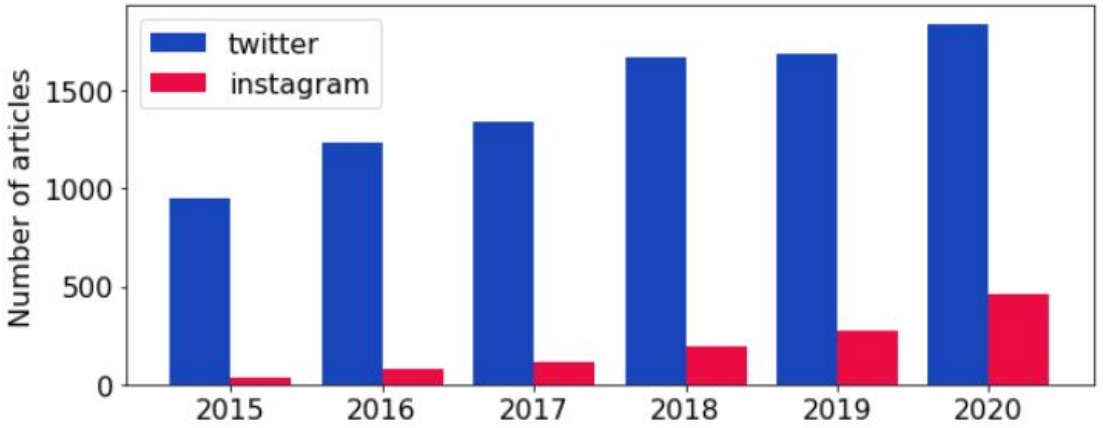
Ковальчук Михаил Андреевич
mkovalchuk@itmo.ru

Популярность Instagram

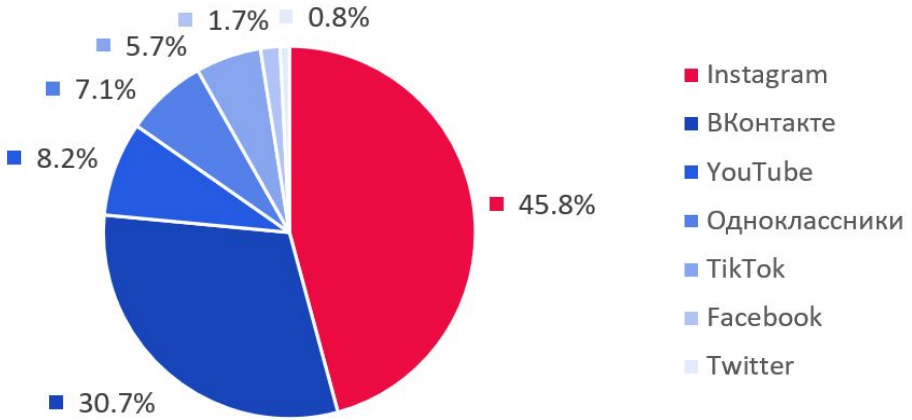
Динамика популярности ключевых слов темы



Упоминания социальных сетей в статьях



Популярность социальных сетей среди активных авторов в России осенью 2020 года



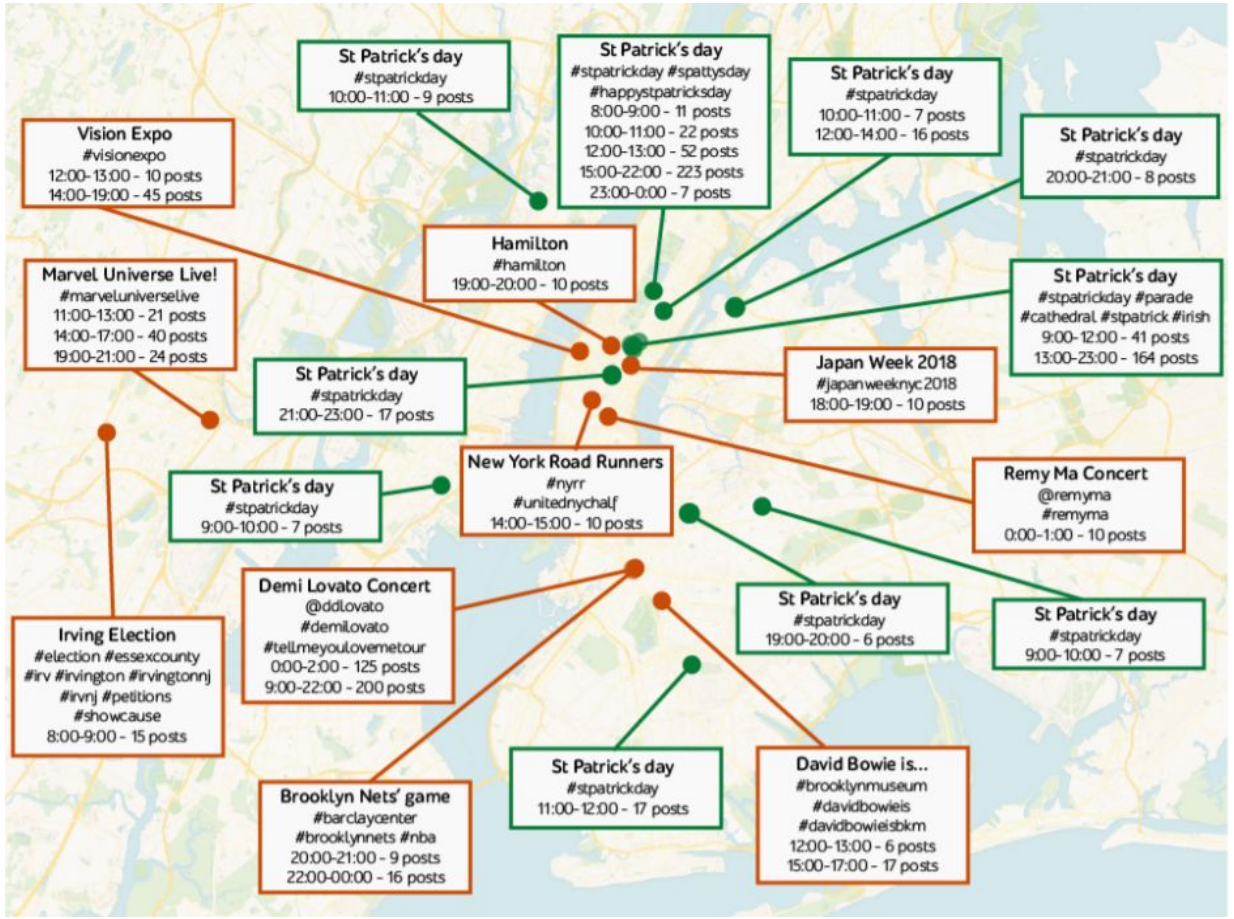
Sources for plots: www.scopus.com; www.rbc.ru

Мотивация

В современном мире городская жизнь все больше переходит в цифровой формат, что предоставляет значительные возможности для анализа с целью решения проблем современного общества через призму анализа больших данных.

Одним из перспективных направлений исследований является анализ городских событий для решения современных проблем городского жителя.

Это подводит нас к проблеме решения тематического моделирования.

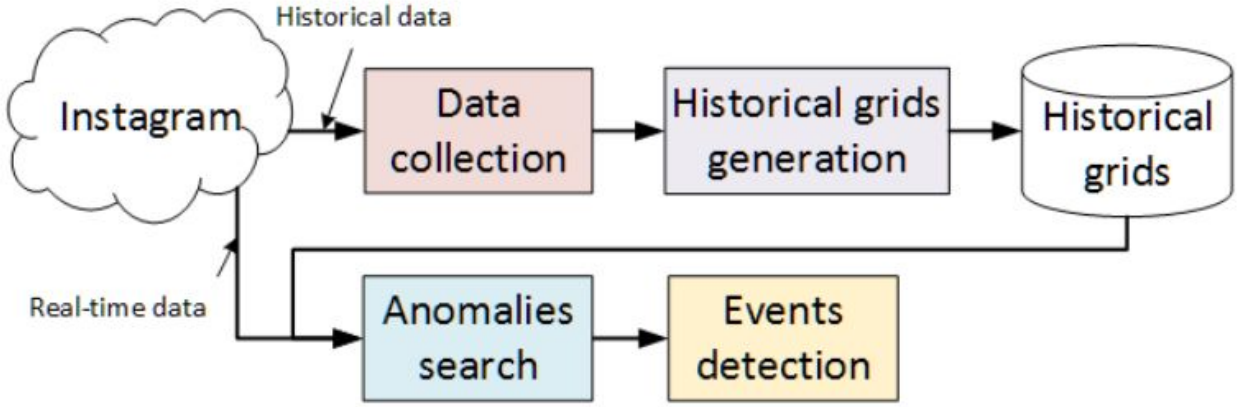


На карте примеры событий, обнаруженных для 17 марта 2018 года в Нью-Йорке, празднование Дня святого Патрика выделено зеленым цветом, все остальные - красным. События были обнаружены с помощью статистического подхода, который не использует семантику событий.

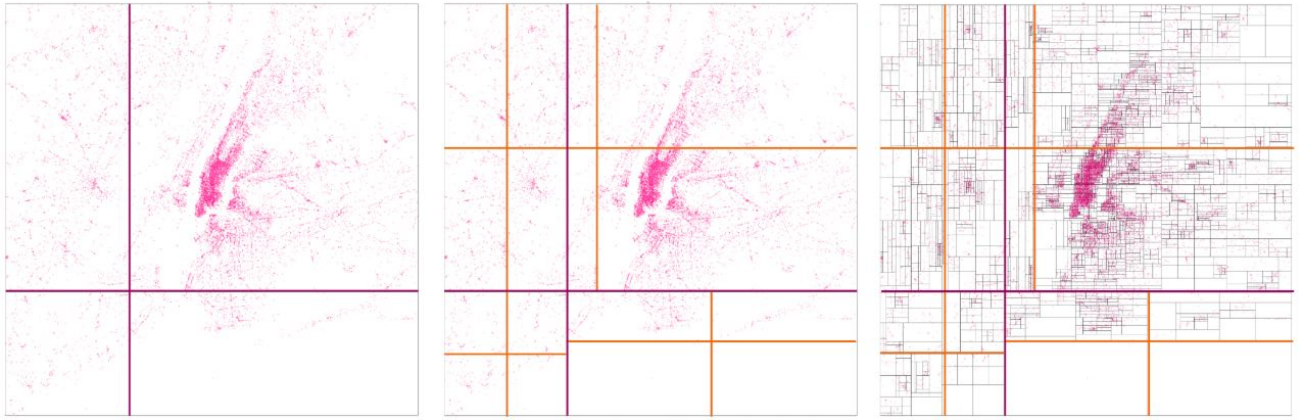
Частотный алгоритм обнаружения событий

Алгоритм:

- Построение исторических сеток за 1 год и более
- Сравнение исторической активности с текущей и обнаружение аномалий
- Использование тэгов для фильтрации кандидатов в события
- Объединение постов относящихся к одному событию

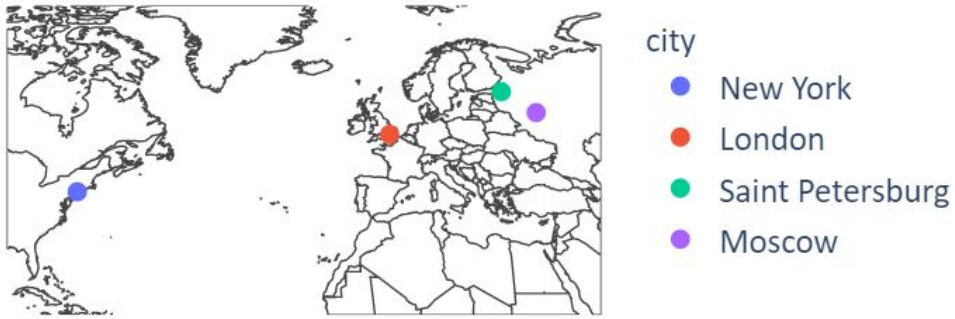


Pipeline алгоритма

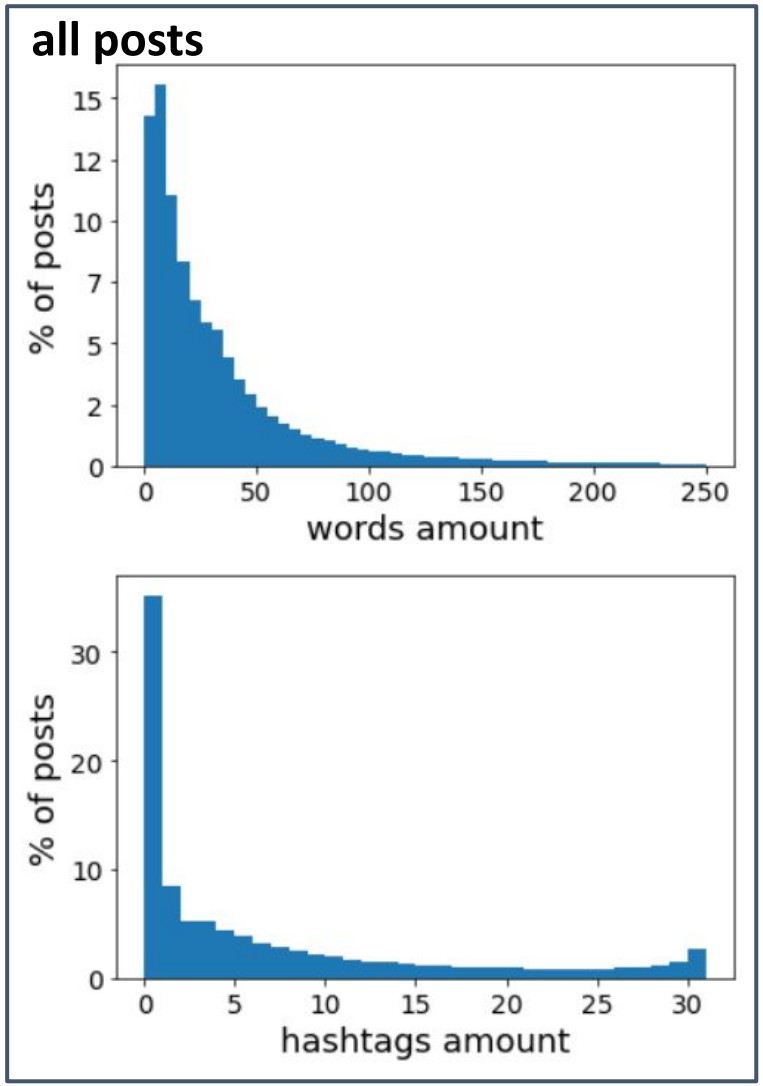


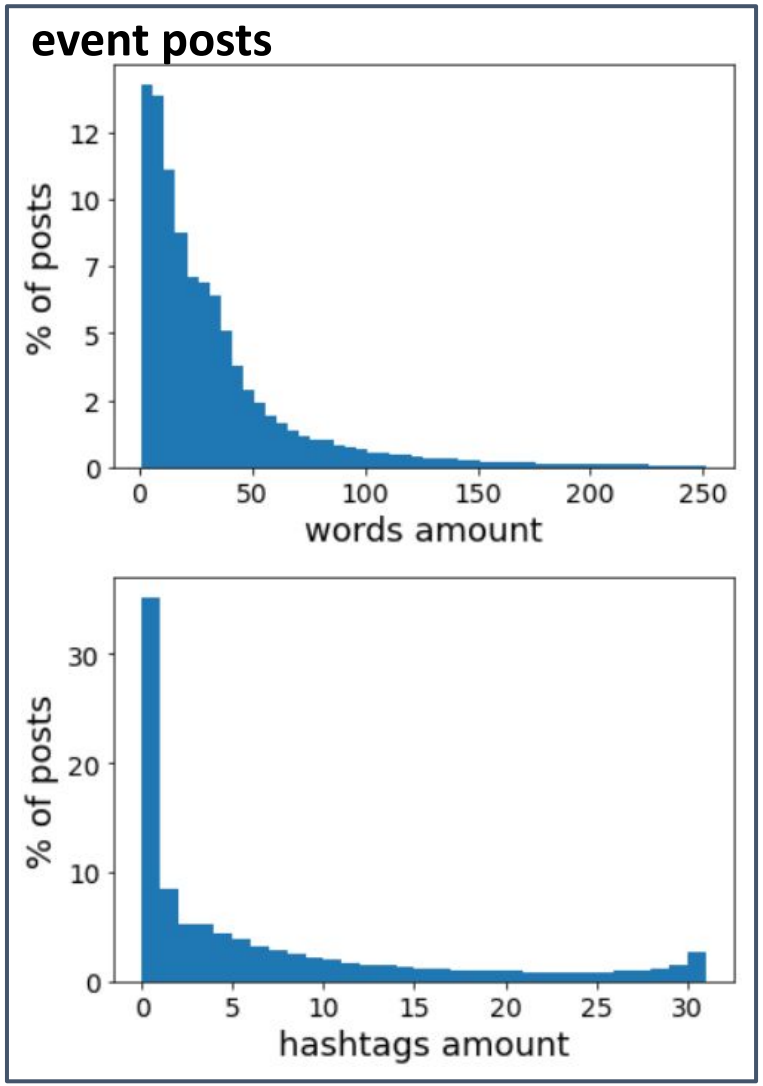
Построение сверточного квадродерева

Собранные данные



- 63 050 464 постов из Instagram из Москвы, Санкт-Петербурга, Лондона и Нью-Йорка.
- Около 2 414 216 266 словоупотреблений.
- 8284 события в России, 7954 в Нью-Йорке, 9231 в Лондоне на 2019 и начало 2020 года.
- Каждое событие - это в среднем 17 постов.
- Среднее количество слов и хэштегов:
  - во всех постах - 37 слов и 7 хэштегов;
  - в событиях - 36 слов и 10 хэштегов.





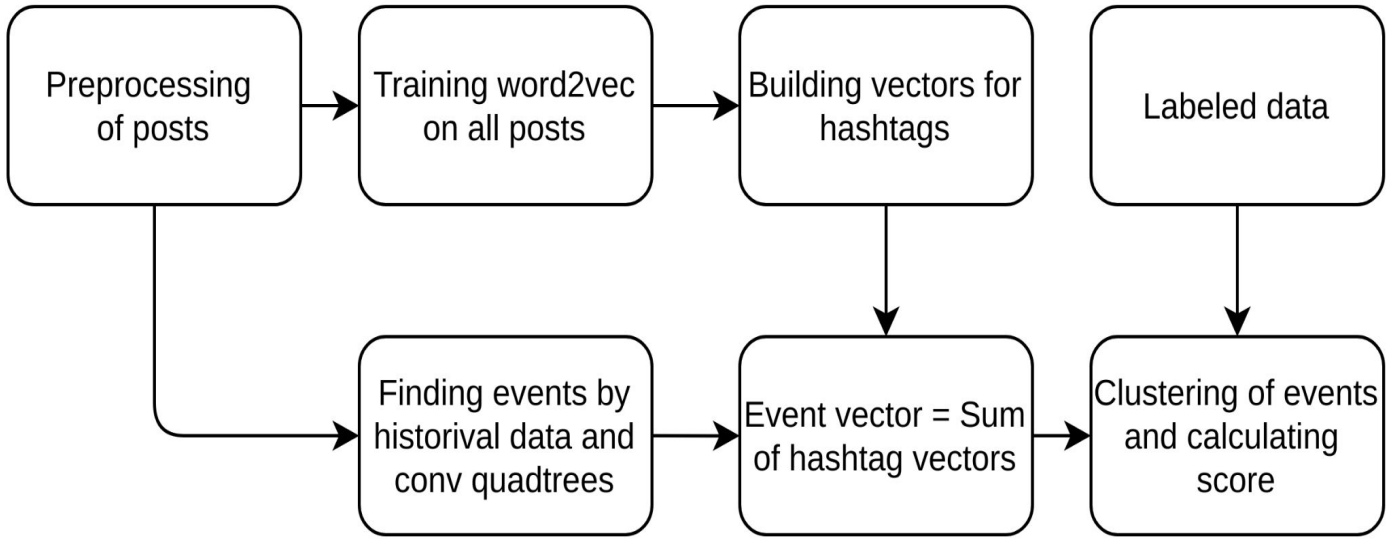
Плотность слов и хэштегов в сообщениях

Метод встраивания хэштегов

Встраивание хэштегов:

- Предположение - хэштеги содержат ключевую информацию в тексте.
- Тематика текста может быть охарактеризована содержащимися в нем хэштегами.

Семантический вектор текста - средняя сумма векторов хэштегов.



Pipeline алгоритма

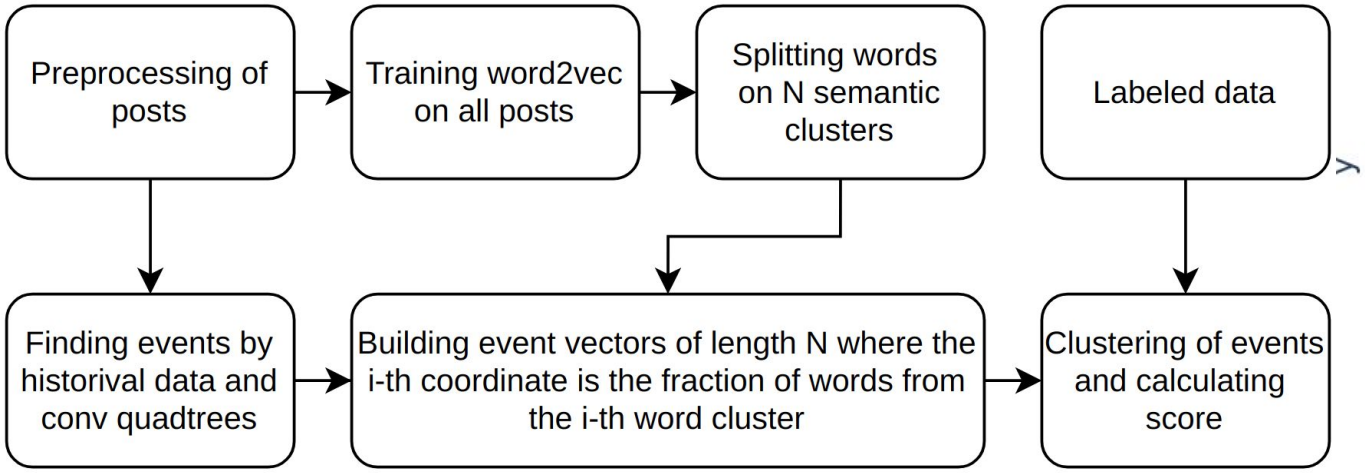
f0.5 оценка встраивания хэштегов

Лучшие результаты:

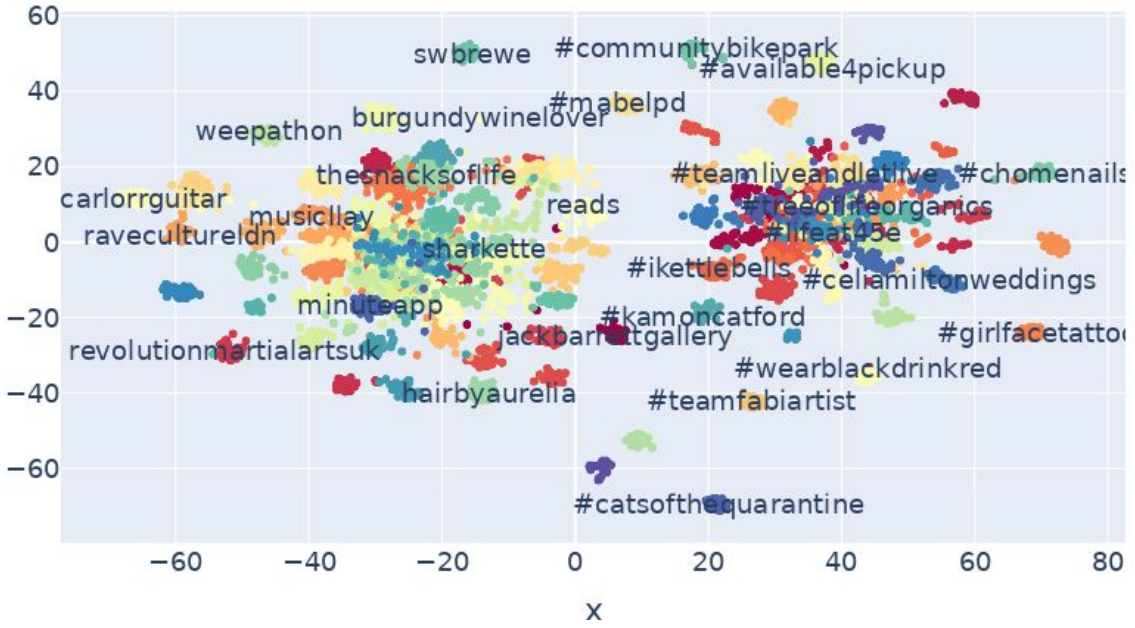
- **0.86** f0.5-score на 20 темах для “#hashtag”;
- **0.84** f0.5-score на 15 темах для “# hashtag”.

Метод семантических кластеров слов

- Использование векторов слов и векторов хэштегов для поиска семантики текста.
- Использование семантических кластеров слов для снижения размерности проблемы.
- Гибкая система настройки и выбора различных подходов.
- Специализированная обработка хэштегов.



Pipeline алгоритма



Семантические кластеры слов, для каждого кластера центральные 50 слов, название кластера - слово, наиболее близкое к центру, координаты получены с помощью word2vec и t-SNE

Метод семантических кластеров слов

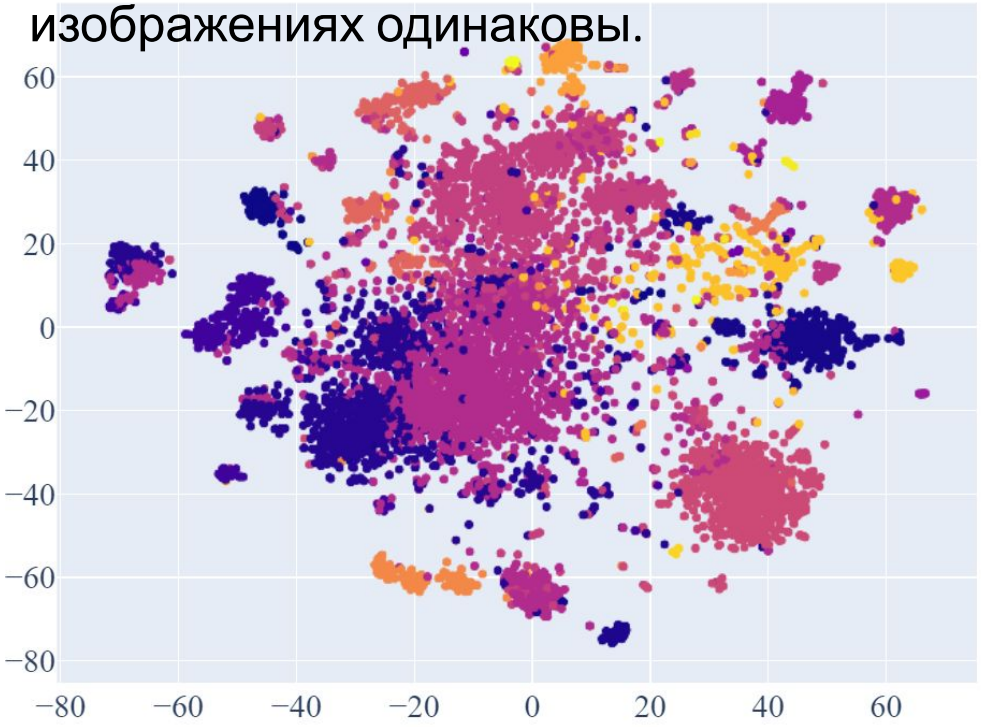
Лучший результат f-score для тематического моделирования событий для различных сценариев хэштегов

model type	f0.5 score	Number of topics for f0.5 score	f1 score	Number of topics for f1 score
‘#’ хэштег - служебная часть речи “# hashtag”	**0.90**	60	**0.88**	55
‘#’ игнорируется “hashtag”	0.86	50	0.83	50
removed all hashtags	0.81	90	0.78	90
only hashtags	0.79	110	0.70	55
trained on wiki	0.56	100	0.56	70

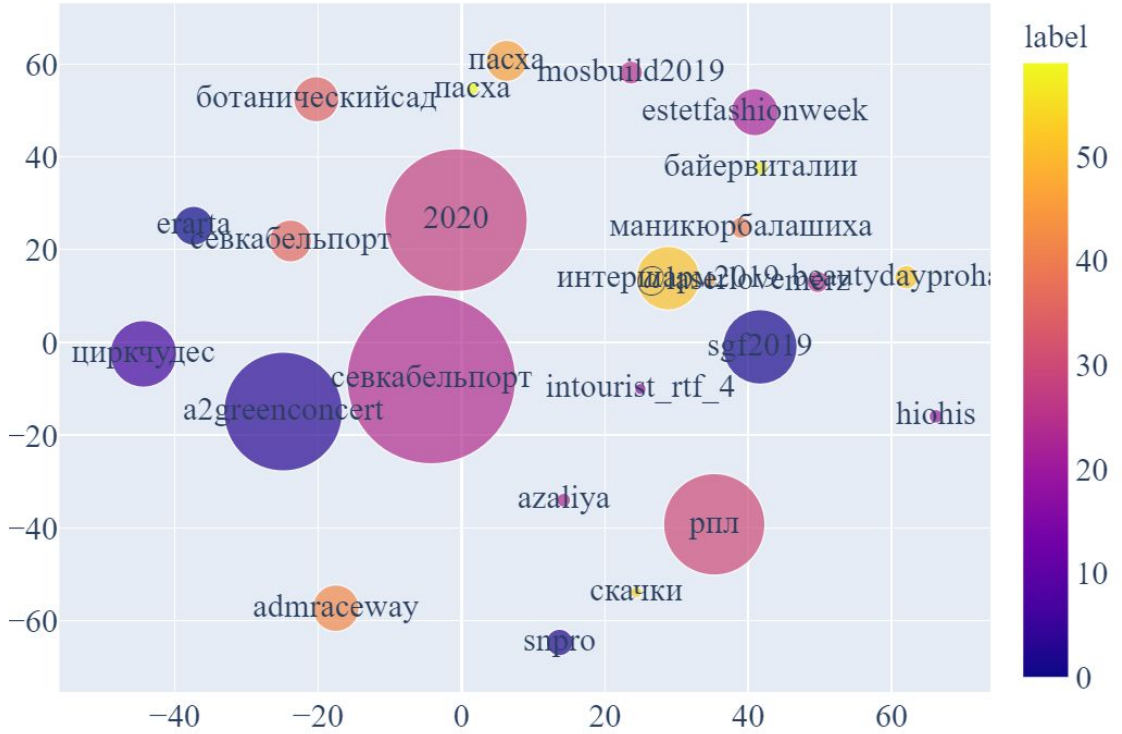
f0.5-score для сценариев хэштегов для метода для тематического моделирования событий

События Санкт-Петербурга и Москвы

Две визуализации тематического моделирования для метода кластеров семантических слов, 300 кластеров слов, 60 тем, f0.5-score 0.90. Цвета одинаковых кластеров на изображениях одинаковы.



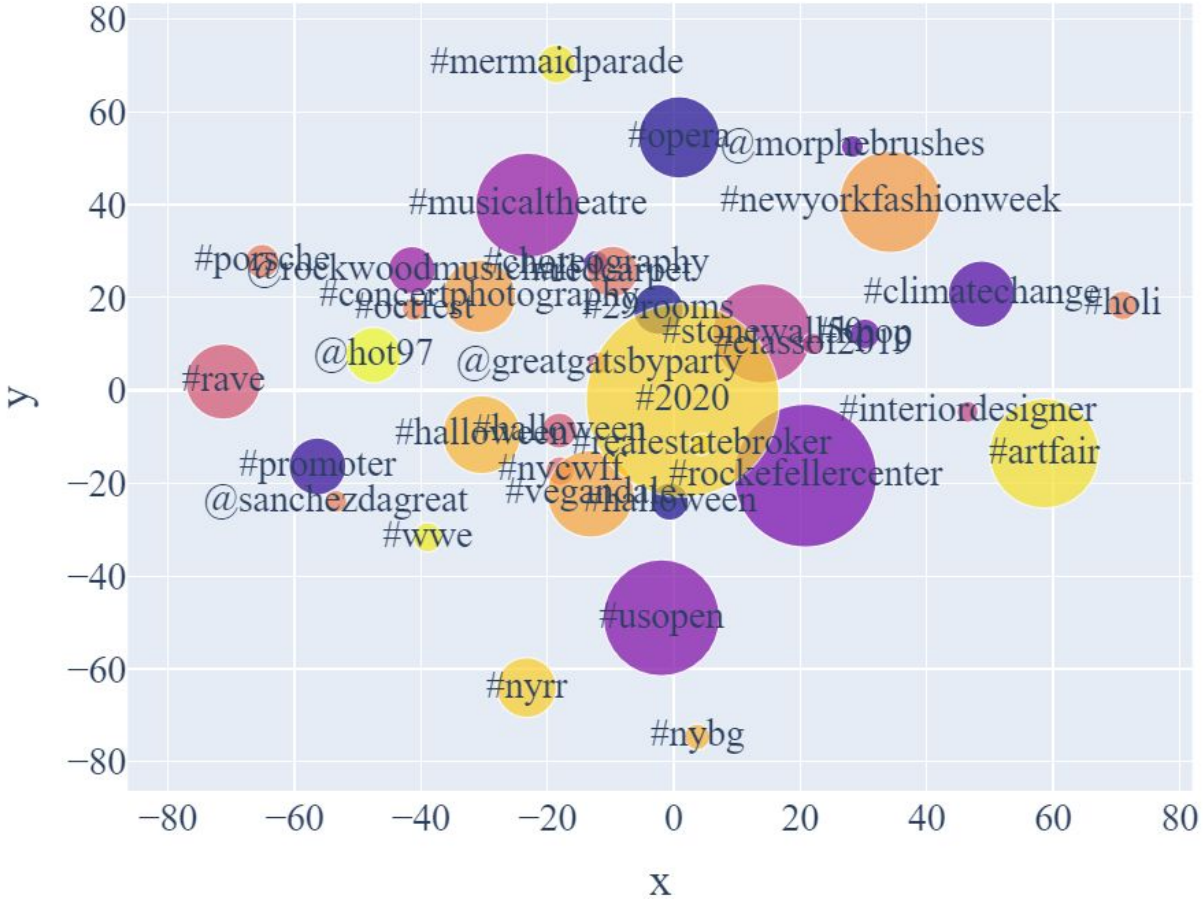
2d встраивание векторов событий полученные t-SNE, каждая точка - некоторые события, цвета - классы событий.



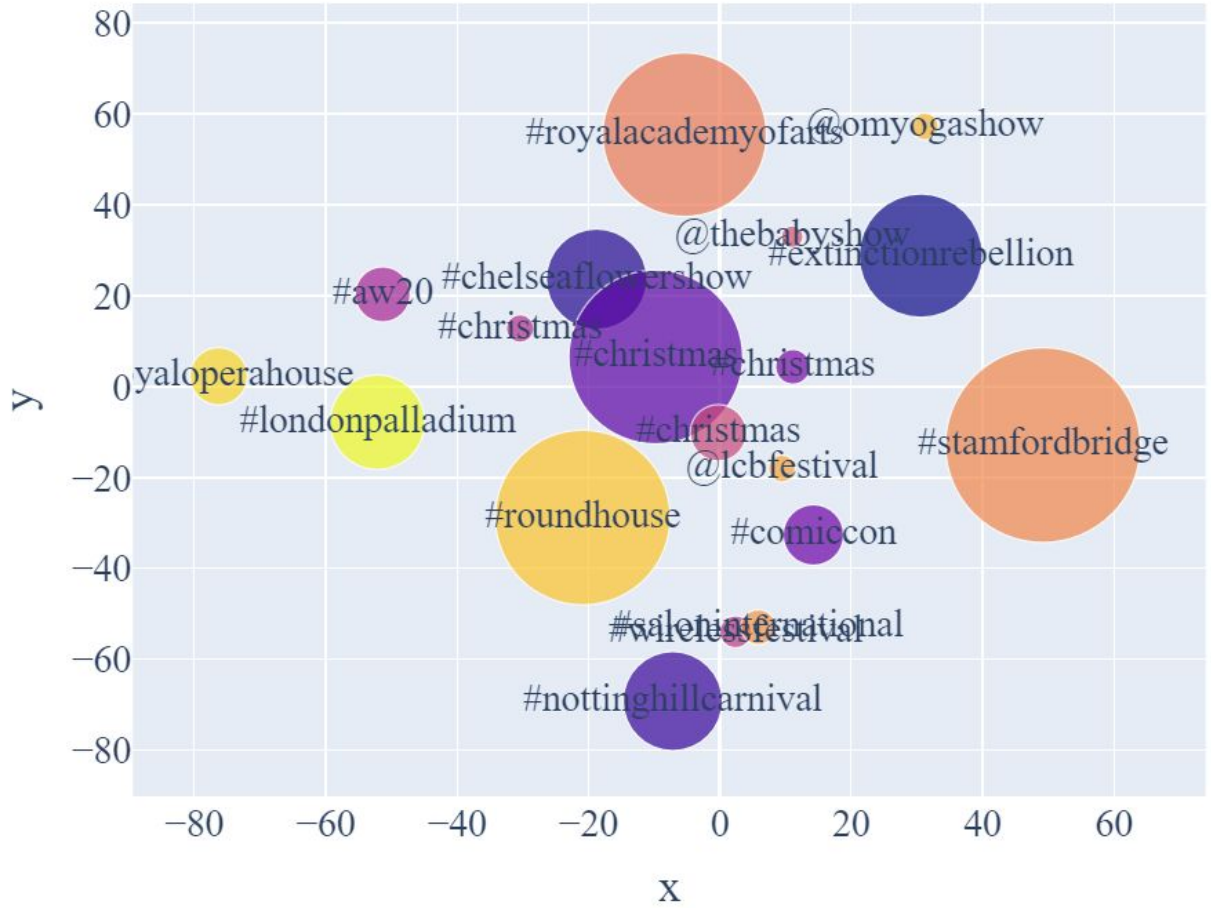
Центроиды, содержащие 20 и более событий с наиболее влиятельными хэштегами, размер точек пропорционален размеру кластеров.

Лондон и Нью Йорк

Результаты использования метода семантических кластеров слов для иностранных городов:



Нью Йорка

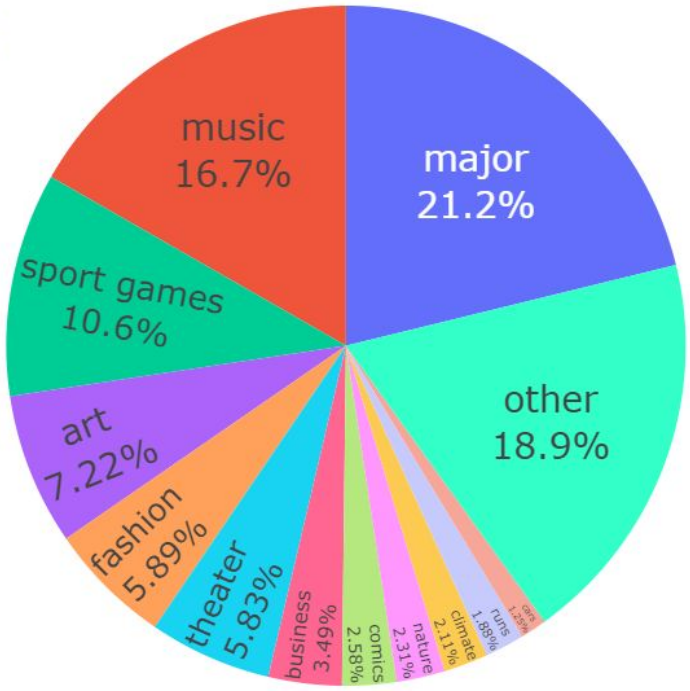


Лондон

Распределение тем событий

В России выделяются мероприятия инфобизнеса, хоккейные матчи и такие специфические события, как парад ледоколов и танковые игры.
В Лондоне популярен крикет, карнавал в Ноттингеме был сильно освещен в Instagram, там же широко распространены футбольные матчи и выставки искусства, дизайна и моды.
Нью-Йорк выделяется парадами в поддержку меньшинств, а музыкальные концерты и футбол гораздо менее популярны.

Распределение наиболее крупных классов событий по странам



Доли классов событий во всех четырех городах

НЦКР

Дальнейшие направления развития

- Область применимости использования хэштегов: события хорошо характеризуются хэштегами, для каких сущностей и объектов это правило верно?

- Обнаружение событий без использования исторических данных и частотного алгоритма на основе хэштегов, семантики и изображений.

- Создание обновляемого дерева знаний хэштегов с их определениями и семантическими векторами для анализа соцсетей.

Thank you for the attention!