Применение интеллектуальных методов для анализа новостных источников

Горюнов Егор, студент гр.БПМ2001
Научный руководитель: к.ф.-м.н., доцент Скородумова Е.А.

МТУСИ
21.06.2024

Постановка задачи

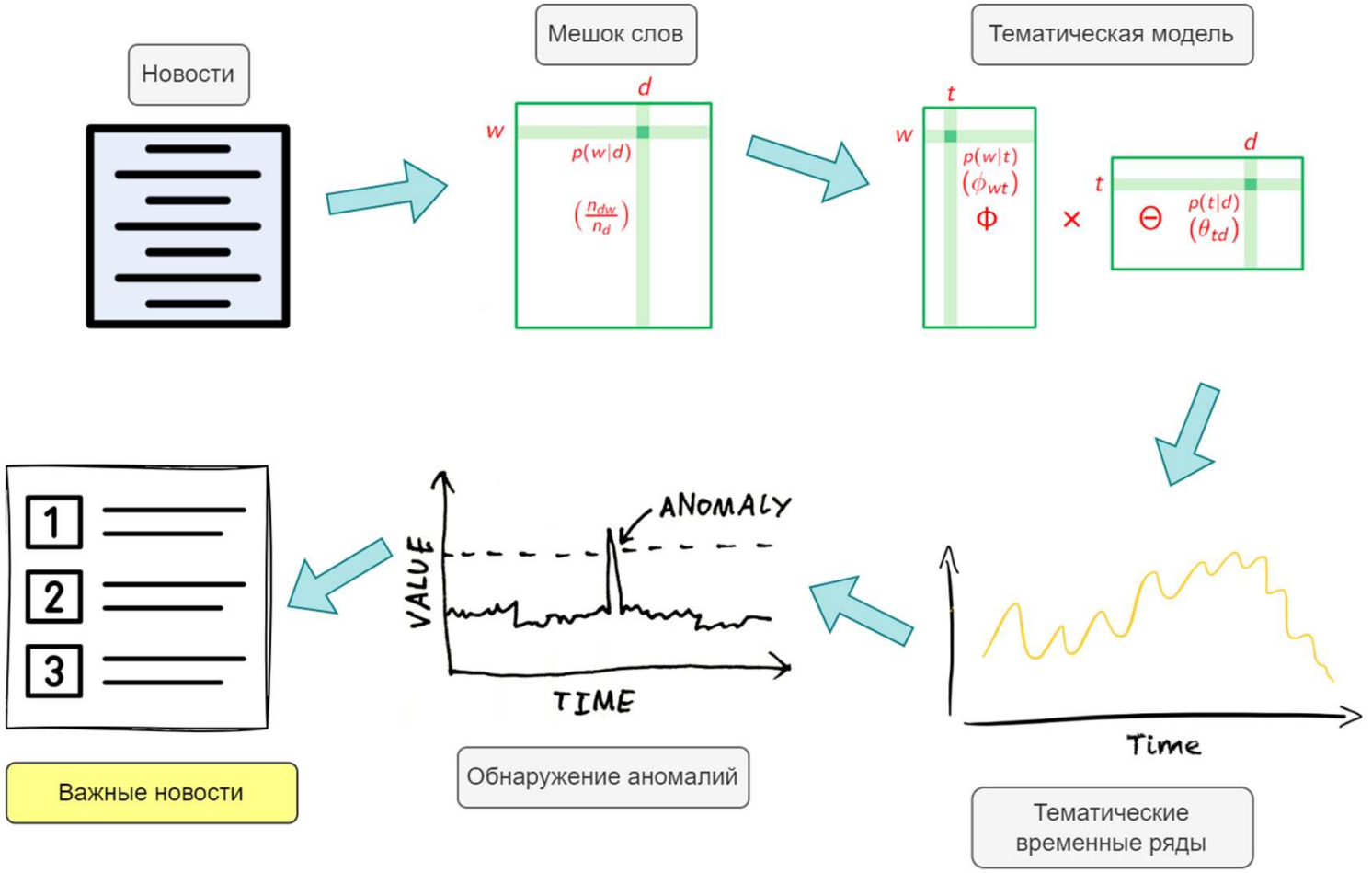
Актуальность: В условиях высокой информационной нагрузки сложно оставаться в курсе важных событий.

Цель: Обнаружения важных событий в огромном потоке новостей.

Проблема: Понятие «важного события» трудно формализуемо.

Решение: Предложить подход на основе интеллектуальных методов (тематическое моделирование и обнаружение аномалий во временных рядах).

Этапы исследования



Данные

Датасет с данными по около 100 000 Телеграм-каналам.

Источник: *La Morgia, Massimo & Mei, Alessandro & Mongardini, Alberto. (2023). TGDataset: a Collection of Over One Hundred Thousand Telegram Channels*.

Мотивация:

- ✓ Оперативность
- ✓ Разнообразие источников
- ✓ Объём данных
- ✓ Неофициальный контент

Предобработка данных

01

Отбор русскоязычных новостных телеграм-каналов.

02

Очистка и нормализация текстов

03

Токенизация и векторизация

Векторизация мешка слов:

1. the red dog
2. cat eats dog
3. dog eats food
4. red cat eats

the	red	dog	cat	eats	food
1	1	1	0	0	0
0	0	1	1	1	0
0	0	1	0	1	1
0	1	0	1	1	0

Задача тематического моделирования

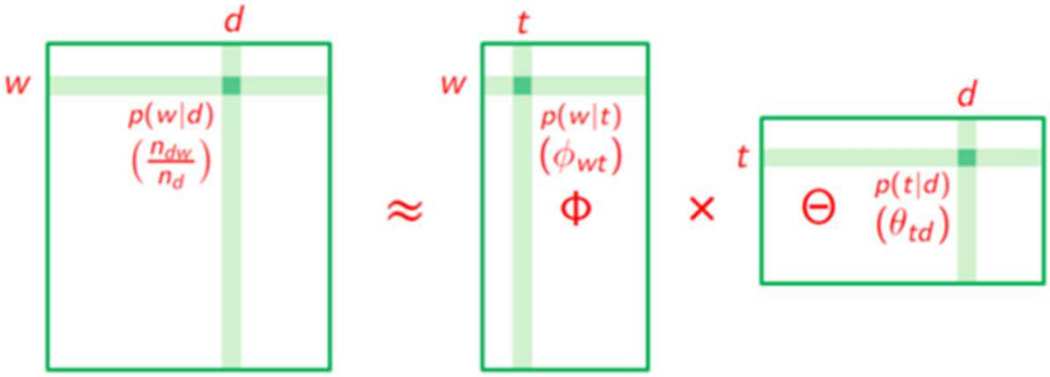
Дано: коллекция текстовых документов

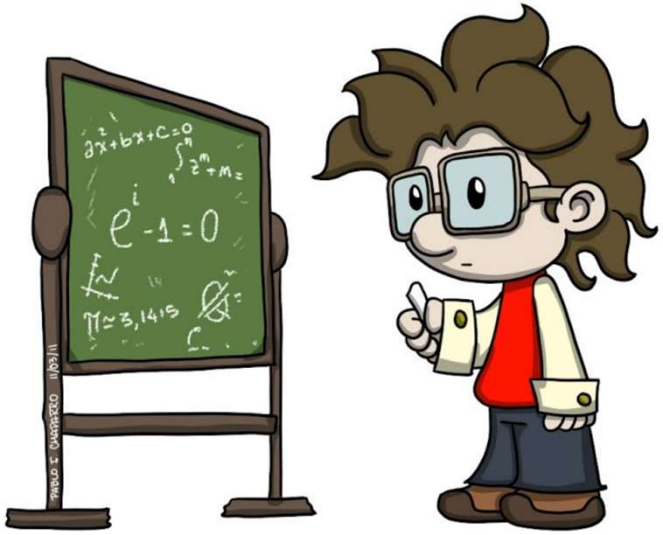
- W — конечное множество термов (слов, токенов)
- D — конечное множество документов
- n_{dw} — частота терма w в документе d

Найти: вероятностную тематическую модель

$$p(w|d) = \sum_{t \in T} p(w|\cancel{d}, t)\, p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$$

где $\phi_{wt} = p(w|t)$, $\theta_{td} = p(t|d)$ — параметры модели





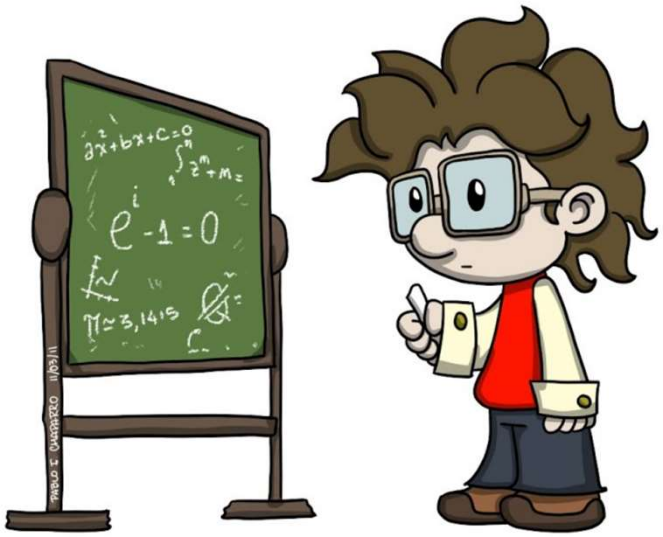
Аддитивная регуляризация тематических моделей (ARTM)

Максимизация логарифма правдоподобия с регуляризатором:

$$\sum_{d,w} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}; \qquad R(\Phi,\Theta) = \sum_i \tau_i R_i(\Phi,\Theta)$$

метод простой итерации для системы уравнений

$$\begin{cases} p_{tdw} \equiv p(t|d,w) = \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right) \\ \phi_{wt} = \underset{w \in W}{\text{norm}}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} \right), \quad n_{wt} = \sum_{d \in D} n_{dw} p_{tdw} \\ \theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}} \right), \quad n_{td} = \sum_{w \in d} n_{dw} p_{tdw} \end{cases}$$



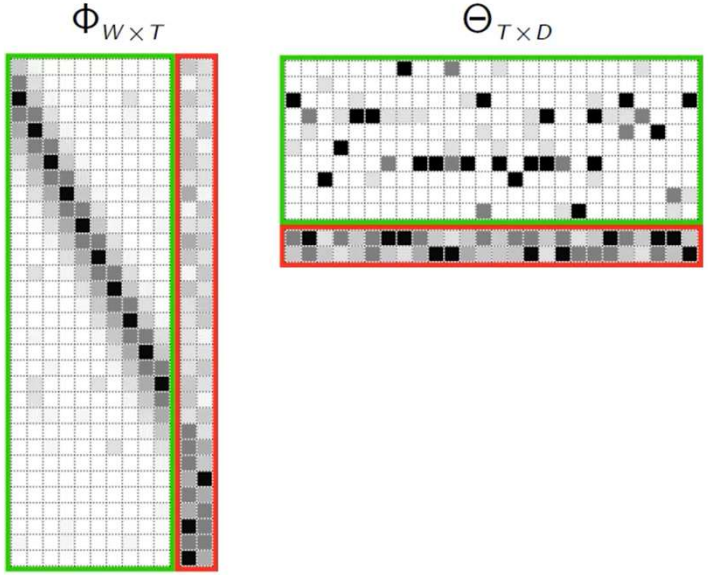
Регуляризаторы ARTM

1. Регуляризатор сглаживания и разреживания:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T} \sum_{w \in W} \beta_{wt} \ln \phi_{wt} + \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_{td} \ln \theta_{td} \to \max,$$

где $\beta_0 > 0$, $\alpha_0 > 0$ — коэффициенты регуляризации,
β_{wt}, α_{td} — параметры, задаваемые пользователем:

- $\beta_{wt} > 0$, $\alpha_{td} > 0$ — сглаживание
- $\beta_{wt} < 0$, $\alpha_{td} < 0$ — разреживание



2. Регуляризатор декоррелирования:

Минимизируем ковариации между вектор-столбцами ϕ_t:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \phi_{wt} \phi_{ws} \to \max.$$

Примеры тем, выделенных с помощью ARTM

```
topic_0:  ['банк', 'деньги', 'счет', 'финансовый', 'покупка', 'клиент', 'карта']
topic_1:  ['путин', 'президент', 'владимир', 'зеленский', 'встреча', 'песок', 'кремль']
topic_2:  ['маска', 'ведущий', 'илона', 'маск', 'твиттер', 'twitter', 'собчак']
topic_3:  ['операция', 'турция', 'специальный', 'карта', 'украина', 'турецкий', 'готовность']
topic_4:  ['цена', 'товар', 'магазин', 'продукт', 'рынок', 'топливо', 'изза']
topic_5:  ['сша', 'американский', 'байден', 'белый', 'посол', 'дипломат', 'штат']
topic_6:  ['донбасс', 'срок', 'денис', 'реальный', 'милиция', 'ростовский', 'короткий']
topic_7:  ['акция', 'поддержка', 'участие', 'проходить', 'участник', 'участвовать', 'мероприятие']
topic_8:  ['компания', 'совет', 'директор', 'руководитель', 'пост', 'председатель', 'глава']
topic_9:  ['пытка', 'заключать', 'скорость', 'колония', 'фсин', 'правозащитник', 'шахта']
topic_10:  ['москва', 'новость', 'риа', 'ресторан', 'регион', 'городской', 'москвич']
topic_11:  ['россия', 'приостанавливать', 'компания', 'прекращать', 'бренд', 'продажа', 'производство']
topic_12:  ['данные', 'база', 'номер', 'адрес', 'яндекс', 'заказ', 'доставка']
topic_13:  ['одежда', 'пол', 'журнал', 'женский', 'мужской', 'владислав', 'красота']
topic_14:  ['сайт', 'заблокировать', 'требование', 'ведомство', 'роскомнадзор', 'говориться', 'фейк']
topic_15:  ['спасать', 'животное', 'собака', 'хозяин', 'кот', 'счастие', 'застревать']
topic_16:  ['иностранный', 'выполнять', 'агент', 'функция', 'лицо', 'данный', 'сообщение']
topic_17:  ['город', 'улица', 'здание', 'протест', 'взрыв', 'митинг', 'центр']
topic_18:  ['российский', 'информация', 'источник', 'июль', 'тасс', 'созданой', 'опровергать']
topic_19:  ['случай', 'сутки', 'коронавирус', 'выявлять', 'новый', 'заболевать', 'умирать']
topic_20:  ['переговоры', 'франция', 'процесс', 'состояться', 'макрон', 'прекращение', 'делегация']
```

Построение тематических временных рядов

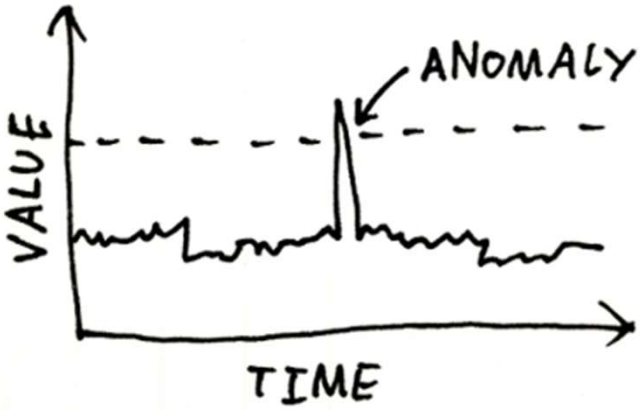
Обнаружение тем-событий

Назовём тему, для которой на данный момент времени обнаружен скачок, темой-событием.

Алгоритмы обнаружения аномалий

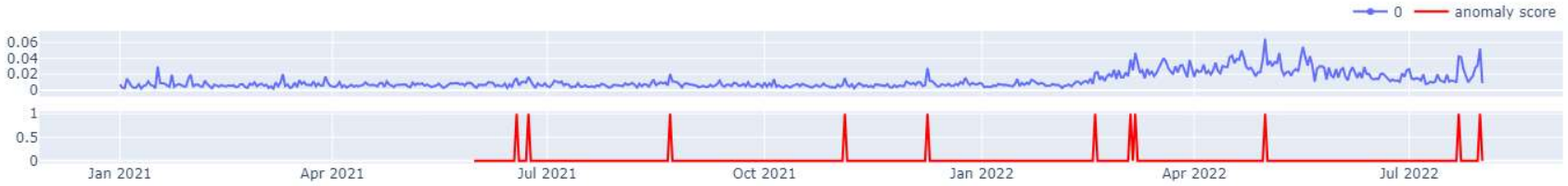
В работе были использованы следующие алгоритмы из библиотеки StreamAD:

- KNNDetector() – детектор на основе алгоритма kNN-CAD;
- SRDetector() – детектор на основе алгоритма SR;
- ZScoreDetector() – детектор на основе z-статистики;
- OCSVMDetector() – детектор на основе One-Class SVM;
- MadDetector() – детектор на основе MAD;
- SArimaDetector() – детектор на основе ошибки предсказания;
- SpotDetector() – детектор на основе алгоритма SPOT.

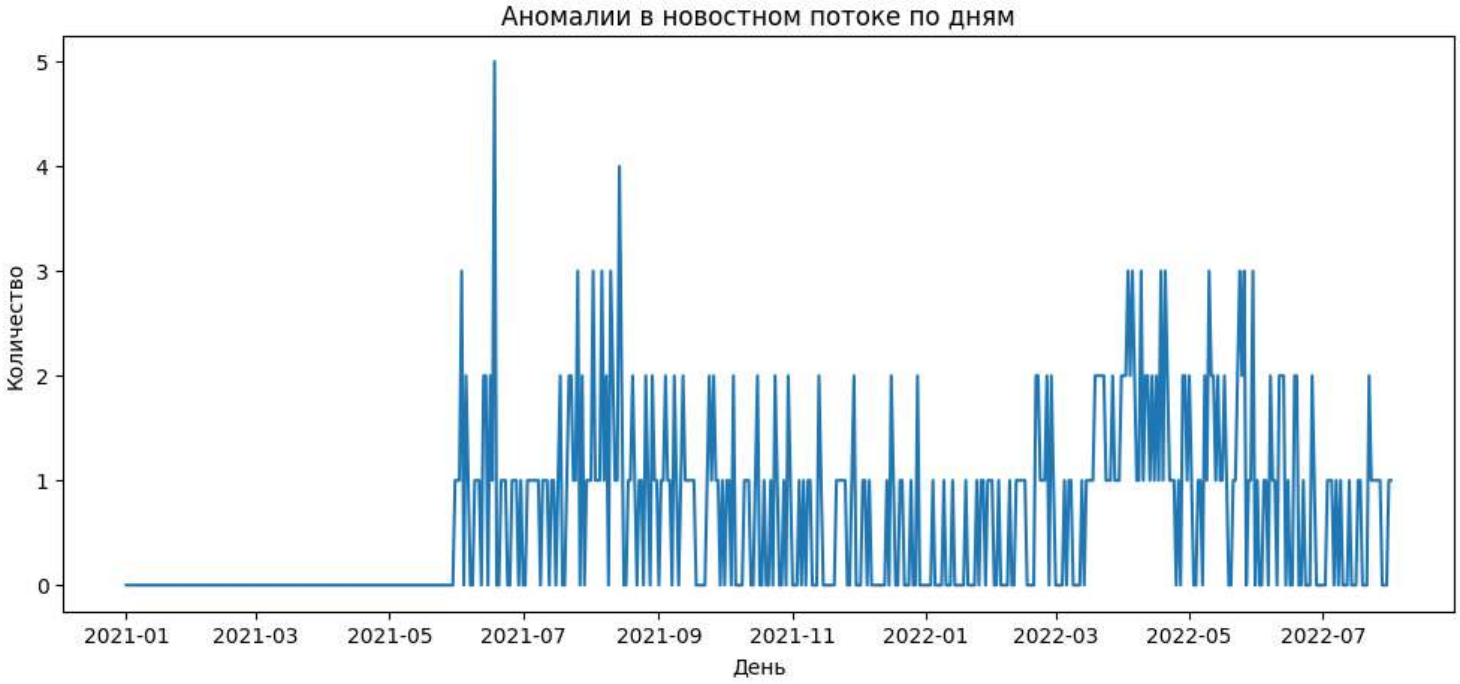


Алгоритмы составляют ансамбль.

Работа ансамбля:



Важные новости есть не всегда



Пример обнаружения «горячей» новости

Топ-3 новостей по вкладу за 2022-03-25 (1 тема-событие):
topic_52: ['**дело**', '**уголовный**', '**возбуждать**', 'квартира', 'факт', 'расследование', 'статья']

1) «На ингушскую журналистку Изабеллу Евлоеву возбудили уголовное дело по статье о фейках про российскую армию. Причиной стали посты в телеграм-канале. По данным «Базы», сотрудники ингушского Центра по противодействию экстремизму МВД и …»

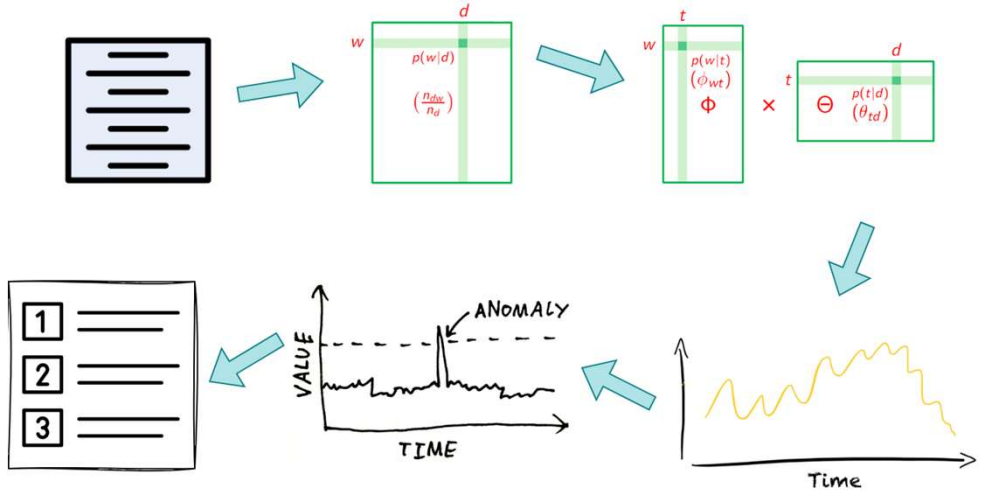
2) «Останкинский суд Москвы зарегистрировал протокол по делу о «дискредитации» российский военных на экс-редактора Первого канала Марину Овсянникову, которая 14 марта в эфире программы «Время» за спиной ведущей Екатерины Андреевой вышла с пацифистским плакатом про Украину. На Овсянникову составили протокол. Ей грозит штраф от 30 до 50 тысяч рублей.»

3) «На бывшую журналистку Первого канала Марину Овсянникову, которая вышла с антивоенным плакатом в прямой эфир, возбудили ещё одно административное дело. На этот раз — по новой статье о дискредитации ВС РФ. По статье 20.3.3 КоАП РФ Овсянникова может получить штраф от 30 до 50 тысяч рублей. Дело на неё зарегистрировали 24 марта. Ранее журналистка — за призыв выходить на антивоенные митинги, который судья увидел в её видеообращении, записанном ещё до выхода с плакатом.»

Заключение

- Проведена качественная **предобработка** текстовых данных;
- Использован метод тематического моделирования **ARTM** для построения **векторных представлений** новостей;
- Использованы методы **анализа временных рядов** для выявления **тем-событий**;
- Разработан и применён алгоритм, который **успешно выявил конкретные новости**, ставшие источниками обнаруженных тем-событий.

В совокупности, проведенное исследование **продемонстрировало работоспособность** алгоритмов для **обнаружения важных новостей** и тем в русскоязычных телеграм-каналах.



Спасибо за внимание!