## Средний тренд роста публикаций (число всех статей по годам).

Среднее количество публикаций в год: 40000 по всем статьям.

Не один из топиков не превышает данный показатель.

Это связано с распределением документов в топиках. Данные содержат выбросы или существенно искажены (например, несколько топиков с очень высоким количеством публикаций 9000, в то время как большинство топиков имеют значительно меньше публикаций 600)

В связи с этим, я решил применить анализ изменения количества публикаций по темам по годам, чтобы определить, какие темы становятся более или менее популярными, и тем самым получить более точное представление о динамике интереса к различным темам. Это позволило выявить темы с наибольшим ростом и снижением интереса, что является более релевантным показателем текущих научных тенденций, чем простое сравнение абсолютных объемов публикаций.

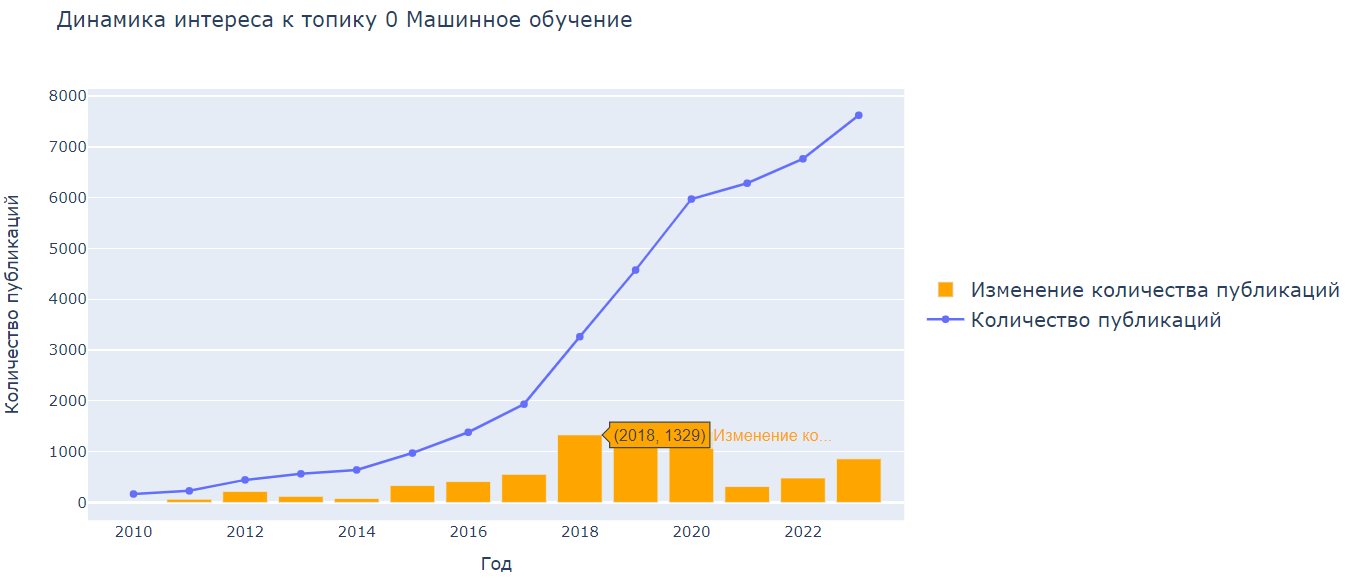


график динамики интереса к топику 0

Динамика количества публикаций по годам

Таблица для 0 топика

| **Год** | **Кол-во Публикаций за год** | **Изменение Кол-ва Публикаций** |
| --- | --- | --- |
| 2010 | 165 |  |
| 2011 | 227 | +62 |
| 2012 | 443 | +216 |
| 2013 | 562 | +119 |
| 2014 | 639 | +77 |
| 2015 | 971 | +332 |
| 2016 | 1381 | +410 |
| 2017 | 1934 | +553 |
| 2018 | 3263 | +1329 |
| 2019 | 4575 | +1312 |
| 2020 | 5973 | +1398 |
| 2021 | 6285 | +312 |
| 2022 | 6765 | +480 |
| 2023 | 7623 | +858 |

Заметен постоянный рост интереса к данной области, особенно скачок наблюдается в период с 2017 по 2018 год, когда количество публикаций увеличилось почти в два раза. Несмотря на замедление темпов роста в последующие годы, общий тренд показывает устойчивый рост количества публикаций, достигая пика в 2023 году с 858 новыми работами и общим количеством 7623 публикаций

Результаты анализа показали, что **топ-5 тем с наибольшим ростом** интереса включают темы

| **Топик** | **Тема** | **Ключевые Слова** | **Изменение Кол-ва Публикаций** |
| --- | --- | --- | --- |
| 0 | Машинное обучение: Байесовские методы, опорные векторы (SVM), переобучение, классификация | bayesian, svm, overfitting, classifiers, arxiv | +7458.0 |
| 2 | Искусственный интеллект: AlphaGo, автоматоны, робототехника, искусственные интеллектуальные системы | alphago, automaton, robots, ais, robotic | +2136.0 |
| 1 | Анализ данных: Кластеризация, метод опорных векторов (SVM), классификация, Вороного диаграммы, компьютерное зрение | clustering, svm, classifiers, voronoi, opencv | +1750.0 |
| 5 | Машинное обучение: Классификация, Байесовские методы, долгосрочная память (LSTM), SVM, переобучение | classifiers, bayesian, lstm, svm, overfitting | +1515.0 |
| 14 | Графическая визуализация: Вороного диаграммы, трассировка лучей, фотограмметрия, воксели | voronoi, raytracing, photogrammetry, voxel, svm | +1251.0 |

**топ-5 тем с наибольшим сокращением** интереса включают

| **Топик** | **Тема** | **Ключевые Слова** | **Изменение Кол-ва Публикаций** |
| --- | --- | --- | --- |
| 140 | Связь и передача данных: Реле, ретрансляция, передача сигналов | relay, relays, insuch, relaying, transceiver | -49.0 |
| 93 | Канальные связи, специализированные исследования в машинном обучении: байесовские методы, архив публикаций (arXiv), квантование, сверточные сети, латентный анализ | bayesian, arxiv, quantization, convolution | -42.0 |
| 477 | Распределенные системы: Peer-to-peer, BitTorrent, децентрализованные сети, BGP, блокчейн | peer, bittorrent, decentralized, bgp, blockchains | -21.0 |
| 203 | Сетевые технологии: Кластеризация, подсети, ZigBee, управление контейнерами | clustering, subnets, cluster, zigbee, kubernetes | -18.0 |
| 590 | Технологии баз данных: XML, XPath, NoSQL, RDF, структуры данных | xml, xpath, nosql, rdf, datastructure | -14.0 |

Эти результаты могут служить основой для дальнейшего изучения причин изменений в популярности тем и для определения стратегии исследовательской работы в соответствии с текущими научными тенденциями.

## Краткое описание top2vec

[Top2Vec](https://github.com/ddangelov/Top2Vec) — это алгоритм, разработанный для автоматического обнаружения тем в текстовых данных, создания векторных представлений для тем, документов и слов, которые совместно вложены в единое семантическое пространство.

Особенностью Top2Vec является его способность автоматически определять количество тем в наборе данных без предварительной настройки или указания их числа исследователем. Это достигается благодаря комбинированию нескольких современных методов машинного обучения и обработки естественного языка, среди которых:

* Использование моделей, таких как [Doc2Vec](https://radimrehurek.com/gensim/auto_examples/tutorials/run_doc2vec_lee.html), [Universal Sentence Encoder](https://www.tensorflow.org/hub/tutorials/semantic_similarity_with_tf_hub_universal_encoder?hl=ru) или [BERT Sentence Transformer](https://www.sbert.net/), для создания совместных векторов документов и слов.
* Применение алгоритма уменьшения размерности [UMAP](https://umap-learn.readthedocs.io/en/latest/) для выделения плотных кластеров документов, что облегчает последующее их группирование.
* Использование алгоритма кластеризации [HDBSCAN](https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html) для обнаружения и определения границ плотных областей документов в сокращенном пространстве.
* Определение центроидов для каждой обнаруженной плотной области, которые представляют собой векторы тем, и выбор ключевых слов для каждой темы на основе близости к этим центроидам.

Преимущества Top2Vec включают в себя не только автоматическое определение количества тем, но и отсутствие необходимости в ручной предварительной обработке текста, такой как удаление стоп-слов, стемминг или лемматизация. Кроме того, алгоритм эффективно работает с текстами различной длины, что делает его универсальным инструментом для анализа больших наборов текстовых данных и обнаружения в них скрытых семантических структур.

## Как были выбраны ключевые слова.

Выбор ключевых слов для исследования экологических проблем с помощью БПЛА базируется на актуальности и специфике проблем в различных сферах экологического мониторинга:

1. **Стихийные бедствия и чрезвычайные ситуации ["uav", "disasters", "emergency"]**: ключевые слова выбраны из-за роли БПЛА в оперативном мониторинге и оценке ущерба в зонах бедствий, что важно для своевременного реагирования и минимизации последствий.
2. **Загрязнение воздуха ["uav", "pollution", "emissions"]**: слова отражают значимость БПЛА в изучении воздушного загрязнения и выбросов, которые оказывают влияние на здоровье и климат, обеспечивая точные данные для анализа.
3. **Загрязнение воды ["uav", "pollution", "water"]**: выбор обусловлен возможностями БПЛА в выявлении источников загрязнения воды, что критически важно для защиты водных ресурсов.
4. **Бытовые отходы ["uav", "recycling", "urban"]**: ключевые слова подчеркивают потенциал БПЛА в мониторинге управления отходами и рециркуляции в условиях городской среды.
5. **Инфракрасное и тепловое картографирование ["uav", "infrared", "heat", "thermal"]**: слова отражают использование БПЛА для обнаружения тепловых загрязнений и картографирования с помощью инфракрасных и тепловых датчиков.

Для поддержки этих выборов можно опираться на следующие статьи:

"[Applications of drone in disaster management: A scoping review](https://www.sciencedirect.com/science/article/pii/S1355030621001477) 2022": Статья обзорно исследует применение дронов в управлении стихийными бедствиями, выделяя их ключевые области использования и указывая на необходимость дальнейших исследований.

"[A Review on Air Quality Measurement Using an Unmanned Aerial Vehicle](https://link.springer.com/article/10.1007/s11270-020-04973-5) 2021":

Статья рассматривает использование (БПЛА) с различными датчиками для мониторинга качества воздуха, подчеркивая их потенциал в обеспечении более высокого пространственного и временного разрешения данных по сравнению с традиционными методами.

"[Unmanned Aerial Vehicles in Hydrology and Water Management: Applications, Challenges, and Perspectives](https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021WR029925) 2021": Статья обсуждает применение БПЛА в гидрологии и управлении водными ресурсами, акцентируя внимание на их эффективности и инновационном потенциале для сбора данных.

"[Drone technology in municipal solid waste management and landfilling: A comprehensive review](https://www.sciencedirect.com/science/article/abs/pii/S0956053X21006516) 2022": Статья рассматривает применение БПЛА для управления свалками и отслеживания выбросов, подчеркивая их растущее использование и перспективы исследований.

"[A Review of Modern Thermal Imaging Sensor Technology and Applications for Autonomous Aerial Navigation](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8540138/) 2021": Статья представляет обзор тепловизионных датчиков и их интеграции в системы автономной навигации для БПЛА, подчеркивая их эффективность в условиях ограниченного освещения и в GPS-запрещенных зонах.

## Картинка со словами.

В этот раз использовал ключевые слова. ["emissions","disasters", "monitoring","emergency"]



## Использовать другие тематических моделей?

Я могу попробовать использовать BERTopic, но пока не знаю как я смог бы воссоздать примерно такие же топики, т.к алгоритмы отличаются, получится другое количество топиков.

Если необходимо я могу попробовать.

## Сгладить тренды путем применения регрессии.

Применил полиномиальную регрессию из библиотеки **sklearn**

Y - количество публикаций

X - дата публикации год-месяц

Точки это количество публикаций по определенной тематике с указанной датой.

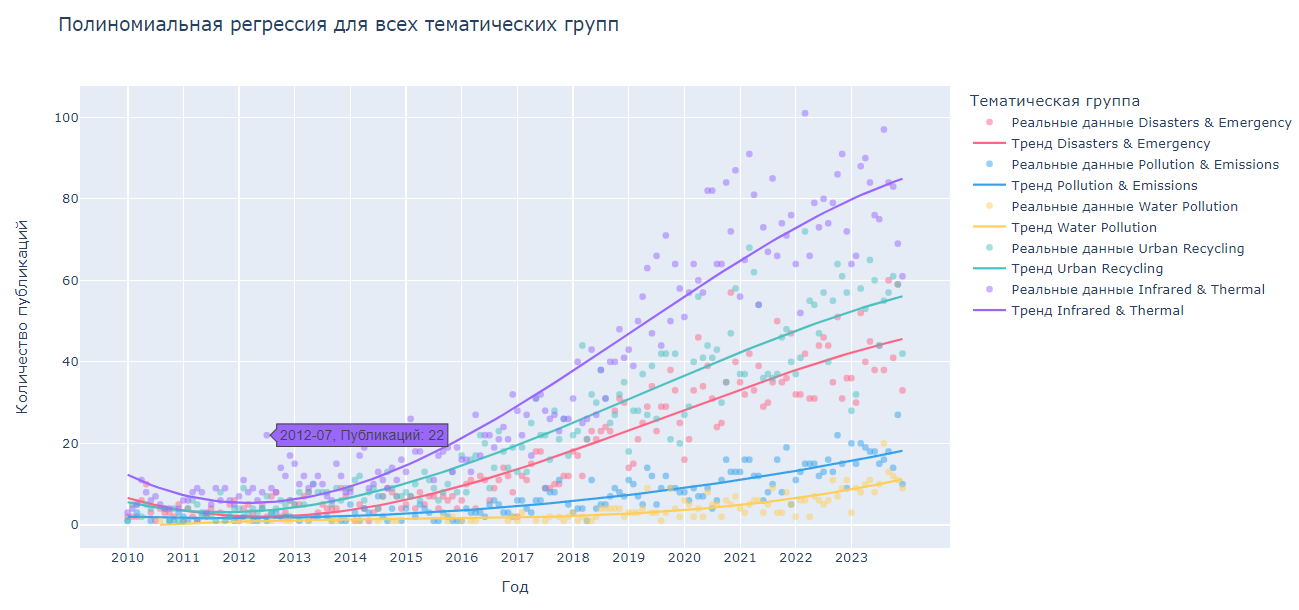


график 5 топиков с регрессией

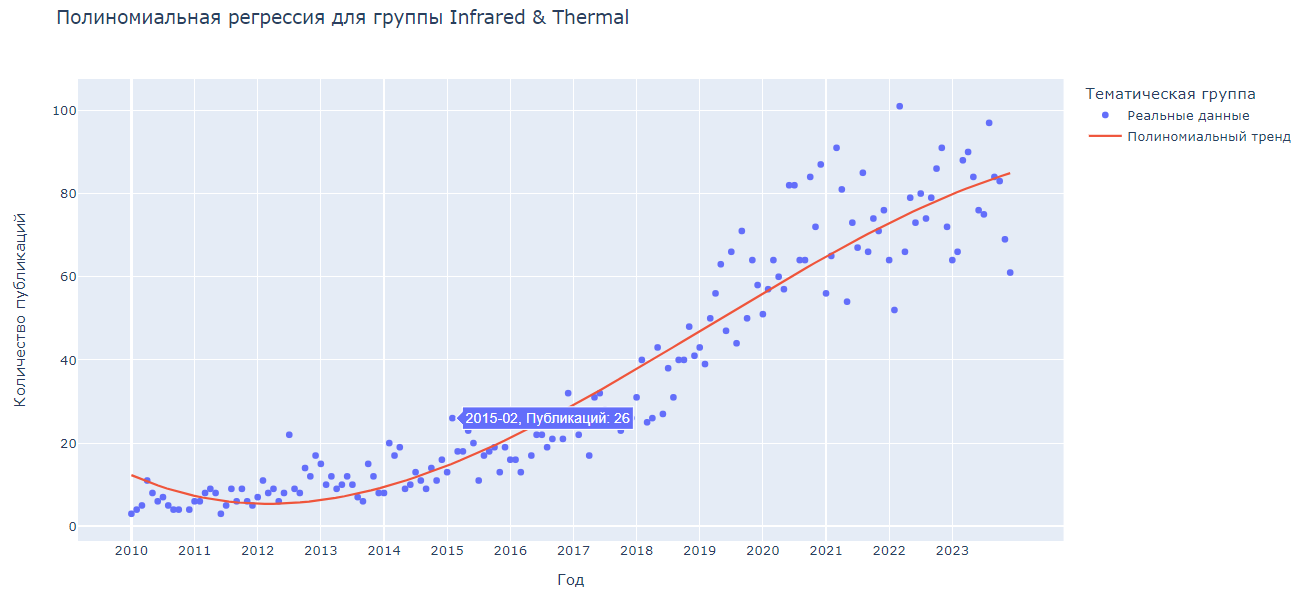


график 1 конкретного топика