**Задача.** Мониторинг экологический проблем с помощью БПЛА

Найти эталонные абстракты статей по данной тематике, чтобы создать более точный топик.

**Основные задачи:**

1. Какие тематики могут относится к нашему топику.
2. Как тематическая группа развивалась во времени несколько тематических групп развивались (стихийные, воздух, вода, бытовые отходы, инфракрасное )
3. Показать ключевые слова для данного топика

**Процесс.**

1. Найти эталонные абстракты статей по тематике БПЛА.
2. Скрапер

2.1 Собрать данные по нужной нам тематике

2.2 Собрать данные по всем статьям, чтобы сравнить тренды развития

1. Использовать алгоритмы тематического моделирования BERTopic или Top2vec.

статьи и нужно ещё поискать.

[*https://doi.org/10.3390/rs15235544*](https://doi.org/10.3390/rs15235544)

[*https://doi.org/10.3390/rs15174269*](https://doi.org/10.3390/rs15174269)

[*https://doi.org/10.3390/drones7060357*](https://doi.org/10.3390/drones7060357)

[*https://doi.org/10.3390/math10152552*](https://doi.org/10.3390/math10152552)

[*https://doi.org/10.1109/ACCESS.2022.3149625*](https://doi.org/10.1109/ACCESS.2022.3149625)

[*https://doi.org/10.3390/app11125541*](https://doi.org/10.3390/app11125541)

[*https://doi.org/10.3390/app112110171*](https://doi.org/10.3390/app112110171)

[*https://doi.org/10.3390/sym12121945*](https://doi.org/10.3390/sym12121945)

[*https://doi.org/10.1109/ACCESS.2019.2937627*](https://doi.org/10.1109/ACCESS.2019.2937627)

Структура статьи

Введение   
Анализ работ (статей)

Метод

Результаты

Фантазии  
Скачать статьи по схожей тематикой.

Используя скрапер я собрал корпус научных публикаций с arxiv за период с 2010 по 2024-01-01 по тематике Computer Science, получилось 555 тысяч статей.   
**1.** **Для определения необходимых топиков среди всего корпуса.**   
Я применил top2vec алгоритм для создания модели. Получилось 690 топиков.

Группы были созданы следующим образом.

С помощью top2vec я на вход подавал ключевые слова и получал топики наиболее схожие с этими словами

Стихийные ["uav", "disasters", "emergency"]

Воздух ["uav", "pollution", "emissions"]

Вода ["uav", "pollution", "water"]

Бытовые отходы ["uav", "recycling", "urban"]

Инфракрасное ["uav", "infrared", "heat","thermal]

В скобках указаны ключевые слова

Таким образом получилось 5 тематических групп состоящих из уникальных топиков.

'Disasters & Emergency': [385, 591, 462, 346, 173, 226, 430, 369],

'Pollution & Emissions': [448, 190, 642],

'Water Pollution': [355, 629],

'Urban Recycling': [454, 679, 161, 201, 479, 595, 562, 508, 262, 155],

'Infrared & Thermal': [338, 291, 91, 172, 580, 611, 625, 88, 264, 620, 342, 457]

В скобках указаны топики для данной группы

Общее количество публикаций для всех тематических групп 14271 документ.

Изначально у меня многие топики попадали в одну и ту же группу, по этой причине я взял только уникальные топики для групп.

1. [385, 285, 100, 591, 577, 549, 176, 452, 120, 572, 489, 462, 358, 178, 346, 173, 226, 444, 430, 369]

2. [455, 285, 448, 100, 176, 577, 678, 661, 453, 572, 497, 489, 549, 444, 190, 642, 603, 402, 286, 485]

3. [285, 100, 455, 176, 577, 444, 402, 355, 286, 485, 120, 489, 371, 453, 661, 629, 572, 603, 678, 497]

4. [661, 285, 100, 454, 679, 678, 161, 176, 455, 201, 497, 178, 479, 595, 562, 508, 286, 371, 262, 155]

5. [176, 285, 338, 100, 291, 91, 577, 172, 580, 485, 611, 625, 452, 88, 264, 603, 620, 358, 342, 457]

Группы до выделения уникальных топиков

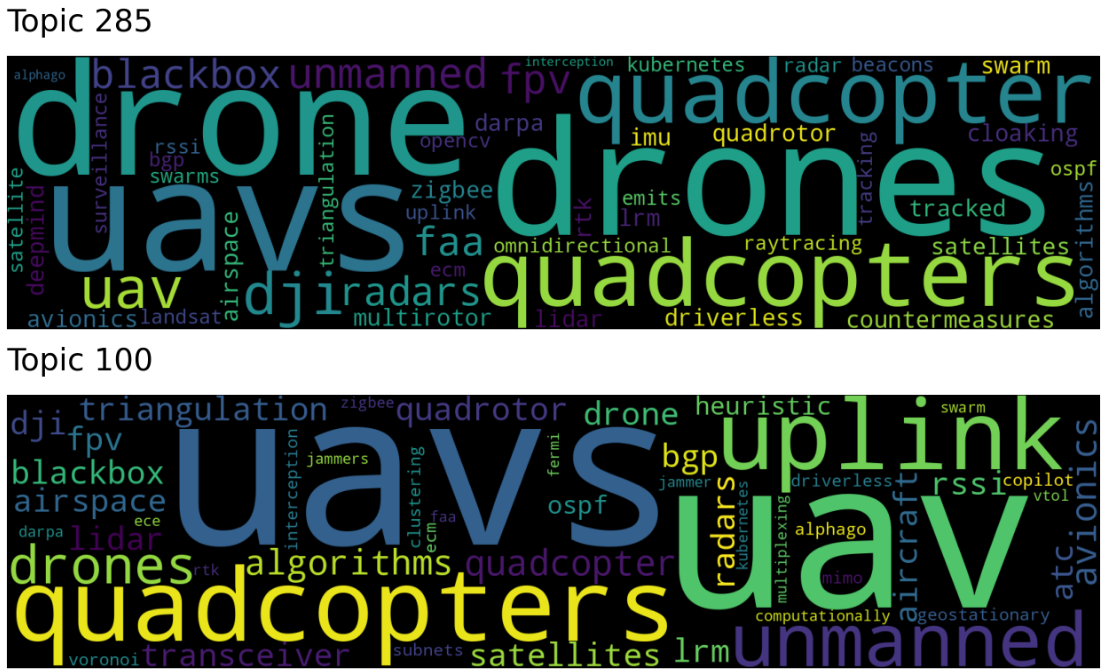
6 группа пересекающиеся топики.

intersecting\_topics = [402, 661, 285, 286, 549, 678, 176, 178, 444, 572, 577, 452, 453, 455, 603, 100, 485, 358, 489, 497, 371, 120]

**2. Для определения схожих топиков**

Я применил wordcloudиз top2vec и таким образом выбрал похожие топики

**Принцип работы.** Я подаю на вход ключевое слово и на выходе получаю наиболее схожие с этим словом топики и их ключевые слова.

****

**3. Для определения как тематическая группа развивалась во времени**

Тематические группы:

1. 'Disasters & Emergency' - количество публикаций: 2968

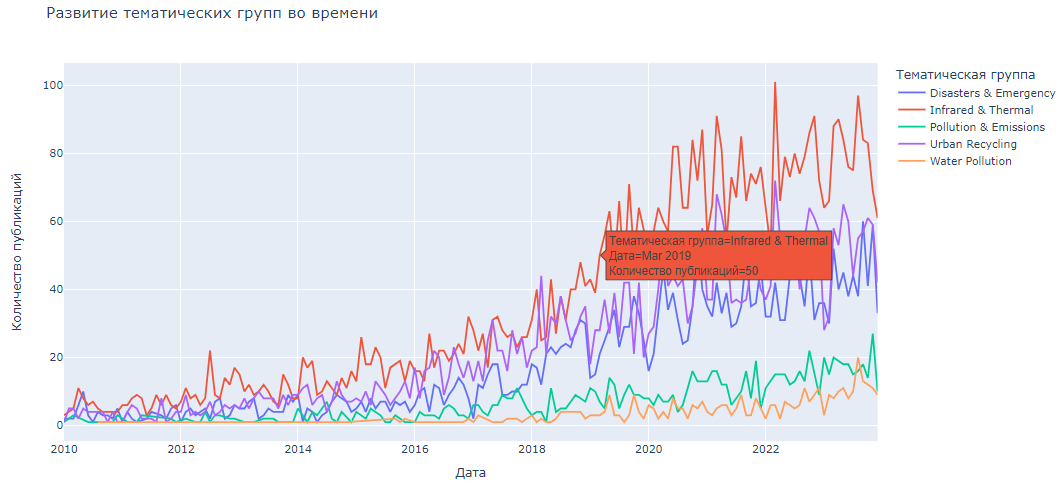
2. 'Pollution & Emissions' - количество публикаций: 1026

3. 'Water Pollution' - количество публикаций: 435

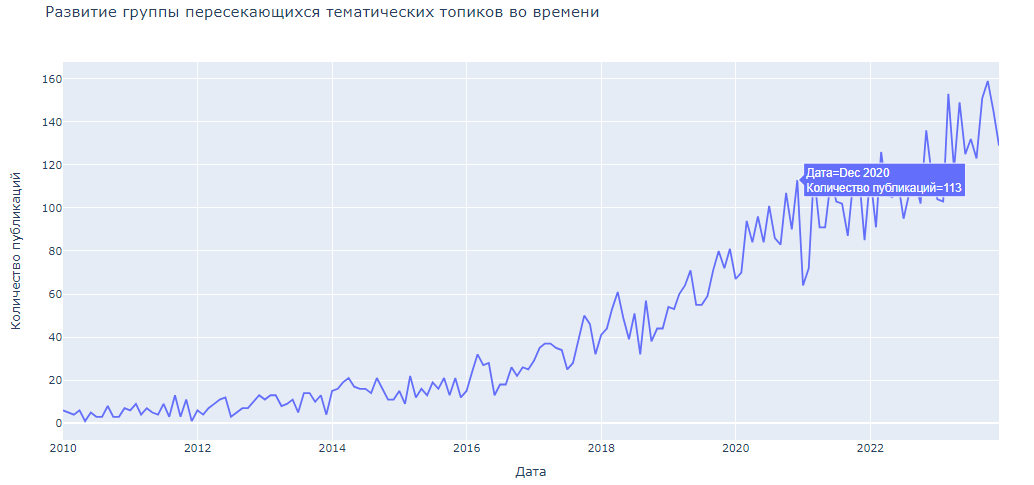
4. 'Urban Recycling' - количество публикаций: 3891

5. 'Infrared & Thermal' - количество публикаций: 5951

Общее количество публикаций для всех тематических групп: 14271



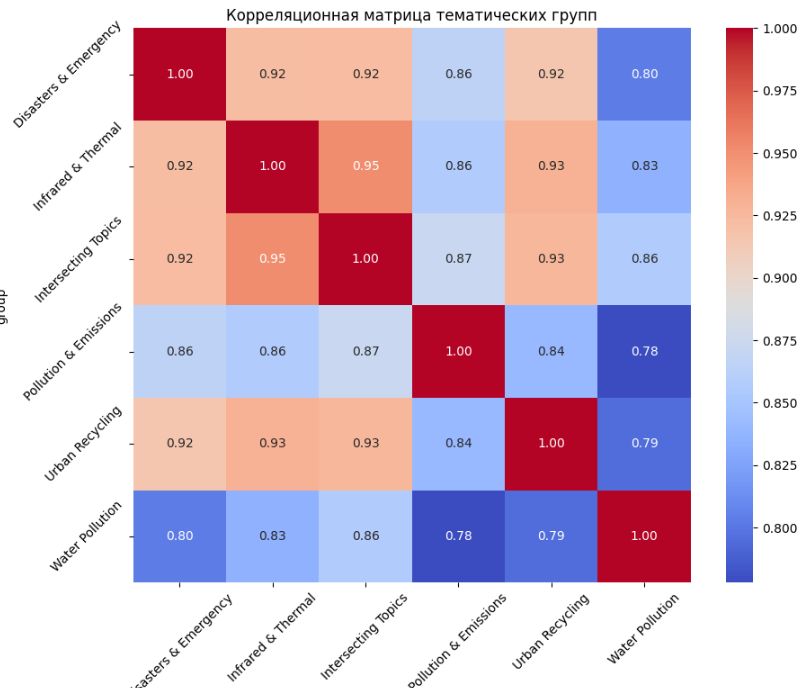
На графике показан тренд развития 5 групп за период с 2010 по 2023-12-31



Тренд развития 6-й пересекающейся группы

Общее количество публикаций для пересекающихся топиков: 7934

Анализ корреляции между тематическими группами, основанный на количестве публикаций в каждой группе.



heatmap для 5 тематических групп