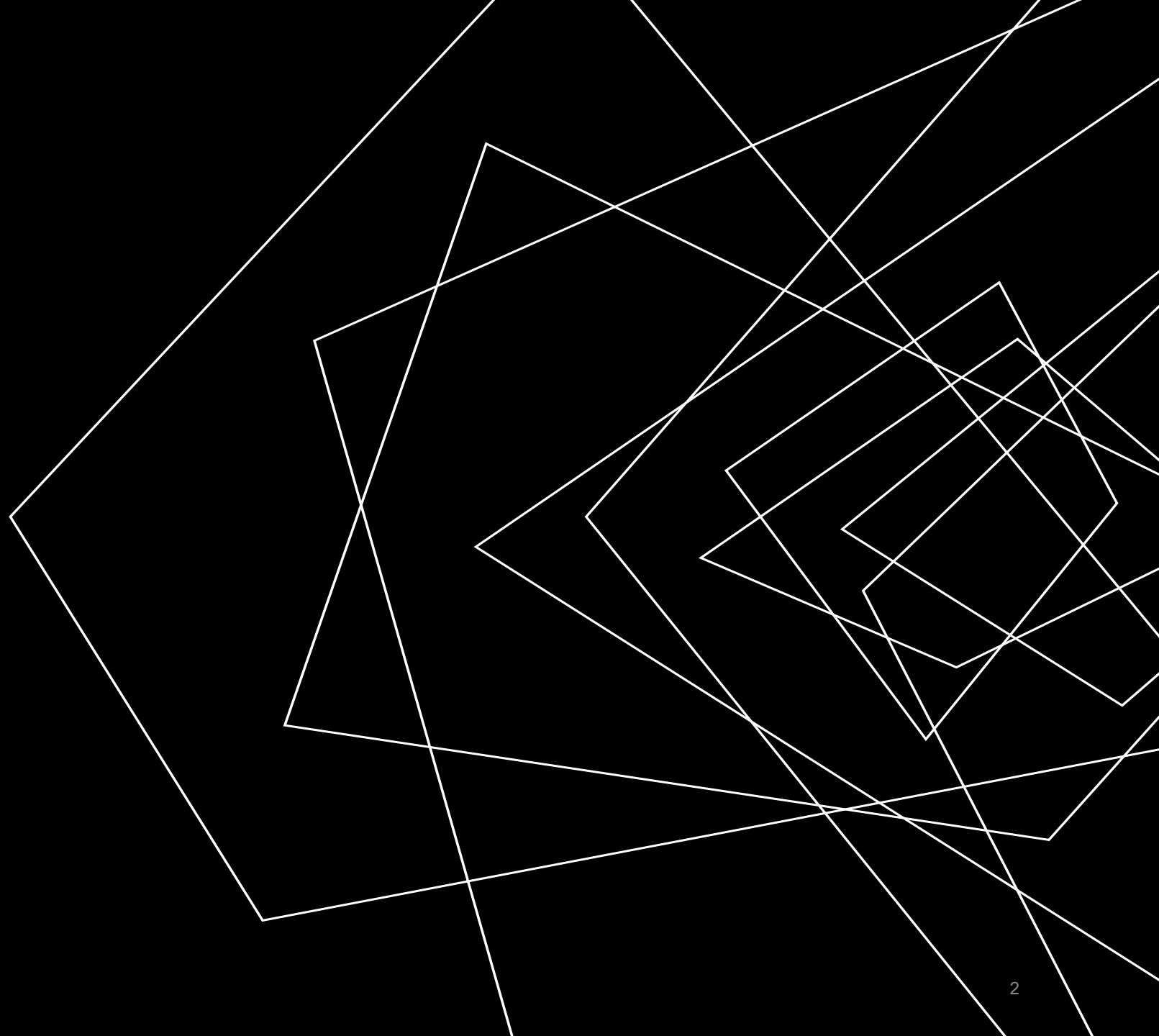
ИССЛЕДОВАНИЕ НОВОСТНОЙ АКТИВНОСТИ В СФЕРЕ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ С ПРИМЕНЕНИЕМ МАШИННОГО ОБУЧЕНИЯ НА ПРИМЕРЕ РОССИЙСКОГО ИНТЕРНЕТ-СМИ

Даргель Юлия

ВНИУ ВШЭ ДПО

Компьютерная лингвистика, 2025

СОДЕРЖАНИЕ



АКТУАЛЬНОСТЬ

Актуальность исследования обусловлена быстрым развитием сектора ИТ и его стратегической важностью для экономики страны и национальной безопасности.

Анализ новостных трендов позволяет отслеживать изменения в технологическом ландшафте, выявлять новые возможности для инноваций и адаптировать маркетинговые стратегии.

Также подобные исследования представляют ценность для мониторинга результатов реализации государственной политики в ИТ-отрасли и общественного восприятия технологий, что способствует принятию более обоснованных решений.

Почему СМИ, а не соцсети?

Россияне больше доверяют новостным, аналитическим и официальным сайтам, чем мессенджерам, блогам и соцсетям.[1] Даже молодежь предпочитает аккаунты СМИ безымянным ТГ-каналам, и эта тенденция наблюдается уже несколько лет.[2]

1 https://wciom.ru/analytical-reviews/analiticheskii-obzor/doverie-smi-v-rossii

2 https://www.cossa.ru/news/328678/

ЗАДАЧА

Проследить, как менялось количество новостей на разные темы, опубликованных на отраслевом ИТ-портале Comnews.ru с января 2022 г. по июнь 2025 года.

ЦИФРЫ

7980 новостей

проанализированы

14 кластеров

сформированы

80+ новостей

месячный максимум на тему «Регулирование ИТ-рынка»

4,5 часа

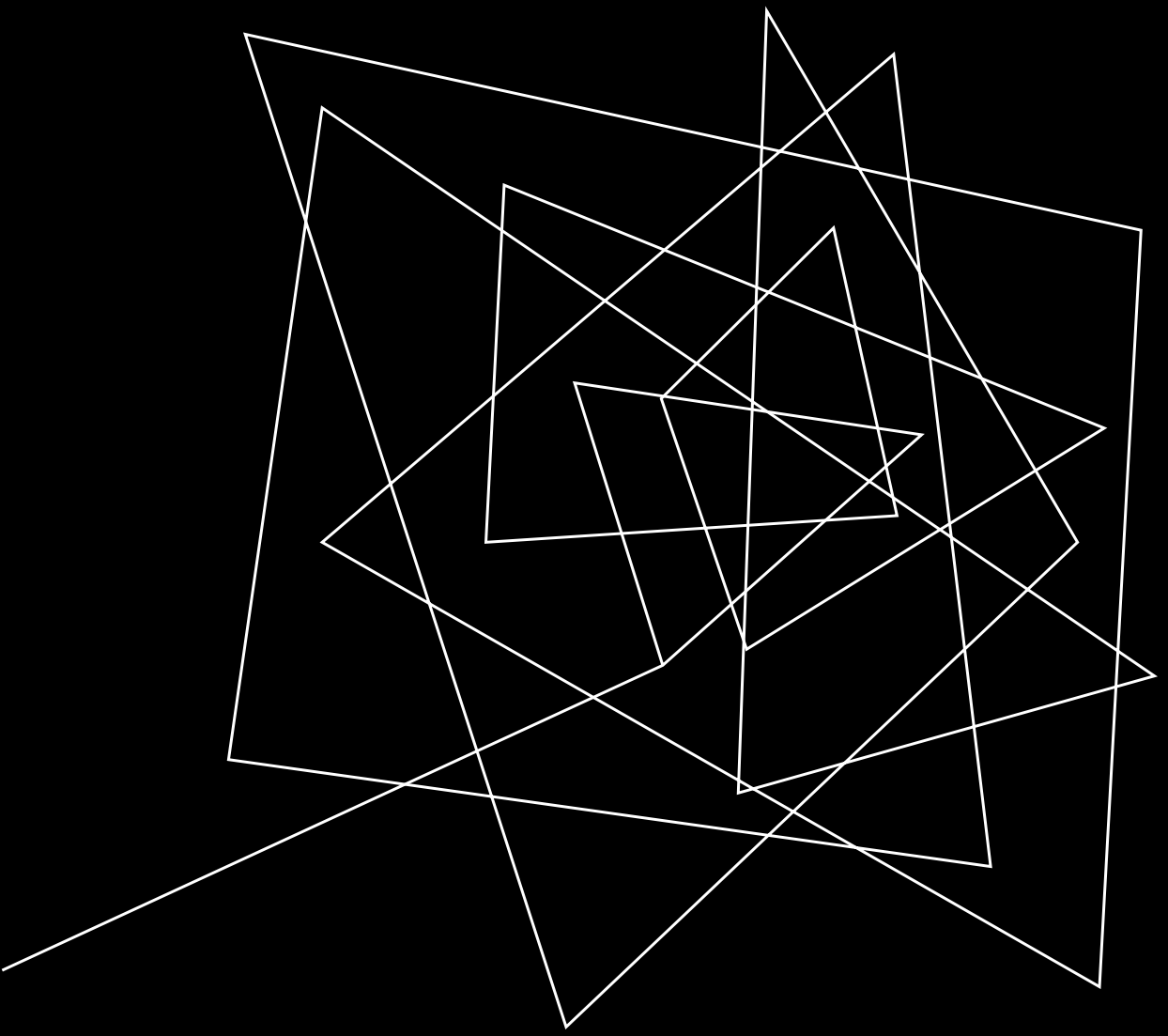
длился парсинг

0.012

лучшее значение коэффициента силуэта

2023 г.

самый активный период



РЕШЕНИЕ

АЛГОРИТМ И ВЫБРАННЫЕ ИНСТРУМЕНТЫ

ЗАДАЧА	ИНСТРУМЕНТЫ
Сбор данных. Для каждой новости за выбранный период собраны следующие данные: дата, заголовок, текст, ссылка	• BeautifulSoup • Pandas
Предобработка. Тексты новостей приведены к нижнему регистру, удалены знаки препинания, спецсимволы, стоп-слова.	• Регулярные выражения • Nltk • MyStem
Векторизация текста. Векторизованы слова, которые встречаются не менее, чем в 5 текстах и не более, чем в 70% выборки.	• Numpy • sklearn • TfidfVectorizer
Выбор оптимального количества кластеров (k). Проведено сравнение качества кластеризации при количестве кластеров от 5 до 15 с определением лучшего варианта.	• K-means • Silhouette Score • Davies-Bouldin Index • Calinski-Harabasz Index
Кластеризация	• K-means
Визуализация кластеров	• Matplotlib/seaborn • TSNE
Тематическое моделирование	• Gensim • Латентное распределение Дирихле

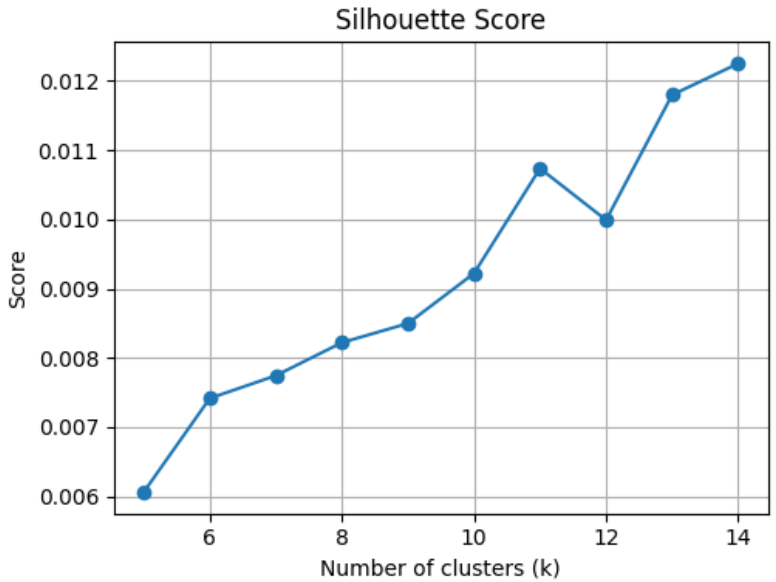
ВЫБОР ОПТИМАЛЬНОГО КОЛИЧЕСТВА КЛАСТЕРОВ

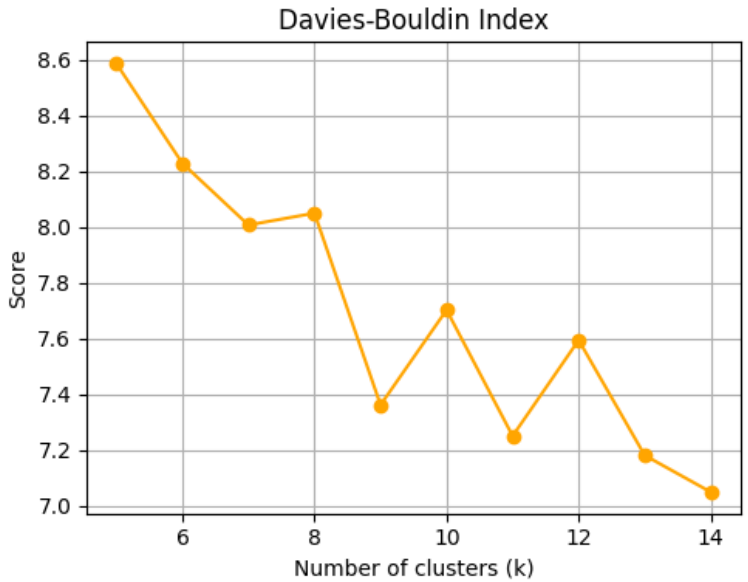
Шаг 1. Для большей объективности были использованы три метрики.

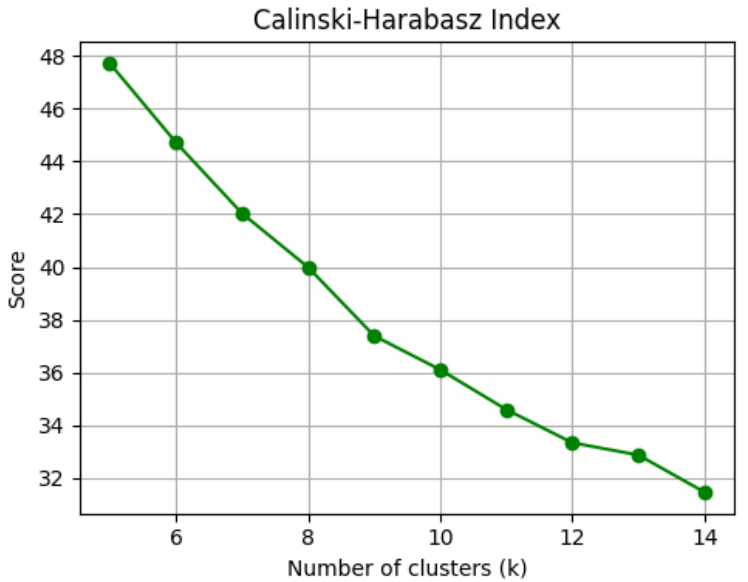
Шаг 2. На втором шаге лучшие k по трем метрикам был нормализованы и проведено взвешенное усреднение.

МЕТРИКА	ЧТО ИЗМЕРЯЕТ	ДИАПАЗОН	ИНТЕРПРЕТАЦИЯ
Silhouette Score (Коэффициент силуэта)	Насколько объекты внутри кластера похожи друг на друга и насколько они отличаются от объектов в других кластерах.	От -1 до 1	Чем ближе к 1, тем лучше объекты кластеризованы.
Davies-Bouldin Index	Насколько кластеры отличаются друг от друга и насколько они компактны.	∞	Чем меньше значение тем лучше; низкое значение означает, что кластеры достаточно сильно разделены и объекты внутри них близки друг к другу.
Calinski-Harabasz Index (Variance Ratio Criterion)	Отношение межкластерной дисперсии к внутрикластерной	∞	Чем выше значение, тем лучше структура кластеров.

ВИЗУАЛИЗАЦИЯ МЕТРИК ДЛЯ ВЫБОРА ОПТИМАЛЬНОГО K



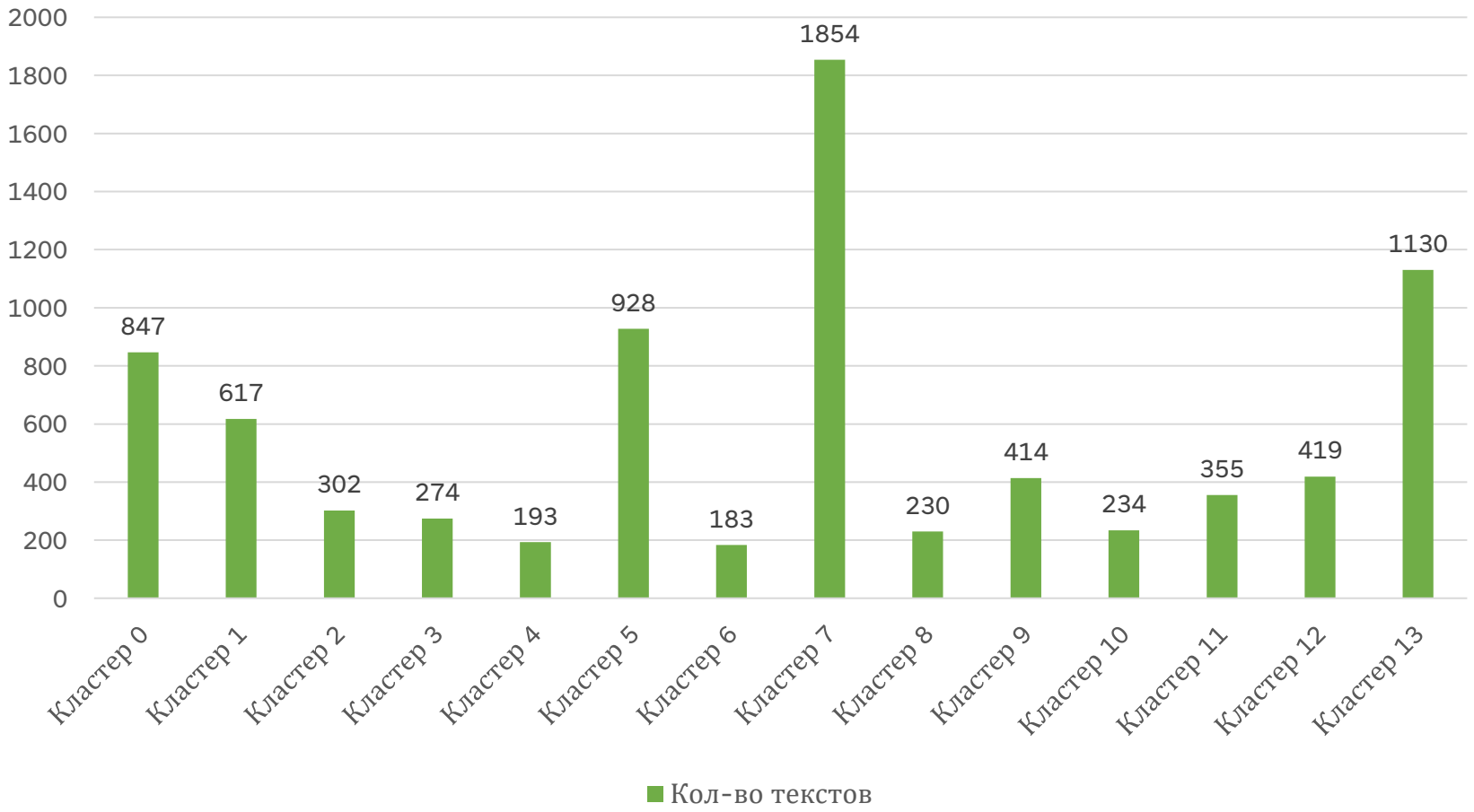




Лучшие значения метрик

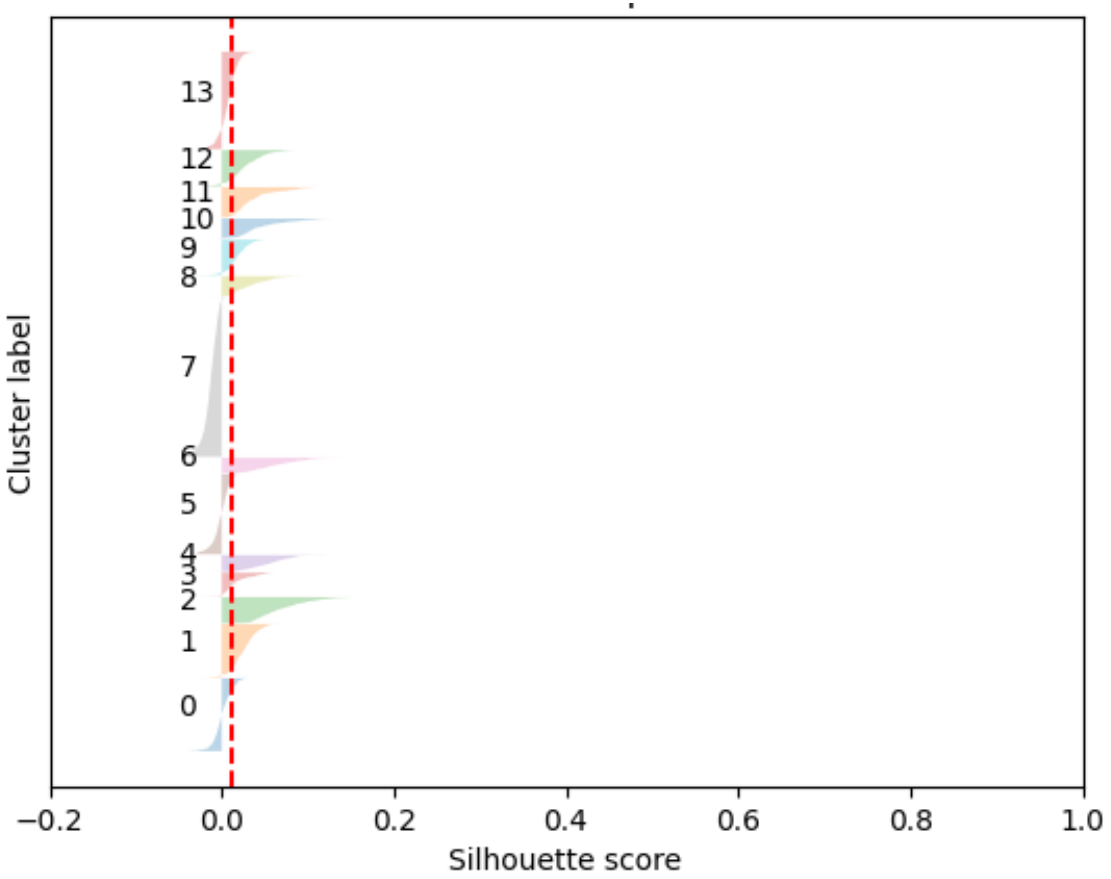
Silhouette Score (k=14)	Davies-Bouldin (k=14)	Calinski-Harabasz (k=5)
0.012	7.051	47.704

РАСПРЕДЕЛЕНИЕ ТЕКСТОВ ПО КЛАСТЕРАМ

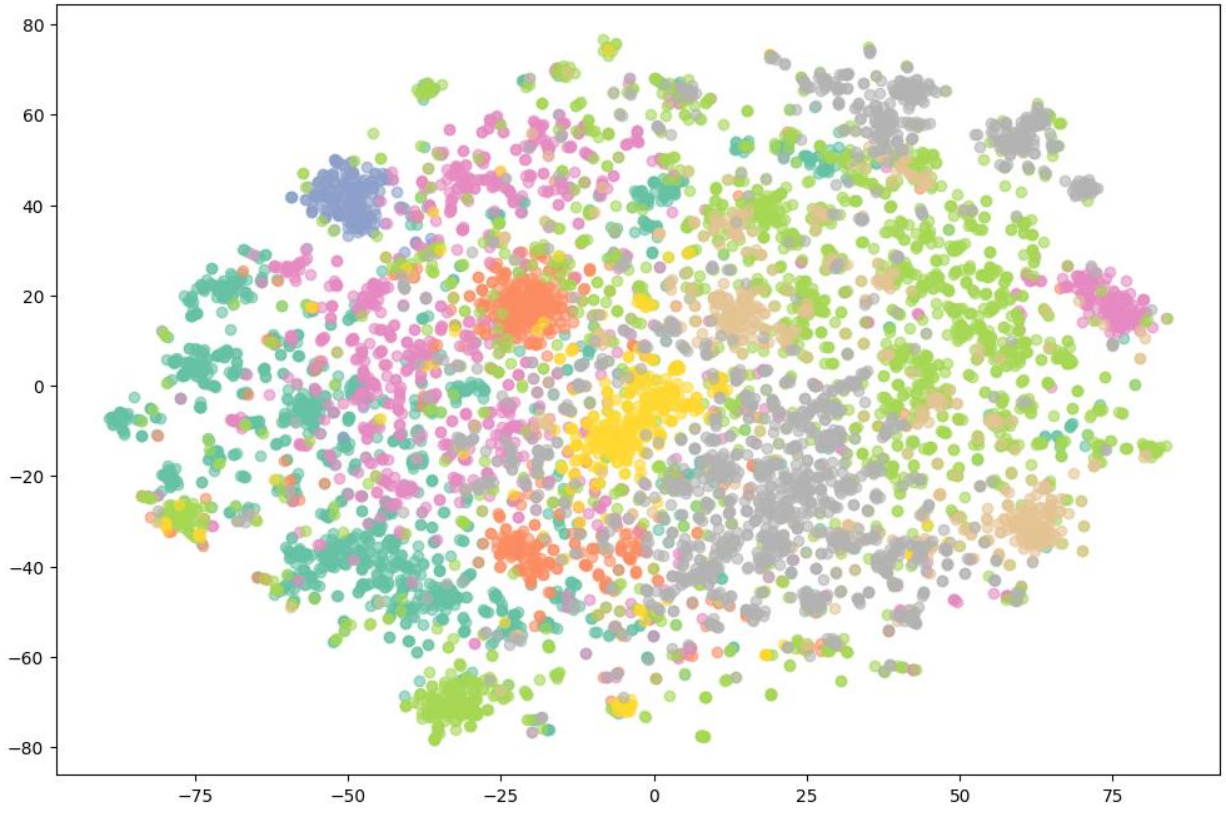


ВИЗУАЛИЗАЦИЯ КЛАСТЕРОВ

Silhouette Plot



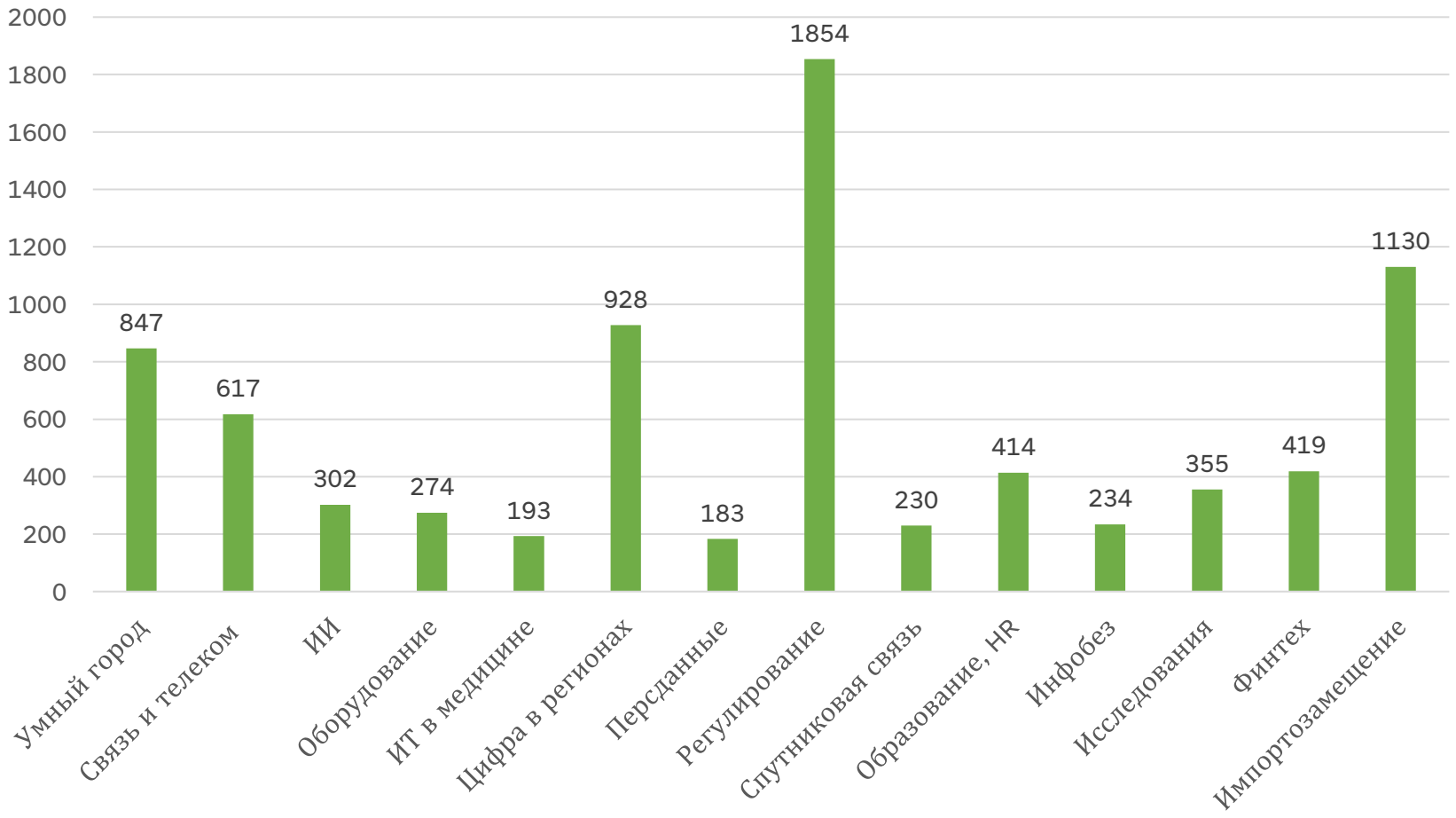
TSNE

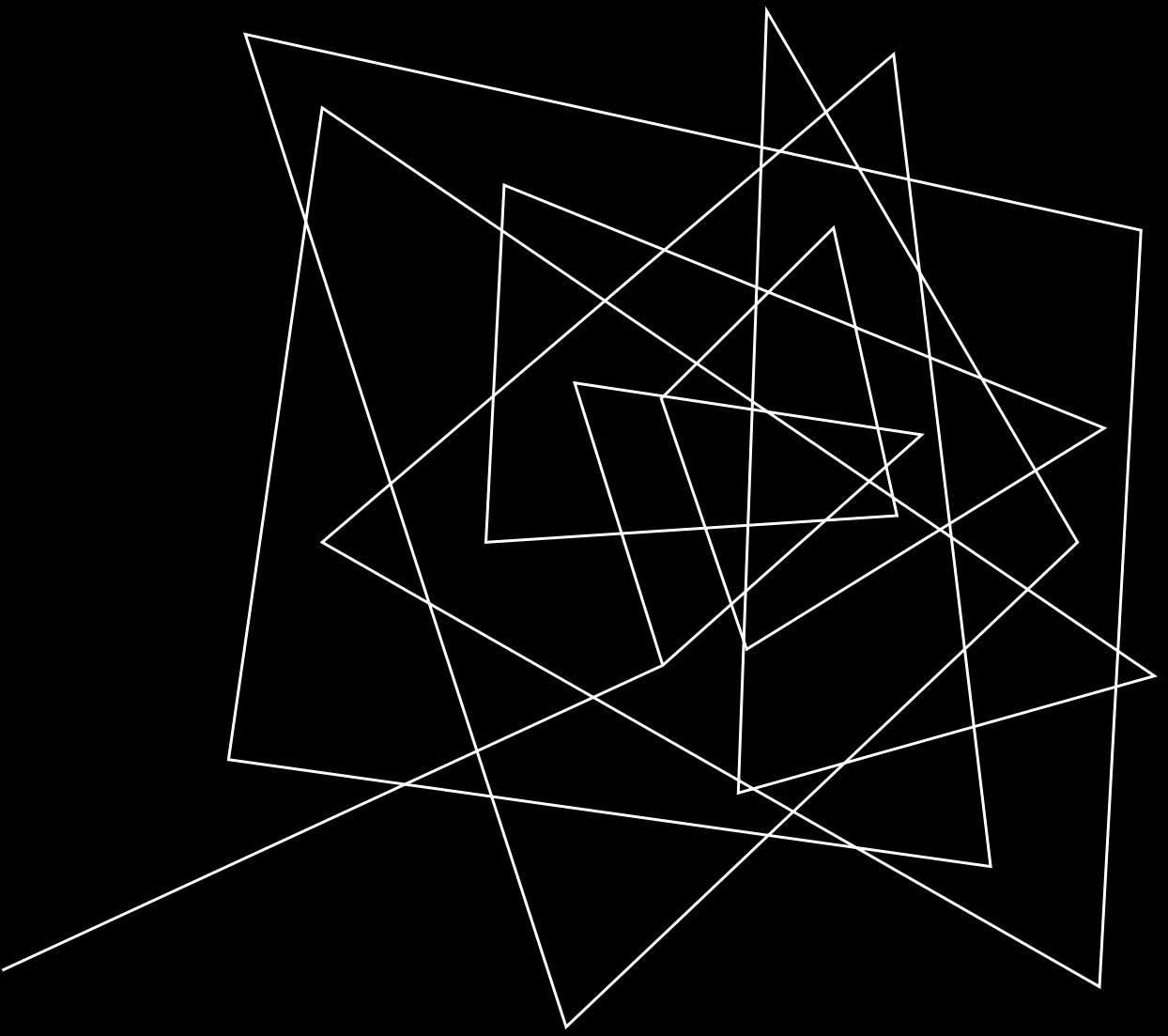


ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

№	ТЕМА	КЛЮЧЕВЫЕ СЛОВА
0	Умный город	система, проект, работа, город, позволять, также, транспорт, цифровой, решение, весь
1	Связь и телекоммуникации	связь, оператор, сеть, г, интернет, мобильный, оборудование, станция, весь, услуга
2	Искусственный интеллект	ия, технология, интеллект, искусственный, решение, данные, мочь, система, развитие, также
3	Оборудование для ИТ	производство, предприятие, система, оборудование, проект, российский, технология, продукция, промышленный, развитие
4	Технологии в медицине	медицинский, врач, пациент, система, здравоохранение, данные, цифровой, весь, сервис, технология
5	Цифровизация регионов	цифровой, система, проект, информационный, развитие, работа, область, государственный, также, регион
6	Персональные данные	данные, утечка, персональный, мочь, информация, защита, штраф, также, система, данный
7	Регулирование ИТ-рынка	россия, российский, мочь, г, также, весь, сервис, рынок, пользователь, новый
8	Спутниковая связь	система, спутник, космический, связь, спутниковый, г, проект, аппарат, российский, россия
9	Образование и кадры	специалист, ита, работа, сотрудник, также, программа, рынок, проект, г, мочь
10	Информационная безопасность	атака, злоумышленник, данные, г, мочь, также, система, защита, безопасность, ddos
11	Исследования рынка	г, руб, млрд, рынок, рост, выручка, составлять, млн, квартал, российский
12	Финансовые технологии	банк, цифровой, россия, мочь, рубль, финансовый, клиент, криптовалюта, система, рынок
13	Импортозамещение	решение, российский, рынок, система, продукт, ита, г, директор, проект, развитие

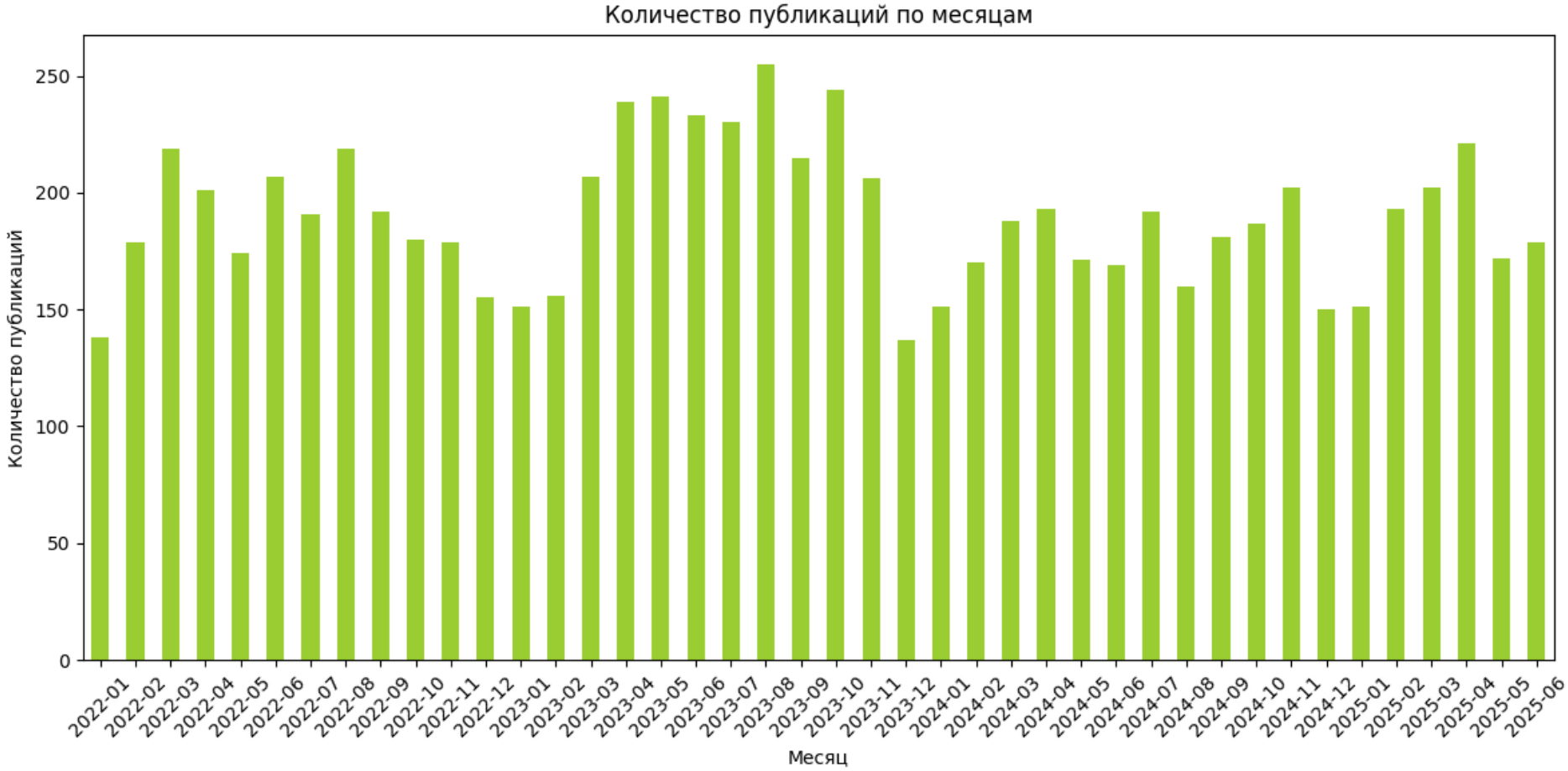
РАСПРЕДЕЛЕНИЕ ТЕКСТОВ ПО ТЕМАМ





РЕЗУЛЬТАТЫ

РАСПРЕДЕЛЕНИЕ ВСЕХ НОВОСТЕЙ ПО МЕСЯЦАМ

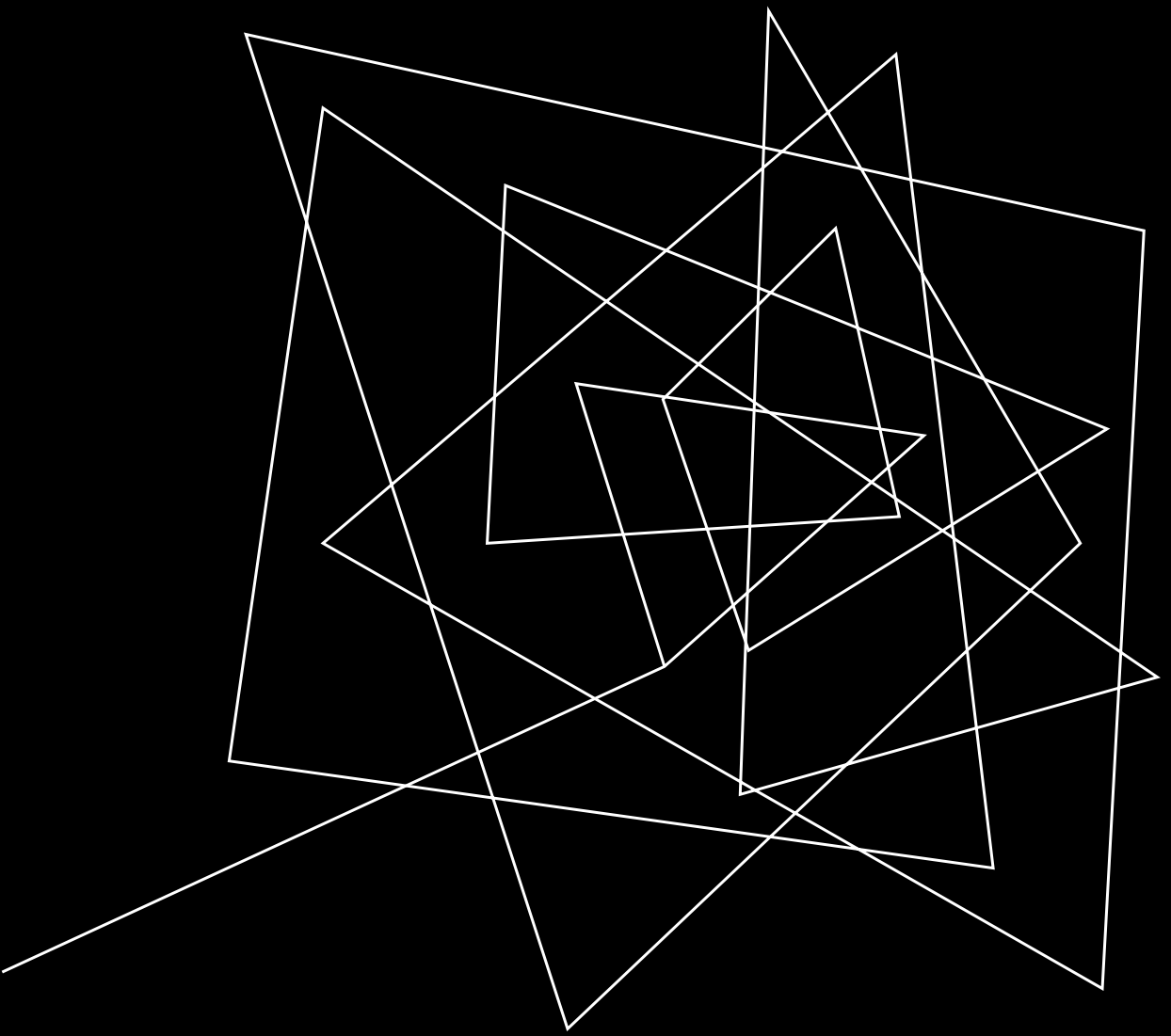


РАСПРЕДЕЛЕНИЕ ТЕМАТИЧЕСКИХ НОВОСТЕЙ ПО МЕСЯЦАМ

РАСПРЕДЕЛЕНИЕ ТЕМАТИЧЕСКИХ НОВОСТЕЙ ПО МЕСЯЦАМ

РАСПРЕДЕЛЕНИЕ ТЕМАТИЧЕСКИХ НОВОСТЕЙ ПО МЕСЯЦАМ

РАСПРЕДЕЛЕНИЕ ТЕМАТИЧЕСКИХ НОВОСТЕЙ ПО МЕСЯЦАМ



ВЫВОДЫ

ИТ-РЫНОК СТАБИЛИЗИРУЕТ АКТИВНОСТЬ

- Больше всего новостей публикуется по темам: Регулирование ИТ-рынка (до 80/мес), Импортозамещение (до 40/мес) и Умный город (до 35/мес)
- Меньше всего освещаются Технологии в медицине и Спутниковая связь. Возможная причина – это технически сложные направления, где объективно меньше инфоповодов.
- В 2022 году заметен всплеск новостей на темы Информационной безопасности и Персональных данных. Вероятно, в связи с СВО
- В 2023 г. был наиболее активным из проанализированных периодов практически по всем темам. Отдельно можно отметить взлет темы «ИИ»
- В 2024 году в целом новостной фон стабилизировался, активность немного уменьшилась по всем темам, кроме Умного города, ИИ и Цифровизации регионов. Эта тенденция продолжилась в 2025 г.
- Интересно, что все три года ноябрь – один из наименее активных месяцев, а летом – всплески публикаций, несмотря на отпускной сезон

СПАСИБО

Проект на GitHub:
https://github.com/Dargel/ICT_News

Датасет на HuggingFace

КОД В COLAB:

Часть 1 – Парсинг

Часть 2 – Основная часть