

Обзор методов в программном решении задачи разведочного анализа текстов через кластеризацию и аппроксимацию числа кластеров

LV МЕЖДУНАРОДНАЯ
НАУЧНАЯ КОНФЕРЕНЦИЯ
аспирантов и студентов
«ПРОЦЕССЫ УПРАВЛЕНИЯ И УСТОЙЧИВОСТЬ»
Control Processes and Stability (CPS'24)

Неронов Роман Михайлович
студент бакалавриата факультета ПМ-ПУ
Санкт-Петербургский государственный университет

Санкт-Петербург, 2024

План доклада

1. Проблематика

- Постановка задачи

2. Решение

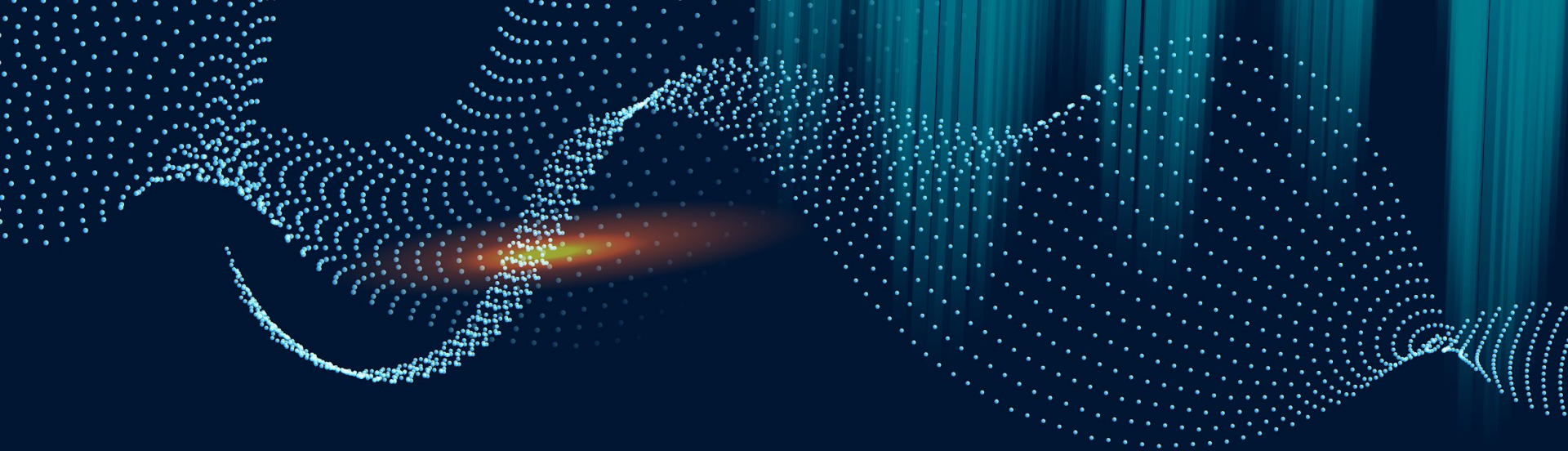
- План решения
- Первичная кластеризация
- Алгоритм подбора количества кластеров
- Ключевые слова для кластеров
- Интерактивное взаимодействие с кластеризацией
- Классификация

3. Сравнение с аналогами

- Тестирование
- Ключевые слова AnaText
- 2D визуализация AnaText
- 2D визуализация BERTopic
- Сравнение по метрикам

4. Подведение итогов

- Список источников



01

Проблематика

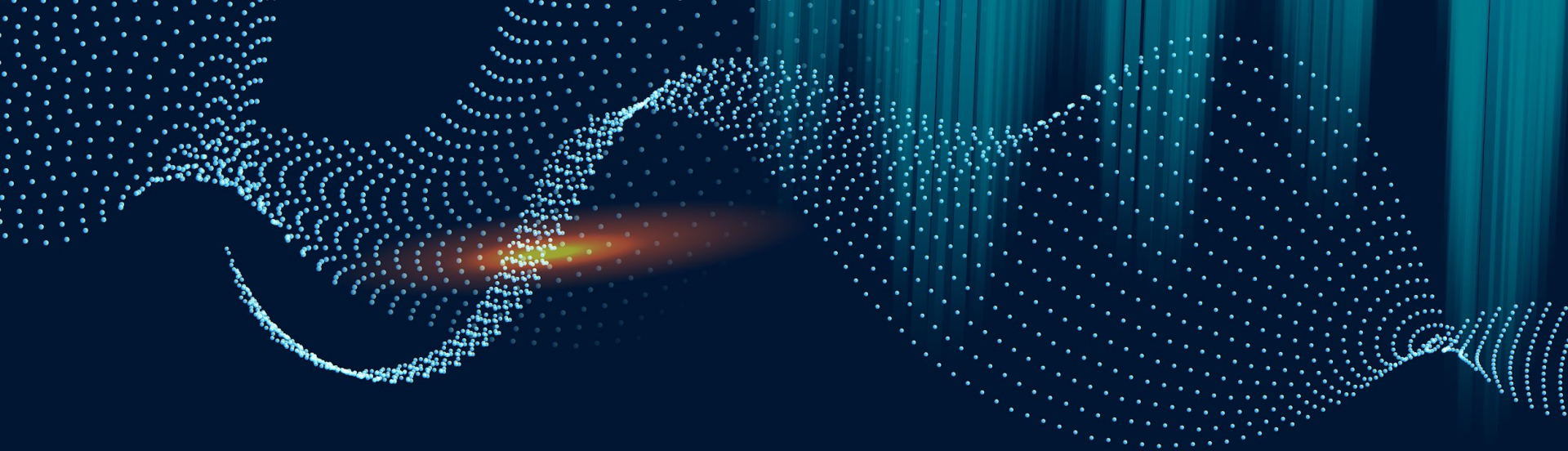
Постановка задачи

Источники данных

- ❑ Новостные ленты
- ❑ Юридические документы
- ❑ Обсуждение и отзывы в соцсетях
- ❑ Обращения клиентов через колл-центр и приложение

Что есть в данных?

Заказчик хочет понять, какие паттерны (темы, кластеры) присутствуют в коллекции текстов.



02

Решение

План решения



Первичная кластеризация

Корпус текстов

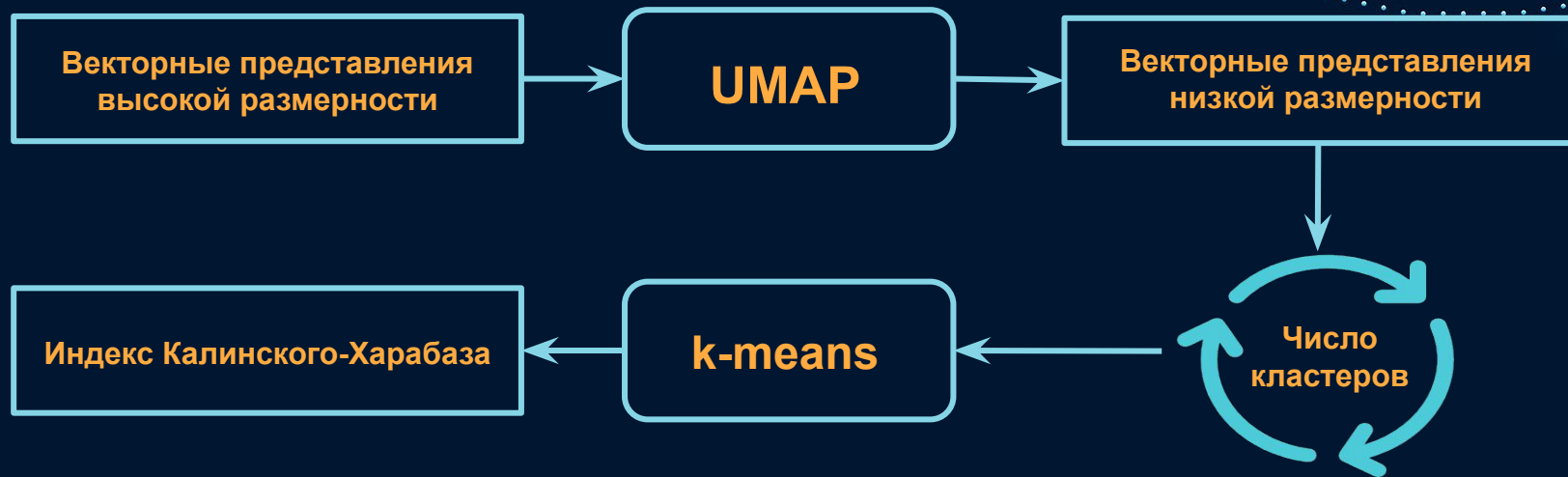
BERT Transformer

Векторное представление

Подбор числа кластеров

Кластеризация

Алгоритм подбора количества кластеров



$$CHI = \frac{B/(c-1)}{W/(n-c)}$$

n — количество сэмплов данных.

c — количество кластеров.

B — матрица внутренней дисперсии.

W — матрица внешней дисперсии.

Ключевые слова для кластеров

- ❑ Находим топ 6 слов для каждого текста отдельно с помощью keyBERT
- ❑ Подсчитываем частоты для каждого слова среди всех ключевых слов для текстов из одного кластера
- ❑ Выбираем 10 наиболее часто встречающихся слов
- ❑ Удаляем слова, которые встречаются в топе более чем у 50% кластеров



Интерактивное взаимодействие с кластеризацией

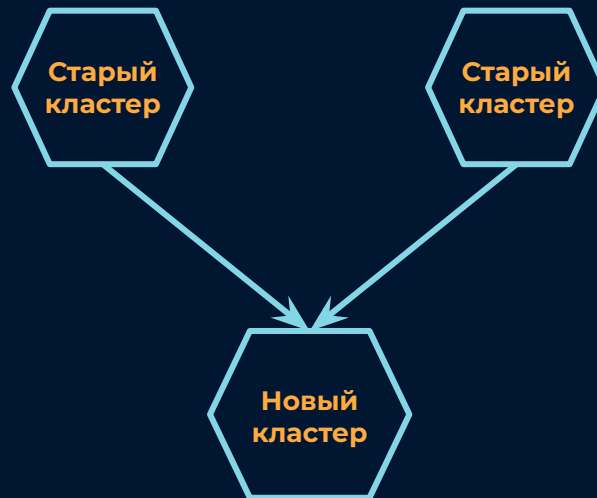
Разделение кластеров

Пользователь задает количество новых кластеров



Объединение кластеров

Пользователь выбирает кластеры, которые хочет объединить



Классификация

- ❑ В случае, если пользователя устраивает качество кластеризации, он может классифицировать новый корпус текстов на основе полученных меток
- ❑ Для классификации используется алгоритм *Random Forest Classifier*.





03 | Сравнение с аналогами

Тестирование

Корпус текстов - 20 newsgroups

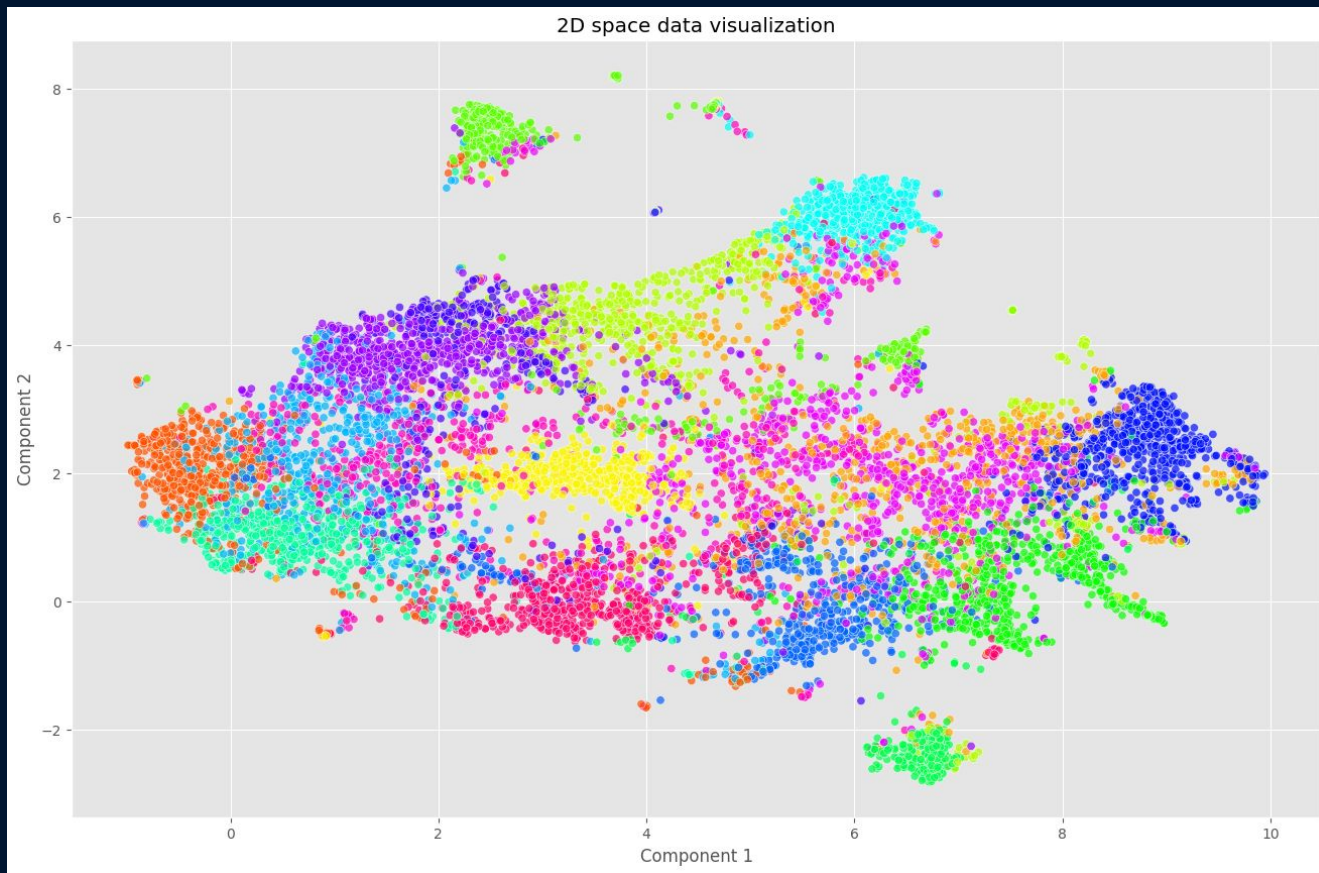
- 20 тем
- 11 тыс. текстов

	Количество кластеров	Время исполнения, сек
AnaText	17	840.48
BERTopic	173	326.84

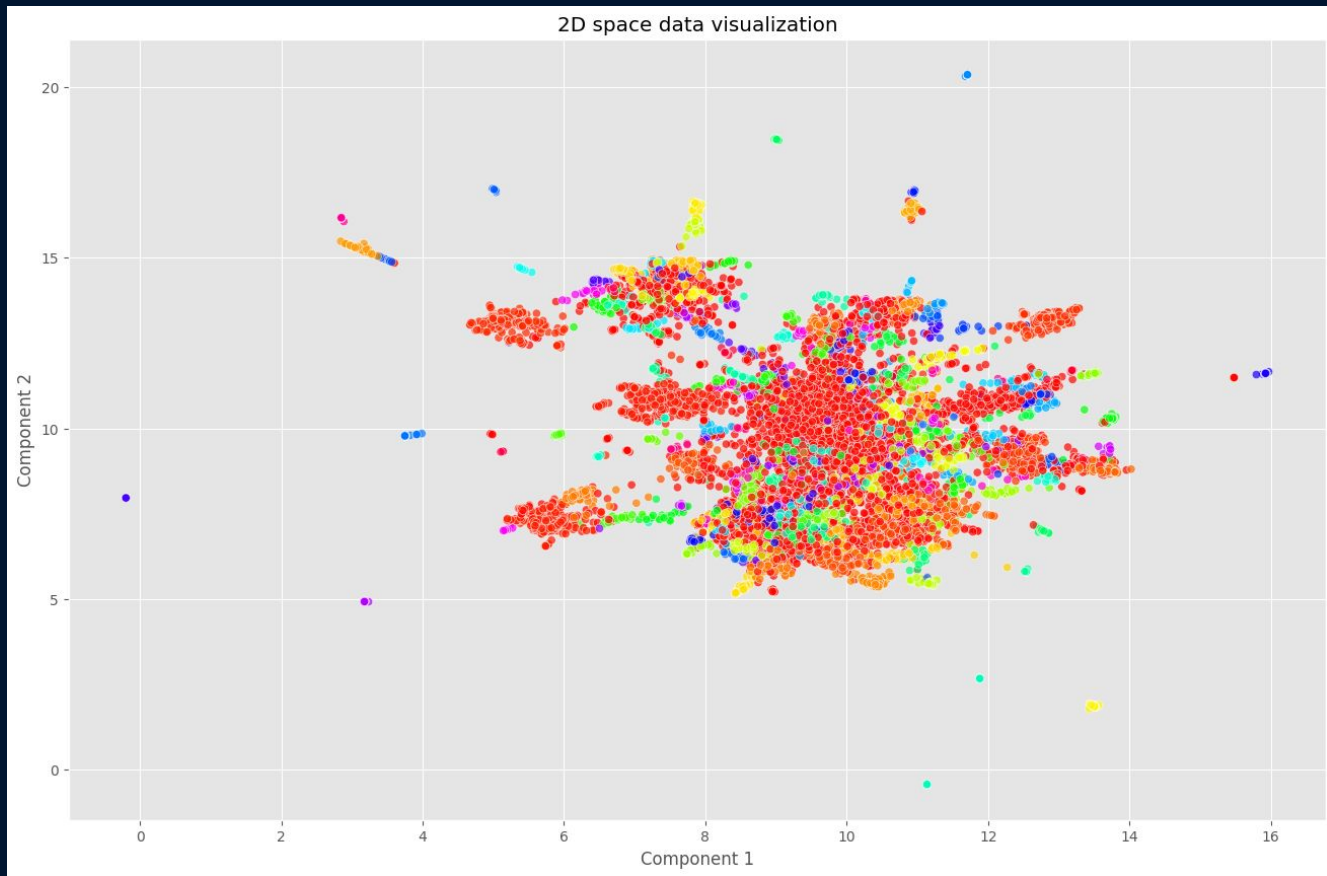
Ключевые слова AnaText

	Keyword 1	Keyword 2	Keyword 3	Keyword 4	Keyword 5	Keyword 6	Keyword 7	Keyword 8	Keyword 9	Keyword 10
16	engine	car	bike	drive	motorcycle	speed	driver	back	slow	gas
15	fax	mail	post	file	host	usenet	sale	ftp	orbit	address
14	oh	back	god	cc	get	post	bmw	hmmm	freenet	newsgroup
13	window	post	card	host	help	organization	nntp	computer	problem	question
12	sale	interested	buy	offer	price	bike	cheap	sell	card	cost
11	god	atheist	bible	faith	morality	islam	heaven	gay	christianity	christians
10	government	encryption	key	chip	tax	phone	pay	federal	moon	cost
9	window	problem	card	help	use	apple	mac	software	computer	disk
8	team	hockey	game	baseball	league	player	win	coach	toronto	goal
7	disk	motherboard	computer	chip	monitor	hardware	controller	problem	fast	modem
6	doctor	disease	symptom	patient	treatment	pain	headache	medical	sugar	diet
5	weapon	government	firearm	israeli	gun	war	genocide	armenia	israelis	kill
4	nntp	oh	cs	system	back	organization	hockey	window	hey	printer
3	organization	nntp	post	baseball	game	god	california	player	virginia	question
2	bank	gordon	fax	back	computer	mail	file	card	disk	oh
1	god	nntp	atheist	post	oh	apple	get	caltech	question	georgia
0	window	file	directory	disk	command	key	memory	error	mouse	fax

2D визуализация AnaText

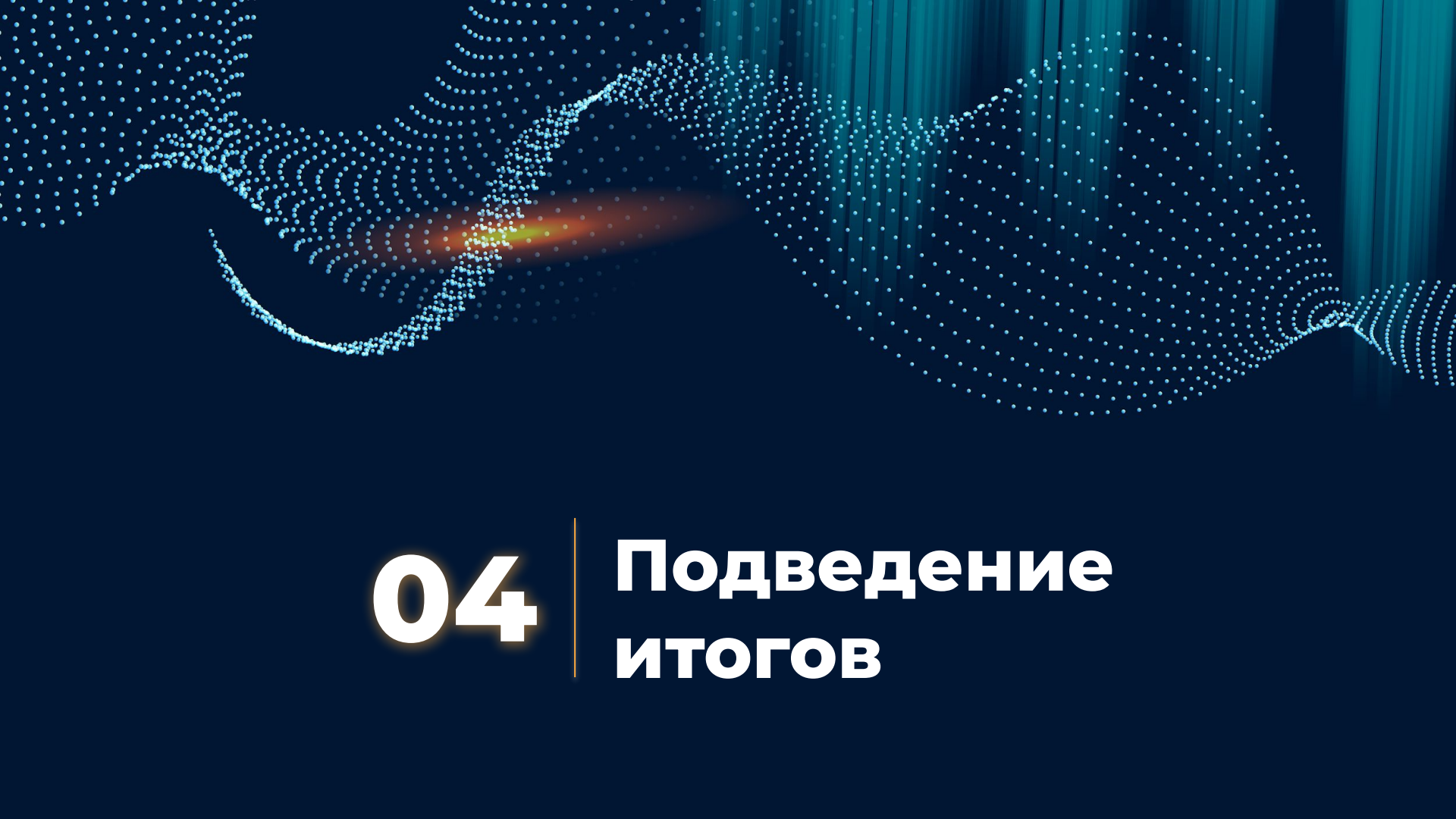


2D визуализация BERTopic



Сравнение по метрикам

	Adjusted Rand Index	Adjusted Mutual Information	V-measure	Fowlkes-Mallows Index	Silhouette Score
AnaText	0.15	0.31	0.31	0.20	0.08
BERTopic	0.06	0.40	0.42	0.14	-0.18



04 | Подведение ИТОГОВ

СПИСОК ИСТОЧНИКОВ

1. UMAP: Uniform Manifold Approximation and Projection / Leland McInnes [и др.] // Journal of Open Source Software. 2018. V. 3. No 29. P. 861.
2. Aman Priyanshu, Supriti Vijay AdaptKeyBERT: An Attention-Based approach towards Few-Shot & Zero-Shot Domain Adaptation of KeyBERT // arxiv URL: <https://arxiv.org/abs/2211.07499> (дата обращения: 17.03.2024).
3. Supporting Clustering with Contrastive Learning / Dejjiao Zhang [и др.] // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021. P. 5419–5430.

Спасибо за внимание

