



МойОфис

От ключевиков до LLM: ретроспектива методов кластеризации коротких текстов



Кирилл Улитин

Кирилл Улитин: От ключевиков до LLM

Кирилл Улитин

- 15+ лет в IT
- 9 лет в МойОфис, 5 лет руковожу направлением UX и исследований
- Преподаватель майнора UX-дизайн, ВШЭ, ex. AccessibilityUnity
- В ПК ResearchExpo, ex. ProfsoUX
- Пишу t.me/ulitin_ru



Кирилл Улитин: От ключевиков до LLM

МойОфис



11+
лет на рынке

Создаём ПО
нового поколения

12500+
клиентов

различных секторов
рынка уже используют
ПО МойОфис

МойОфис — российский разработчик широкой
линейки безопасных офисных приложений
для решения любых бизнес-задач

МойОфис лидер на рынке ИТ
Входит в **топ-3** офисного ПО в России
По версии **TELECOM DAILY**

~4 460
партнёров
во всех регионах РФ

1100+
сотрудников
800 из них — разработчики
Собственные ресурсы
для создания программных
продуктов

Акционер

68,8%
Лаборатория Касперского

Стратегический партнёр, одна из крупнейших
в мире частных компаний, работающих в сфере
информационной безопасности



Продукты МойОфис

МойОфис Документы Настольные

Пакет безопасных решений для работы с документами

МойОфис Документы Онлайн

Корпоративное решение для совместной работы и централизованного хранения документов

МойОфис SDK

Веб-компоненты и библиотеки для встраивания редакторов МойОфис в прикладные системы

МойОфис Squadus

Корпоративный мессенджер с видео конференциями и вебинарами

 ИТ-премия **CNews Awards** в номинации «Цифровое рабочее пространство»

 Премия **TAdviser IT Prize** в номинации «Цифровое рабочее пространство года»

МойОфис Mailion

Корпоративная почта для крупных коммерческих и государственных организаций от 5000 пользователей

 Второе место в рейтинге **CNews** корпоративных почтовых решений

МойОфис Почта

Корпоративная почта для среднего и крупного бизнеса до 5000 пользователей

МояДоска

Виртуальная доска для визуализации проектов и идей с помощью широкого набора инструментов

МойОфис Схема

Графический редактор для описания схем и бизнес-процессов, визуализации идей и проектирования технических чертежей

Кирилл Улитин: От ключевиков до LLM

**Короткие
тексты?**



Открытые ответы в опросах

Чем вам понравился наш продукт?

Простой в использовании,
хорошая цена

Пользовательский фидбек



Здравствуйте, я не могу зайти в аккаунт. Пожалуйста, помогите мне восстановить пароль

Цифровой след



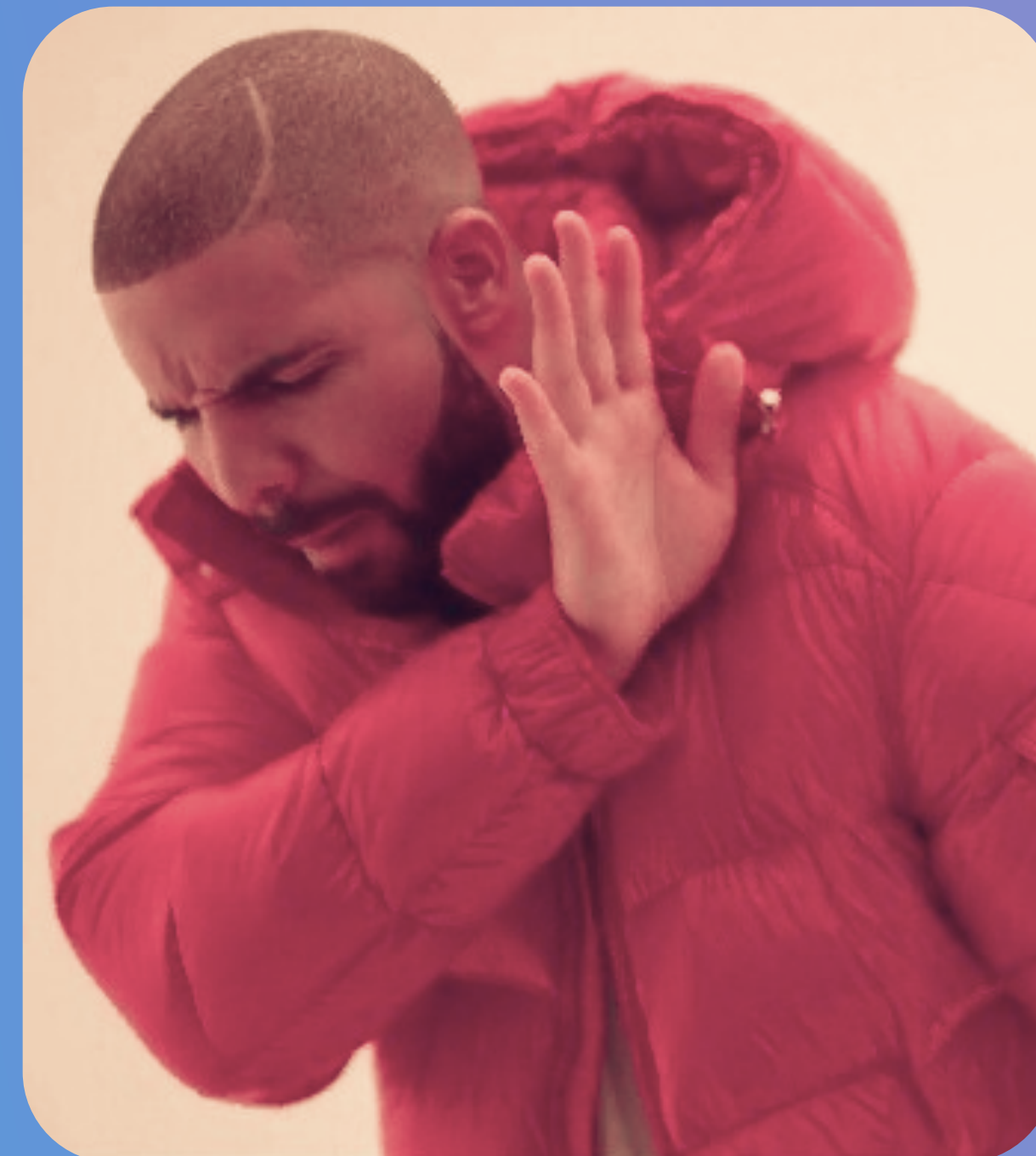
Классное приложение!
Это лучшее что я пробовал.

Кирилл Улитин: От ключевиков до LLM

Вам нравится обрабатывать короткие тексты?



Давайте
проведем
опрос!



Давайте
добавим
“открытый”
вопрос

Чем больше
ответов, тем
сложнее
и дольше

Home Insert Page layout Data Macros Protection View			
B16	Σ	fx	=IF(OR(ISNUMBER(SEARCH(B\$5; \$A16)); ISNUMBER(SEARCH(B\$2; \$A16)); ISNUMBER(SEARCH(B\$4; \$A16))); 1; "")
	A	B	
1	Что стало для вас основным поводом посетить WIAD?	нетворкинг	ин
2		встре	жа
3		знакомств	же
4		нетво	ин
5		движ	лю
16	Темы докладов, встреча со старыми знакомыми)		1
17	у вас всегда классно		
18	доклад делала		
19	встреча с важными людьми (слабо относится к конференции)		1
20	сложность, специализированность докладов		
21	плохое качество записи предыдущих WIAD		
22	интерес к смежной теме, другой взгляд на продукты		
23	была в прошлом году, понравились доклады		
24	доклады коллеги		
25	интерес к деятельности коллег		
26	жажда знаний		

Кирилл Улитин: От ключевиков до LLM

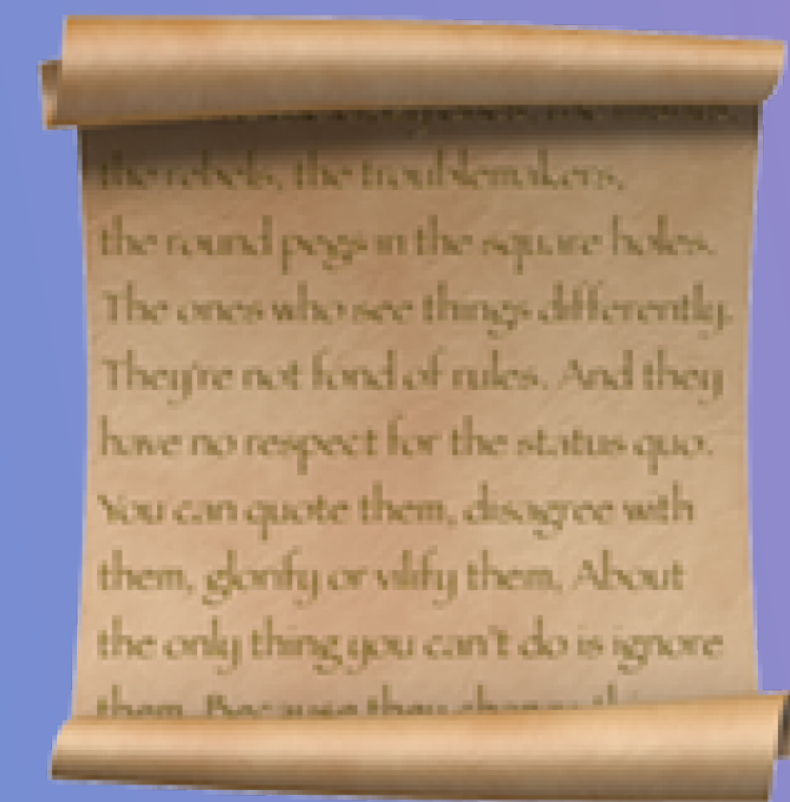
База

LLM

BERT

Ключевики

Итоги



Как было до 2017 года

- Статистические подходы, не понимают контекст и связи слов, зависят от списка стоп-слов.
- Старые нейросети (Word2Vec, CNN, RNN итд.) для кластеризации, хорошо только с очень короткими фразами, не учитывали многозначность слов.

...но в последнее время
начали сильно раздражать
постоянные ошибки
в работа корзины,
когда добавленные товары
~~просто~~ исчезают,
~~и мне~~...

Attention is all you need

- 2017 архитектура новой нейросети трансформер
- “Видит” связи всех слов со всеми
- Подробнее о трансформерах [1](#) [2](#) [3](#)

команда
изменила
кнопку
потому
что
она
изучила
карту
кликов

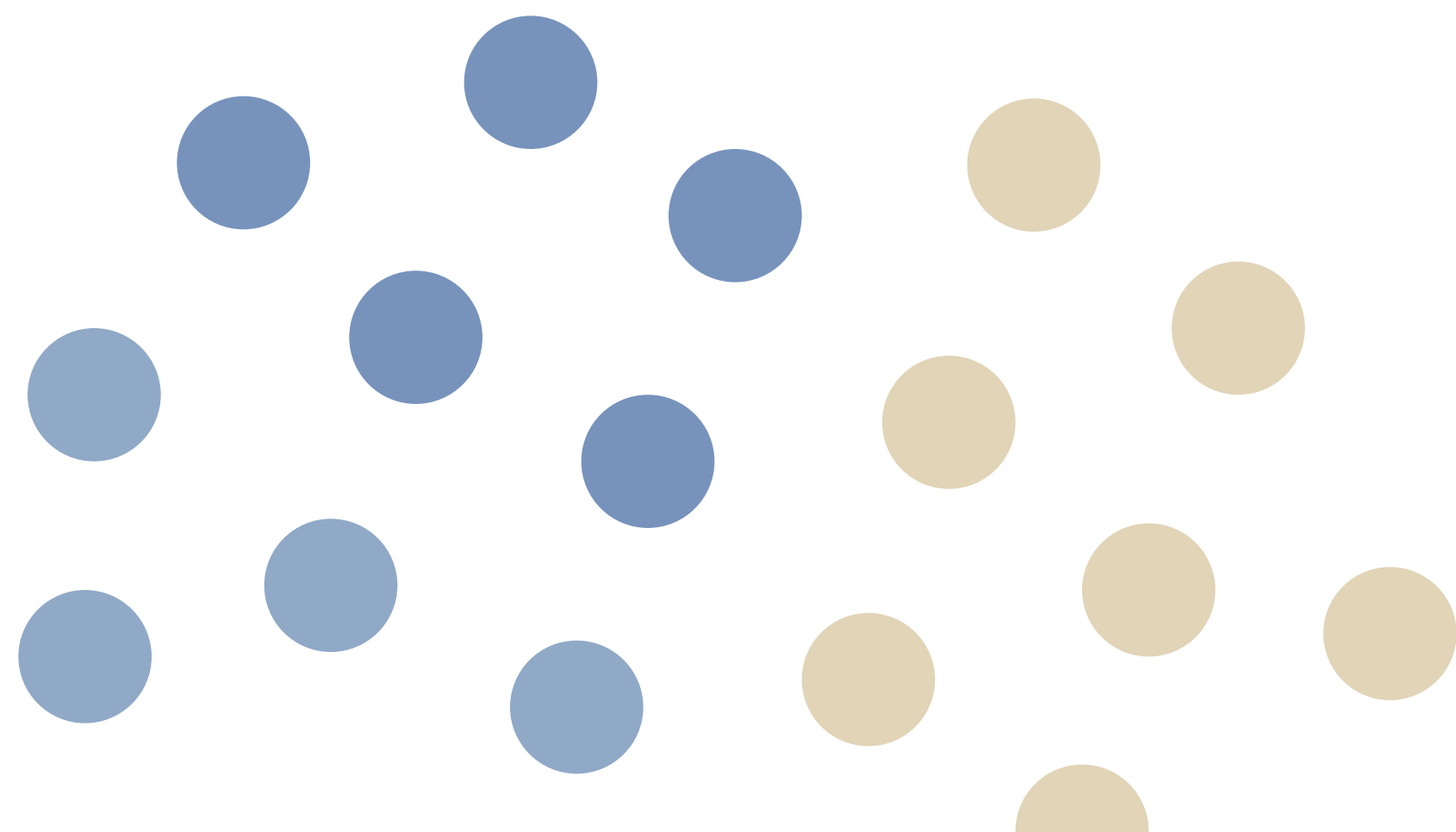
команда
изменила
кнопку
потому
что
она (команда)
изучила
карту
кликов

команда
изменила
кнопку
потому
что
она (кнопка)
была
не заметна
пользователям

команда
изменила
кнопку
потому
что
она
была
не заметна
пользователям

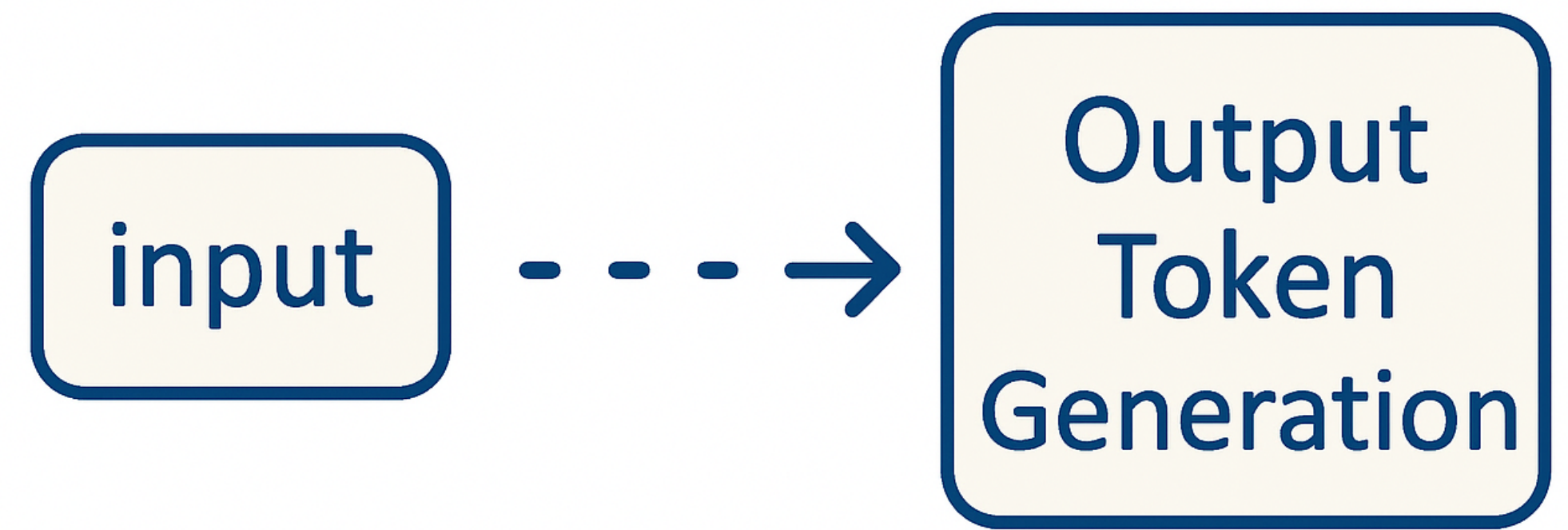
BERT

- Bidirectional Encoder Representations from Transformers
- Кодировщик трансформера
- Понимания языка (NLU), классификация, NER, поиск, как часть RAG



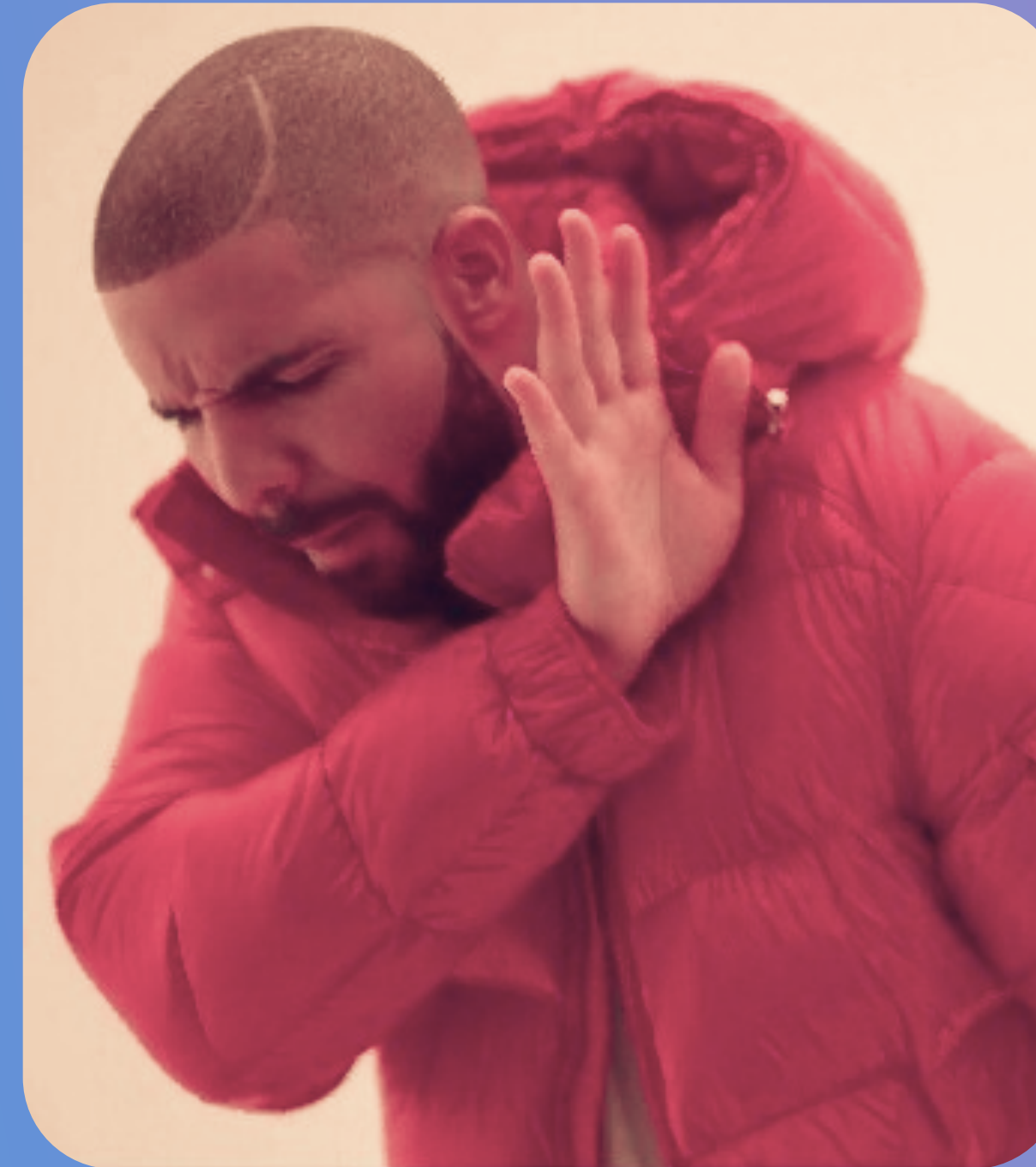
LLM (GPT)

- Generative Pre-trained Transformer
- Декодировщик трансформера
- Генерация текста



Кирилл Улитин: От ключевиков до LLM

**Вам нравится
обрабатывать
короткие тексты?**



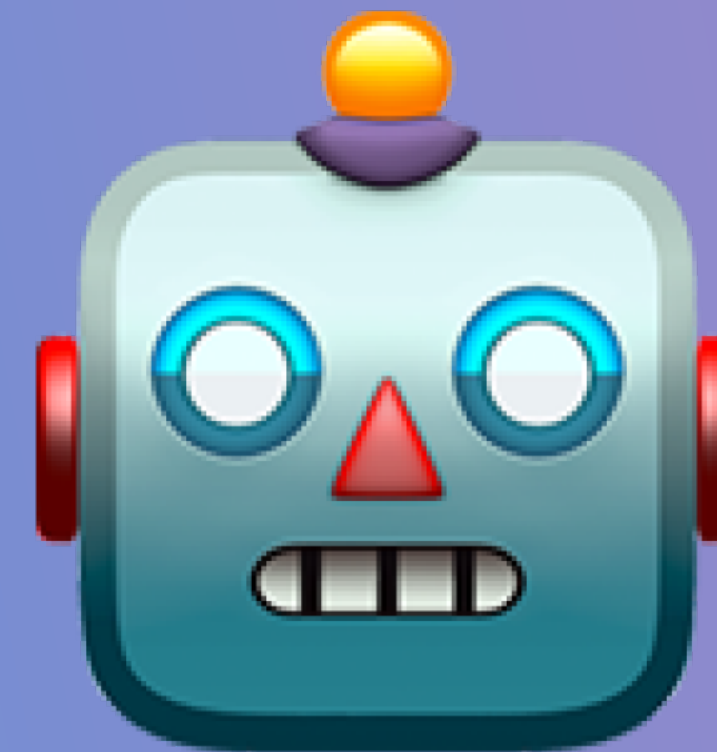
**Обработка
вручную**



**Пойду
в ChatGPT!**

Кирилл Улитин: От ключевиков до LLM

Кластеризация с помощью LLM



Посчитаем повторы в ответах- ассоциациях с помощью ChatGPT?

Привет, рандом!

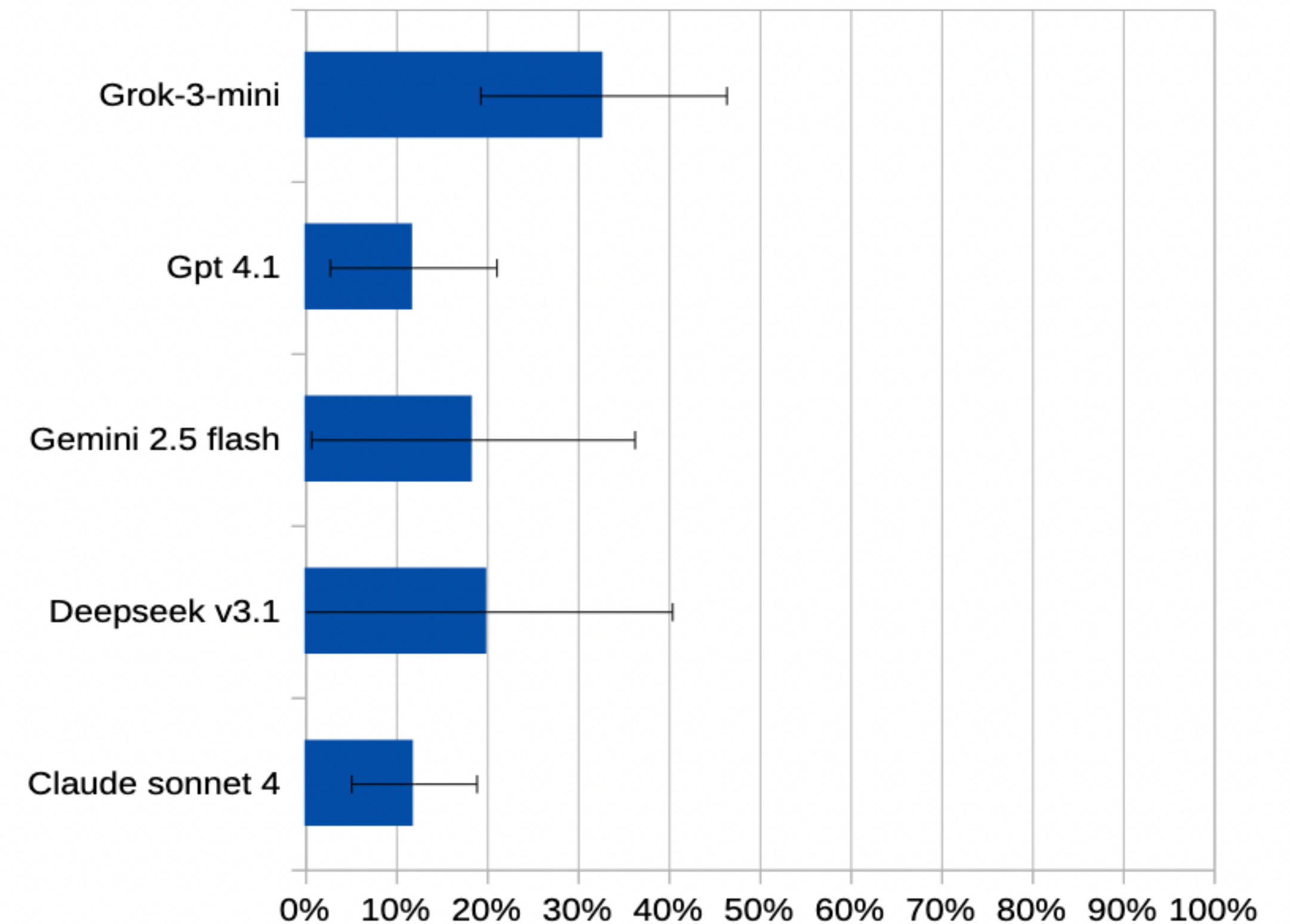
1	калькулятор	23
2	поднос	20
3	игра	13
4	детский	10
5	квадрат	6
6	джойстик	6
7	тетрис	6
8	конструктор	5
9		5



**Ср. ошибка
подсчета
в синт. тесте
19% SD 15%**

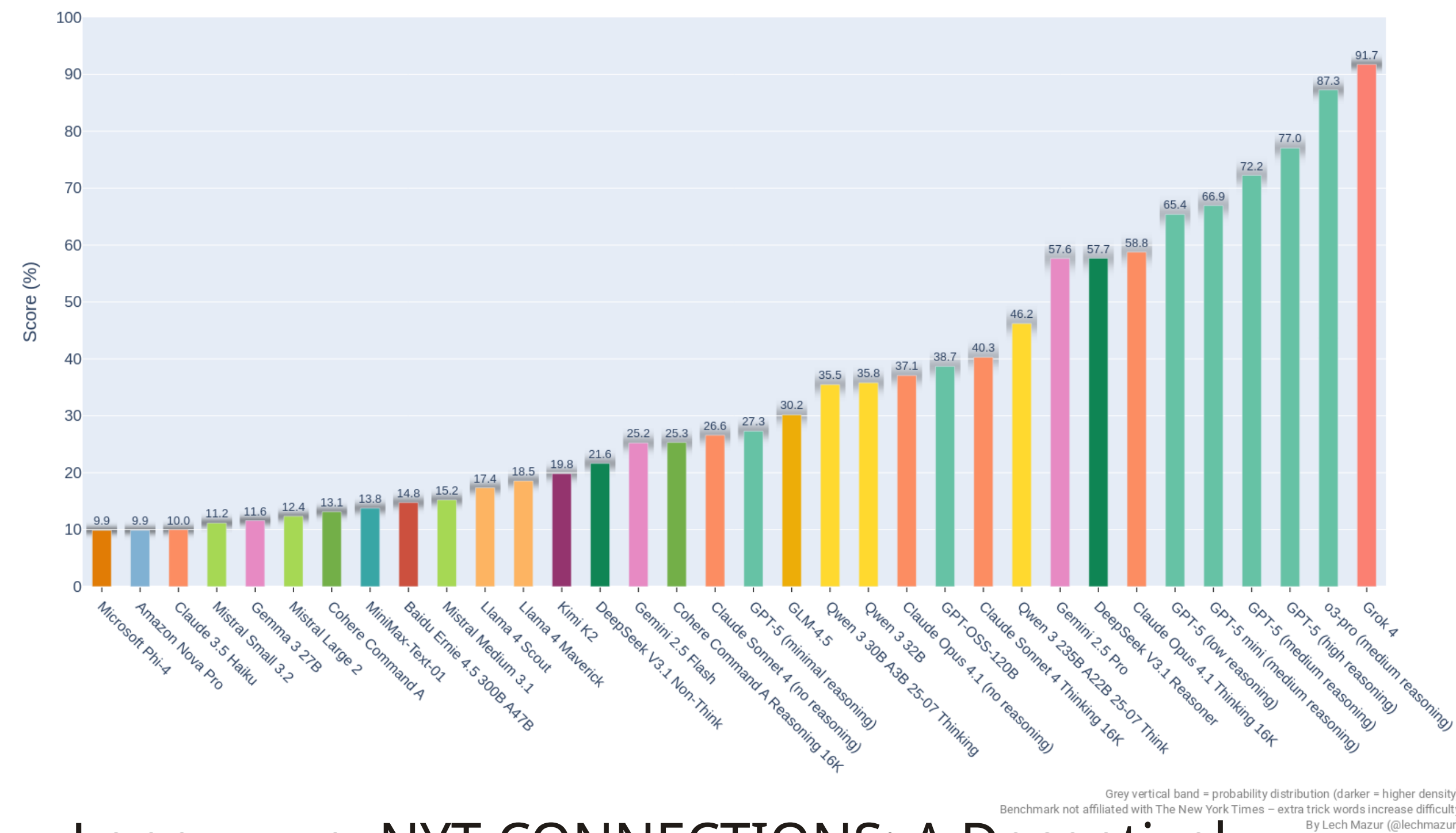
Абс. средняя разница между фактом и результатом llm

случайный список 300 строк из 12 слов, 30 итераций

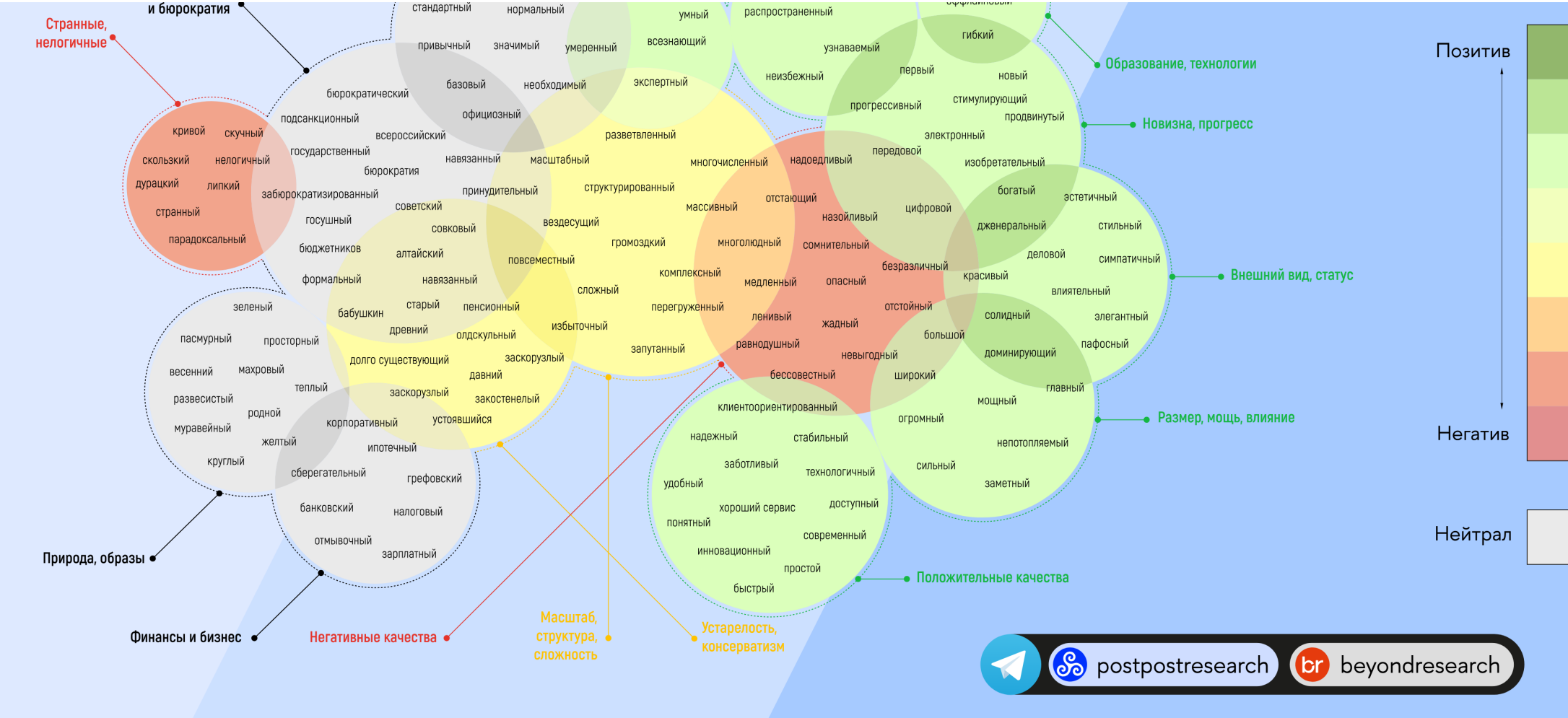


А качество?

Extended Word Connections: Scoreboard (759 puzzles)



Lopez и др. NYT-CONNECTIONS: A Deceptively Simple Text Classification Task that Stumps System-1 Thinkers



Жичкина, Ефимов, Боде - Заменит ли AI маркетинговых исследователей?

LLM ~~кластеринг~~ классификация

Chen, Guoxiu - Text Clustering as Classification with LLMs



**Генерация
тем (GPT)**



**Проверка,
обобщение
тем**



**Классификация
данных (GPT)**



**Подсчет
кластеров**

Контекст	Ниже ответы на вопрос ...
Задача	оцени каждую строку с ответом по наличию в ней следующих тем ...список тем...
Результат	Результат отдай в виде csv, где первая колонка это ответ, а последующие это темы. Укажи 1, если тема есть в ответе и 0 если ее там нет. Для разделителя используй точку с запятой
Примеры	
Данные	...список ответов...

csv

Ответ;Калькулятор;поднос;игра;детский;ке
Мои документы;0;0;0;0;0;0;0;0;0;0
Доходы;0;0;0;0;0;0;0;0;0;0
Обозначение офиса;0;0;0;0;0;0;0;0;0;0
Иконка калькулятора;1;0;0;0;0;0;0;0;0;0
Калькулятор без цифр;1;0;0;0;0;0;0;0;0;0
Не понятно;0;0;0;0;0;0;0;0;0;0
Приложение;0;0;0;0;0;0;0;0;0;0
Калькулятор;1;0;0;0;0;0;0;0;0;0
Отлично;0;0;0;0;0;0;0;0;0;0
Документы;0;0;0;0;0;0;0;0;0;0
Универсальное приложение;0;0;0;0;0;0;0;0;0;0
Калькулятор;1;0;0;0;0;0;0;0;0;0
Пусто пусто плюс;0;0;0;0;0;0;0;0;0;0
Зона отдыха;0;0;0;0;0;0;0;0;0;0

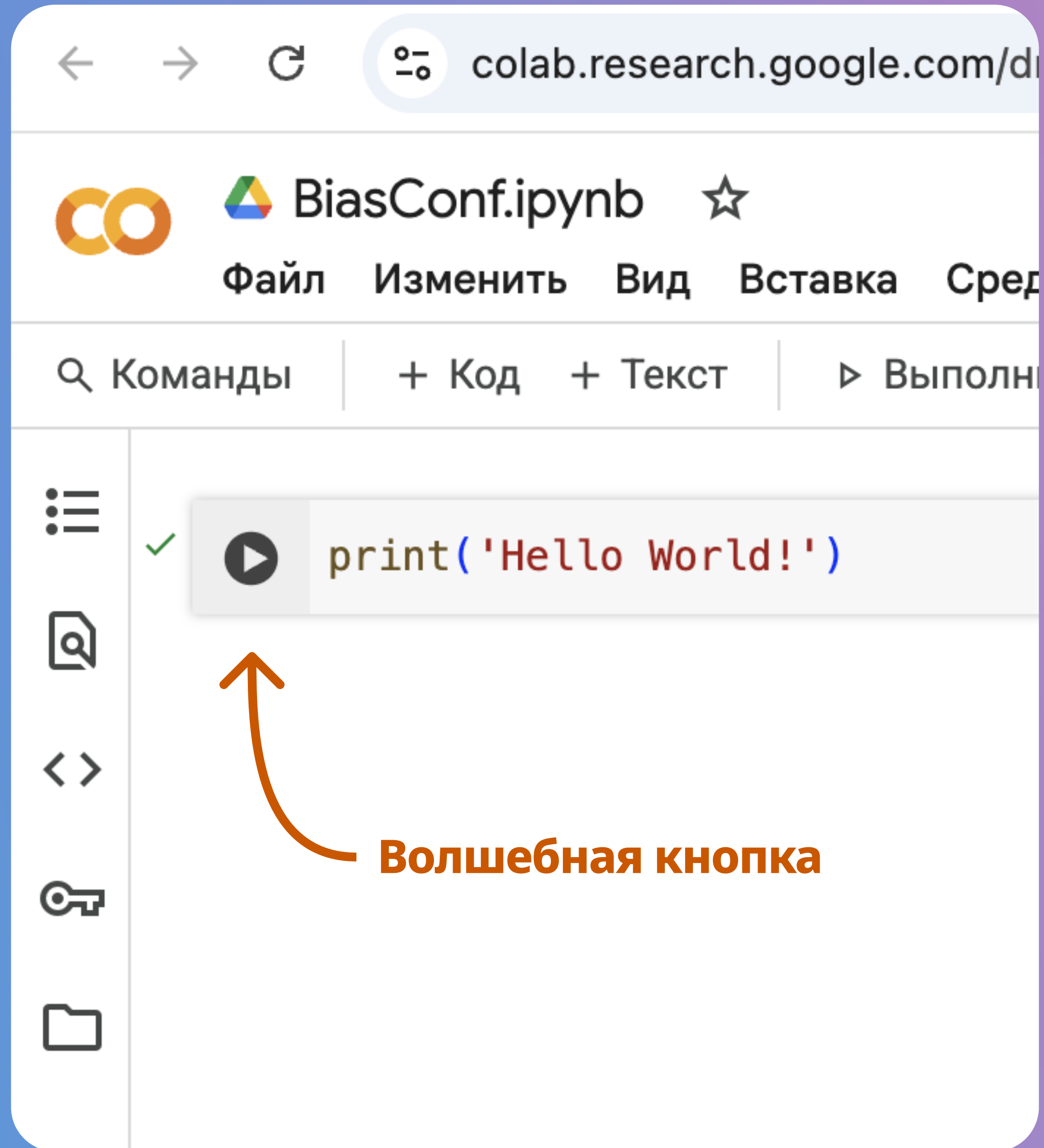
Выводы: LLM

- **Проверять количество токенов в запросе**, если запрос не влезает в ограничение, то он обрезается и мы можем получить частичный ответ
- **Предпочитать чистое api** из-за влияния системных промптов, которые добавляют невидимые инструкции
- **Использовать построчную классификацию вместо прямого подсчета**, так как он производит результаты с большой ошибкой
- **Помнить про приватность данных**, если использовать LLM локально дорого, сложно, то использовать другие методы

Кирилл Улитин: От ключевиков до LLM

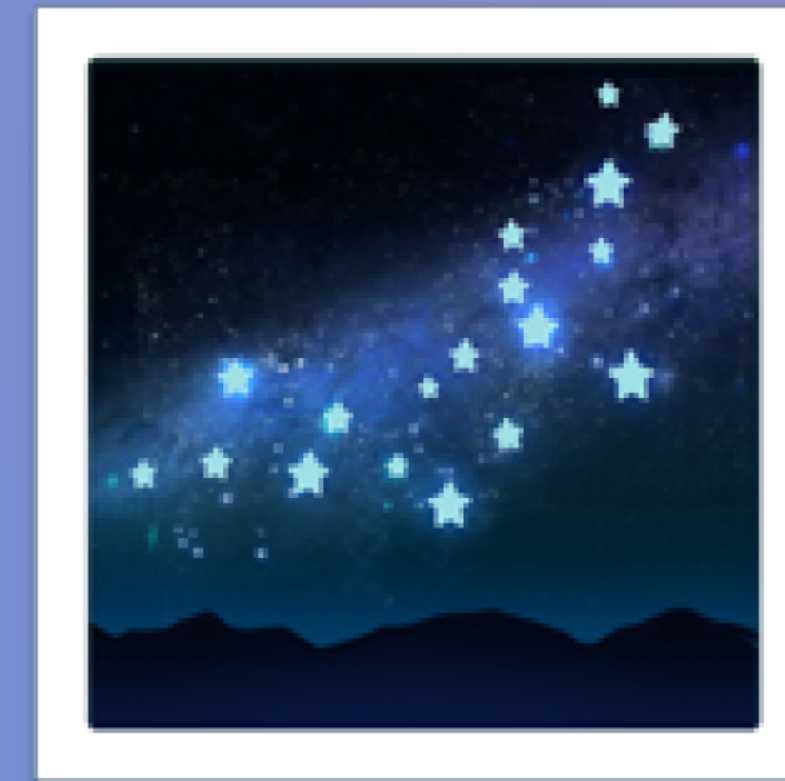
Быстро, приватно, не дорого?

- Python блокноты как инструмент анализа неструктурированных данных
- Онлайн: [Google Colab](#)
- На своем устройстве: [JupyterLab Desktop](#)
- Курс [pycode.ru](#)



Кирилл Улитин: От ключевиков до LLM

Кластеризация с помощью Sentence BERT моделей

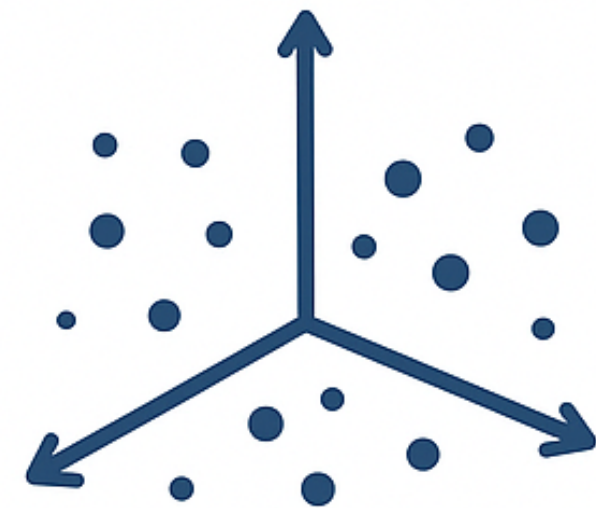




Sentence BERT кластеризация



**Загрузка
данных
в Sent. BERT
модель**



**Получение
векторного
представле
ния текстов**

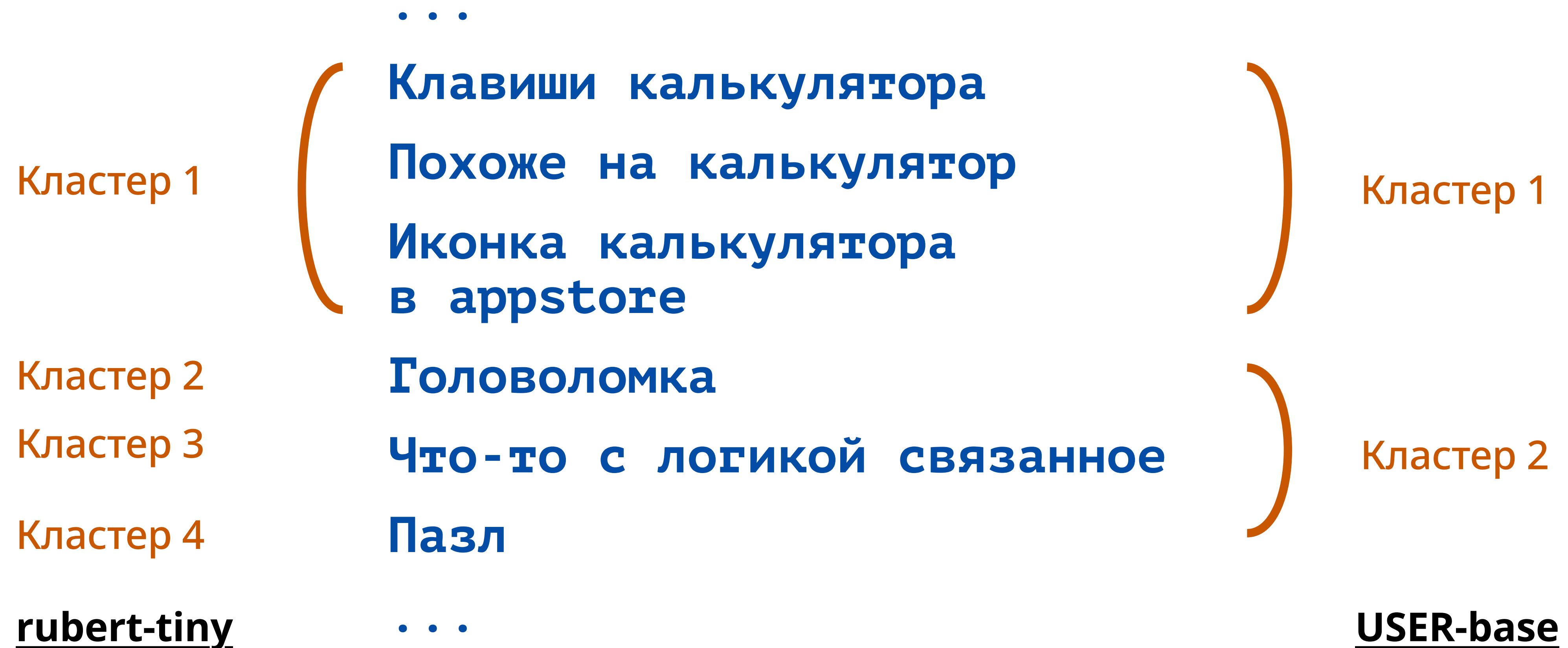


**Кластеризация
текстов
готовыми
алгоритмами**



**Подсчет
кластеров**

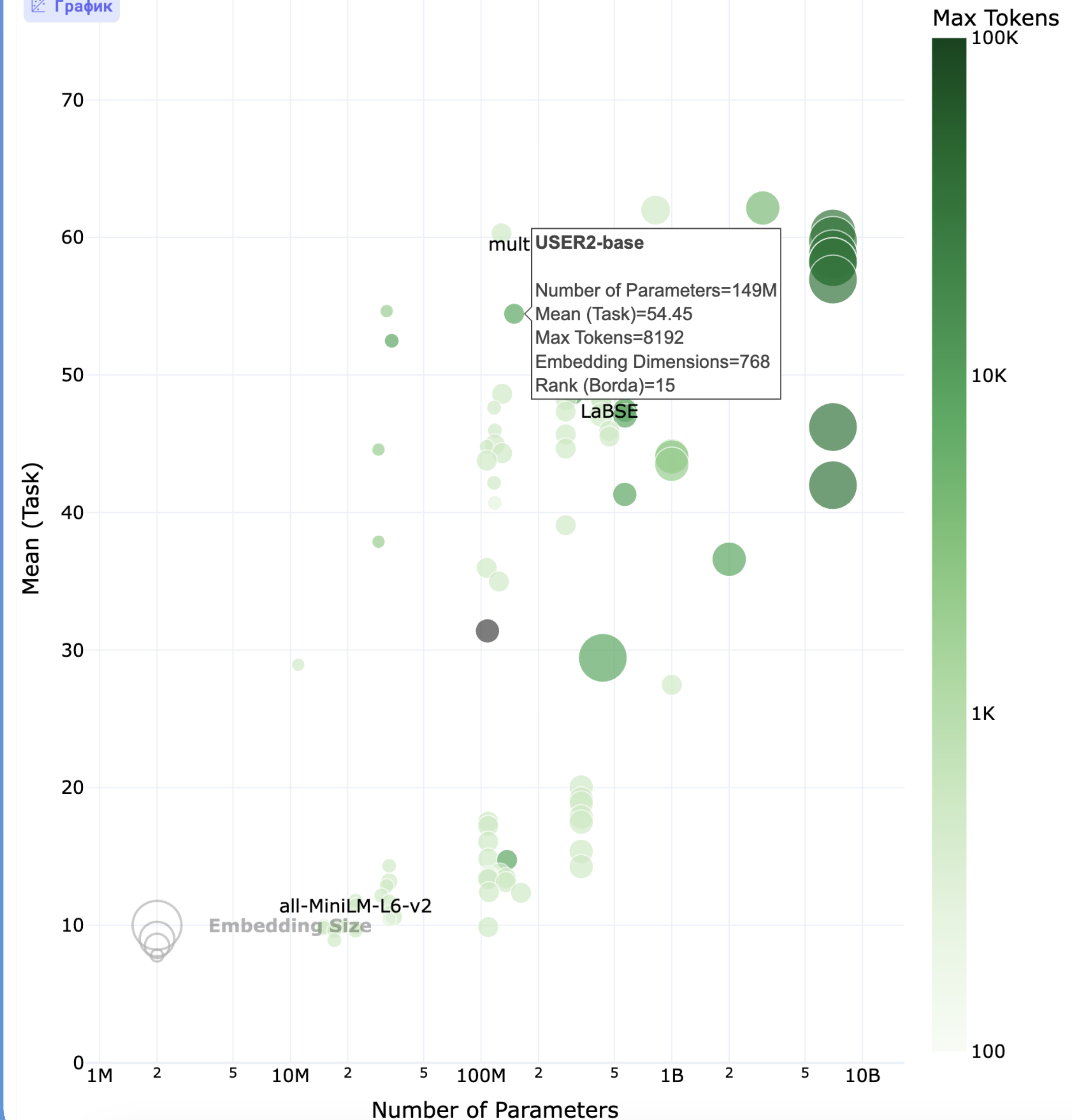
Влияние моделей



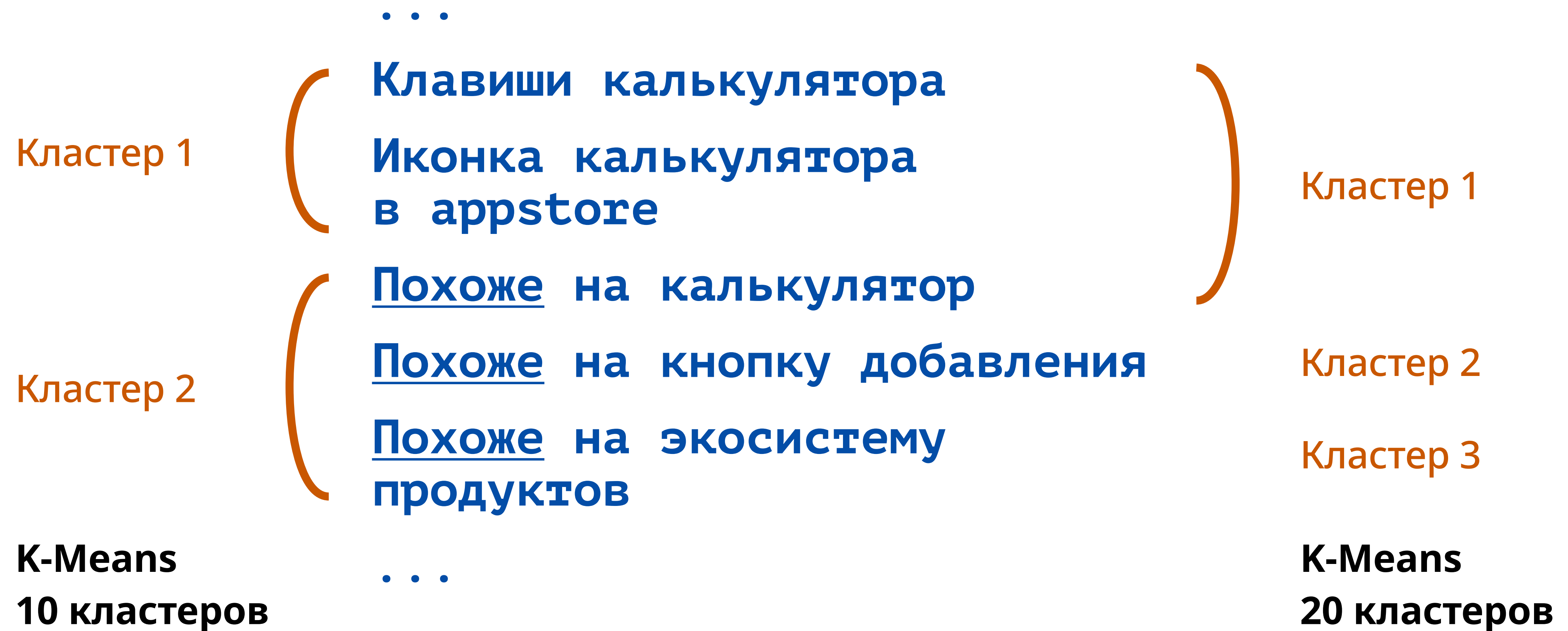
Где искать модели?

- HuggingFace (Бенчмарк MTEB)
- SentenceTransformer
- Русский язык или многоязычная
- Максимальный размер токенов

График



Влияние параметров кластеризации



Chat Intents

- Библиотека в которой реализован автоматический подбор параметров кластеризации
- <https://github.com/dborrelli/chat-intents>
Статья

```
import chatintents
from chatintents import ChatIntents
from hyperopt import hp
from sentence_transformers import SentenceTransformer

st = SentenceTransformer('deepvk/USER-base')
embeddings = st.encode(list(df['text']))

model = ChatIntents(embeddings, 'model1')

hspace = {
    "n_neighbors": hp.choice('n_neighbors', range(3,16)),
    "n_components": hp.choice('n_components', range(3,16)),
    "min_cluster_size": hp.choice('min_cluster_size', range(2,16)),
    "min_samples": None,
    "random_state": 42
}

label_lower = 20
label_upper = 50
max_evals = 100

model.bayesian_search(space=hspace, label_lower=label_lower, label_upper=label_upper,
                      max_evals=max_evals)

model.best_params

df_summary, labeled_docs = model.apply_and_summarize_labels(df[['text']])
display(labeled_docs)
labeled_docs.to_excel('results.xlsx')
```

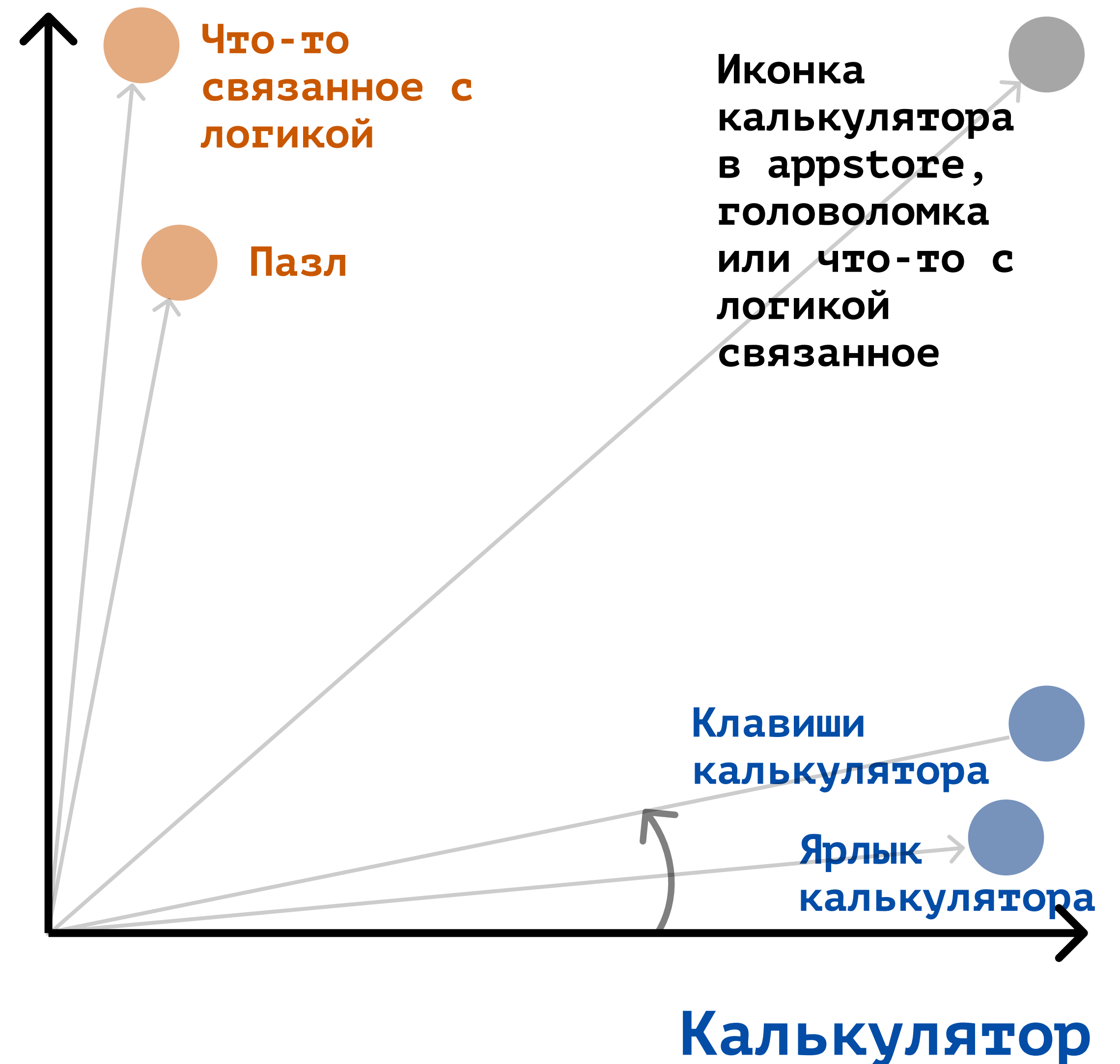
100%|██████████| 100/100 [00:53<00:00, 1.87trial/s, best loss: 0.05785123966942149]
best:
{'min_cluster_size': 3, 'min_samples': None, 'n_components': 11, 'n_neighbors': 6, 'random_state': 42}
label count: 36

	text	label_st1	label
0	Мои документы	25	документы_фото_любой
1	Доходы	25	документы_фото_любой
2	Обозначение офиса\n\n	-1	игра_включить_сложения_похоже

Качество данных

- Каждому тексту Sent. BERT присваивает один общий вектор, чем больше там разных тем, тем хуже результат
- **Решение** Если возможно использовать множественные поля ввода
- **Решение** Явно указывать как разделить несколько мыслей в ответах (например, через запятую, на новой строке итп.)

Головоломка



Метрика успеха

- Добавим параметр согласованности кластера (cos_sim.)
- Оставляем кластера у которых согласованность выше X
- Повторим, пока процент неразобранного не остался минимальным
- Объединим дубликаты кластеров

Высокий ср. cos sim. (83)

разбить текст на колонки
текст по колонкам
колонок созданных в других
...

Низкий ср. cos sim. (47)

работа с полями данных openxml в документах как в ms
использование формул в ячейках для запуска макроса
связь макросов с внешними файлами и настройками
...

Выводы: BERT

- **Проверять количество токенов в запросе**, BERT модели именю сильно меньший контекст 512-4k токенов для каждого текста
- **Выбирать модель под задачу** на основе данных на которых (до)обучена модель
- **В идеале один текст = одна мысль**, давать инструкции для респондентов или дополнительно разделять данные
- **Пробовать разные настройки алгоритмов, валидировать результаты**

Кирилл Улитин: От ключевиков до LLM

Кластеризация с помощью ключевых слов

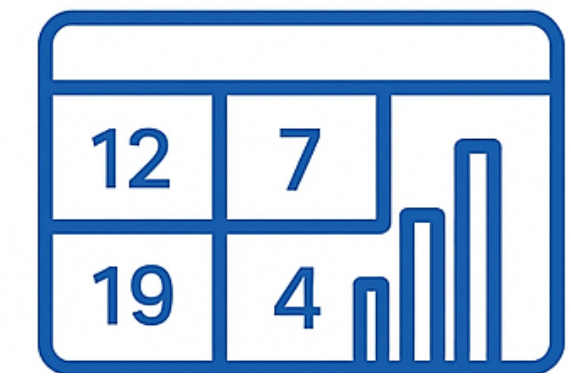


Популярные библиотеки

- **RAKE** 2014 — статистический метод, очень быстрый, на основе частоты слов и стоп-слов
- **YAKE** 2018 — более современный статистический метод, учитывает еще регистр, позицию слова в тексте и другие лингвистические особенности
- **KeyBERT** 2020 — семантический метод, который использует BERT модели



**Загрузка
данных в
библиотеку**



**Подсчет
кластеров
на основе
ключевых слов**

Исходный текст

Подсчет функциональности и работы сотрудников из других филиалов. Аналитическая таблица для анализа каждым своих ошибок и недочетов. Указываю, кто сколько выполнил плана в месяц.

RAKE

работы сотрудников
подсчет функциональности
других филиалов
аналитическая таблица
указываю

YAKE

Подсчет функциональности
функциональности и работы
работы сотрудников
Подсчет
филиалов

KeyBERT

подсчет функциональности
аналитическая таблица
таблица анализа
филиалов аналитическая
сотрудников других



Очень короткие тексты

- Алгоритм не может найти хотябы одно ключевое слово, если в тексте нет “мяса”
- Решение 1** объединить все тексты, получить ключевые слова, а потом посчитать их повторяемость
- Решение 2** Использовать ngram и KeyBERT с анализом кандидатов

```
rake = Rake(  
    language_code='ru',      # Язык текста  
    max_words=1,            # Макс. слов в ключевой фразе  
    min_freq=1,             # Минимальная частота слова  
)  
  
# Функция для извлечения топ-3 ключевых слов  
def extract_keywords(text, n=3):  
    keywords = rake.apply(text)[:n] # Берем топ-N ключевых слов  
    return ', '.join([kw[0] for kw in keywords]) # Возвращаем строку  
  
# Применяем функцию к колонке 'text'  
df['keywords'] = df['text'].apply(extract_keywords)  
display(df[['text', 'keywords']])
```

	text	keywords
1	Мои документы	документы
3	Доходы	доходы
5	Обозначение офиса\n\n	
8	Иконка калькулятора	
10	Калькулятор без цифр	калькулятор, цифр
...	...	...
474	"МойОфис Стандартный"	

Ключевые слова помогают очистить данные

'возможность работы',
'возможность',
'функциональную',
'функциональность',
'функциональностей',
'функциональности',
'поддержка',
'поддержать',
'поддерживать',
'предоставлять',
'должен',
'добавить',
'реализовать',
'мойофис для дома',
'мойофис таблица',
'мойофис текст',
'мой офис таблица',
'мой офис текст',
'мойОфис профессиональный',
'мойофис стандартный домашняя версия',
'мойофис стандартный домашняя',

Выводы: ключевики

- **Формы слов и стоп-слова влияют на результат**, пробовать приводить текст к одной форме, оценивать влияние стоп-слов
- **Сравнивать разные алгоритмы для ваших данных**, что бы выбрать оптимальный вариант так как результаты могут отличаться
- **Использовать классификацию для очень коротких текстов** на основе ключевиков из всего набора данных и последующей разметкой через KeyBert

Кирилл Улитин: От ключевиков до LLM

Итоги



Для задач кластеризации	Ключевые слова	Sentence. BERT	LLM
Понимание контекста	🧠/🧠🧠 (*KeyBERT)	🧠🧠	🧠🧠🧠
Простота	😊😊	😊	😊😊😊
Скорость	🚀🚀🚀 / 🚀🚀 *	🚀🚀	🚀
Дешевизна	💰💰💰 / 💰💰 *	💰💰	💰
Контроль результата	✅	✅	❌
Офлайн	✅	✅	✅ (Ресурсы)
Большие данные	✅	✅	✅ (Частями)

Конец

- Не просите LLM считать – просите классифицировать
- Подбирайте методы под задачу, комбинируйте их
- Python блокноты это универсальный инструмент для работы с данными

**Слайды и интерактивные блокноты
с подходами из доклада:**



https://t.me/ulitin_ru/125