Тематическое моделирование K-means в travel-аналитике:

оптимизация числа кластеров

Слайд 1: Общая информация

Входные данные:

Датасет: Go Russia.csv (тексты телеграм-канала "Путешествия по России. Go Russia")

Объём данных:

· Общее количество текстов: 2849

· Средняя длина текста в символах: 381.06

· Среднее количество слов в тексте: 54.08

· Среднее количество уникальных слов: 49.35

· Среднее лексическое разнообразие: 0.9268

· Средняя длина слова: 6.09 символов

Модель эмбеддингов:

Sentence Transformer
(paraphrase-multilingual-MiniLM-L12-v2)

Метод кластеризации: K-means

- Модель с 4 кластерами (назовем её "Model4K")
- Модель с 5 кластерами ("Model5K")

```
df. head(250)
```

	Заголовок	Текст	Дата публикации	Эмодзи	Локация на карте	Название региона
0	Работа на Юге. ТОП-15 Вакансий	Строительная сфера на юге развивается семимиль...	23.07.2023 10:13:35 UTC+04:00	{'👍': 23, '❤️': 1, '👏': 1, '😁': 1}		
1	Куда дети перевозят своих родителей? Где прове...	Климат морской полезен для всех. Очень часто м...	24.07.2023 11:45:50 UTC+04:00	{'👍': 23, '🛋️': 3}		
2	Головой на море, "одним местом" на диване.	Если вы в эти дни не на морене расстраивайтесь...	25.07.2023 12:48:20 UTC+04:00	{'👍': 11, '❤️': 2, '🤔': 1}		
3	Лучший город России. Рейтинг	Ты все ещё ищешь лучший город России? Блогер-у...	26.07.2023 14:52:33 UTC+04:00	{'👍': 53, '👎': 13, '❤️': 5}		
4	Жить у моря полезно.	Я тут наткнулся на статью про воду там ученые ...	27.07.2023 10:55:21 UTC+04:00	{'👍': 17, '❤️': 2, '🔥': 1}		
...	...	...	...	...	...	...
245	Уксинская озовая гряда – чудо природы, оставше...	Это одно из самых красивых мест Карелии ! Ледн...	26.09.2023 17:36:07 UTC+04:00	{'❤️': 268, '👍': 182, '🔥': 38, '🤩': 9, '😁': 3, ...	https://yandex.ru/maps?whatshere%5Bpoint%5D=31...	#республикакарелия
246	Необычные Сырные скалы в Карачаево-Черкессии.	В Карачаево-Черкесии есть одно необычное место...	26.09.2023 19:18:01 UTC+04:00	{'🔥': 260, '👍': 170, '❤️': 42, '❤️🔥': 11, '🤔': ...	https://yandex.ru/maps/org/syrnaya_peshchera/1...	#карачаевочеркесия
247	Усадьба Нероново – погибающее имперское наследие.	Усадьба Нероново относится к числу наиболее ин...	26.09.2023 22:07:06 UTC+04:00	{'❤️': 249, '😢': 204, '👍': 91, '🔥': 30, '😱': 10...	https://yandex.ru/maps/org/usadba_neronovo/202...	#костромскаяобласть
248	Особняк Василия Каншина, содержателя питейных ...	Василий Семёнович Каншин был богатейшим челове...	27.09.2023 09:07:01 UTC+04:00	{'👍': 320, '❤️🔥': 52, '❤️': 32, '🔥': 23, '🤔': 2}	https://yandex.ru/maps?whatshere%5Bpoint%5D=30...	#санктпетербург
249	Горнолыжный курорт «Матлас	в Дагестане. На горнолыжном курорте Матлас раз...	27.09.2023 12:14:01 UTC+04:00	{'👍': 238, '😇': 78, '🔥': 41, '❤️': 31, '😨': 9, ...	https://yandex.ru/maps/org/gornolyzhny_kurort_...	#республикадагестан

250 rows × 6 columns

Слайд 2: Задачи проекта

1. **Оценить чёткость тематического разделения:**
   - Анализ ключевых слов и примеров заголовков
2. **Проверить однородность кластеров:**
   - Отсутствие тематических конфликтов внутри кластеров
3. **Выявить перекрытие тематик:**
   - Сравнение распределения объектов между кластерами
4. **Сравнить модели:**
   - 4 vs 5 кластеров по чёткости и сбалансированности
5. **Определить практическое применение:**
   - Решения для туризма и рекламы

Слайд 3: Методы и визуализация

Определение числа кластеров:

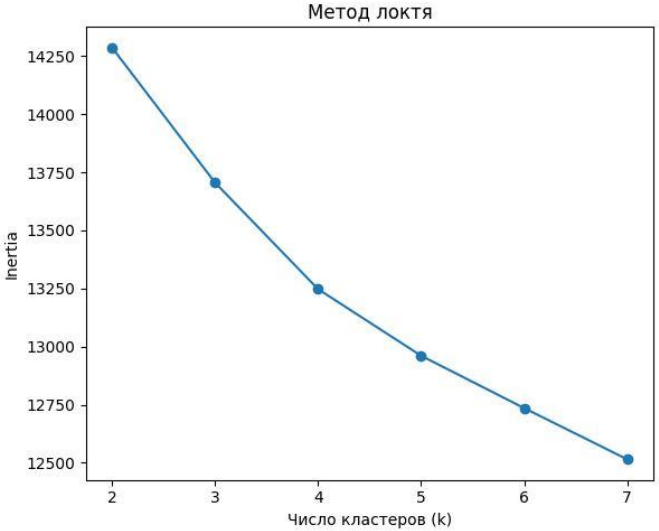
Метод локтя: после k=5 снижение инерции становится более линейным

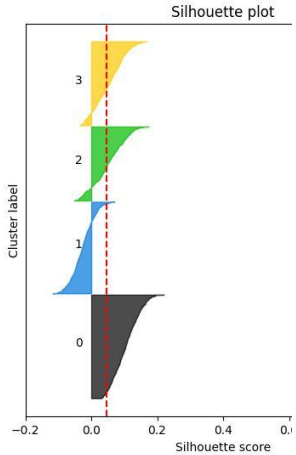
Силуэтный анализ: 5 кластеров (0.051) демонстрируют лучшее качество vs 4 кластера (0.047)

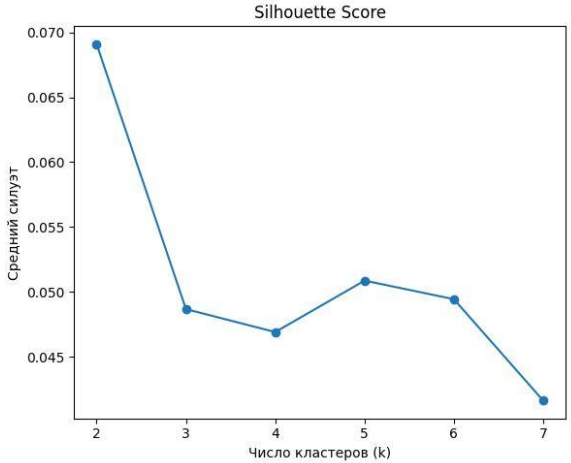
Важные замечания:

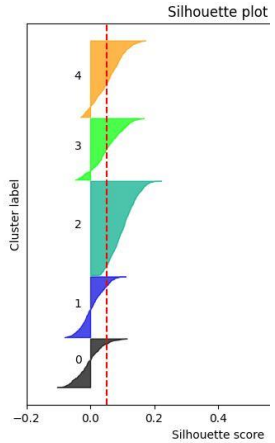
Общая проблема качества: все значения силуэта <0.1 указывают на слабую структуру кластеров

Интерпретация + метрики



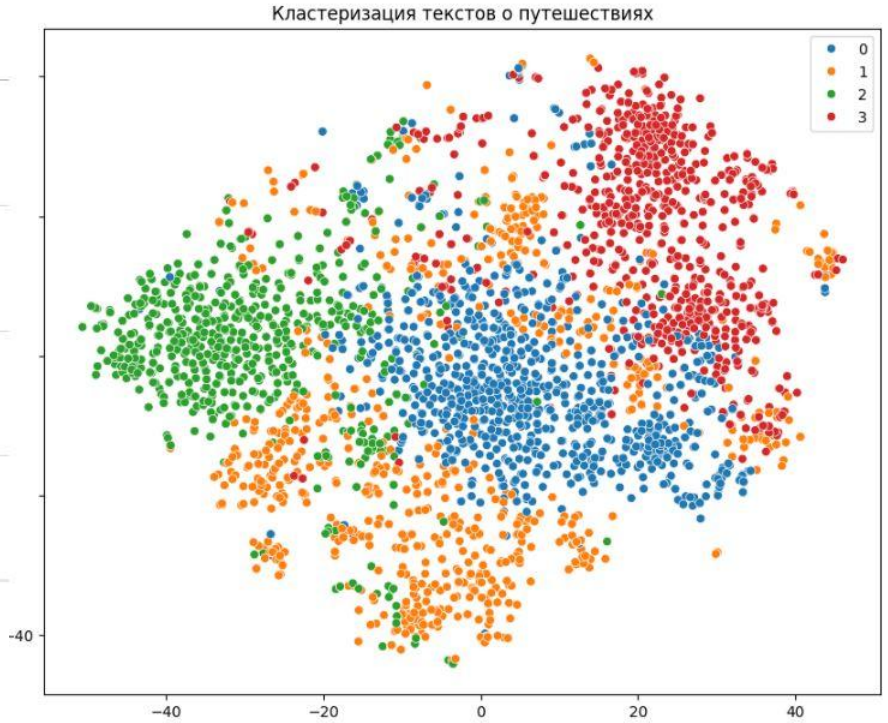






Слайд 4: Модель 4K (4 кластера)

Кластер	Тематика	Доля	Примеры заголовков
0	Горные ландшафты	29%	"Долина гейзеров", "Хребет Малые Бамбаки"
1	Городская экономика	26%	"Переезд в города", "Строительство трасс"
2	Культурное наследие	21%	"Суздаль", "Кинодеревня"
3	Водные объекты	24%	"Жизнь у моря", "Мыс Дооб"

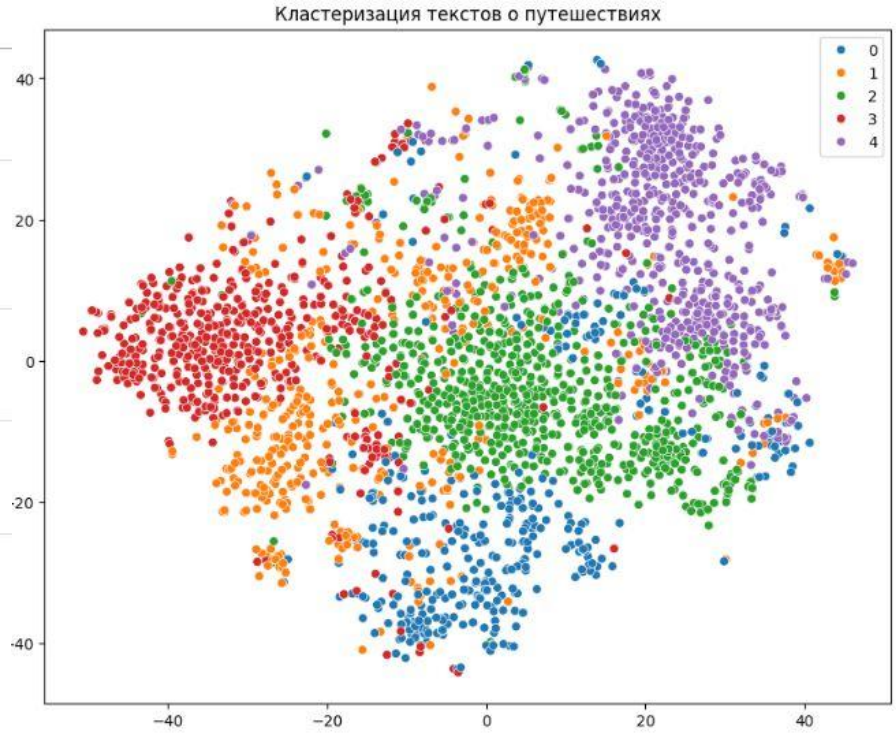


Преимущества:

- Чёткое разделение по темам
- Сбалансированность (разброс размеров: 8.3%)

Слайд 5: Модель 5K (5 кластеров)

Кластер	Тематика	Доля	Примеры заголовков
0	Природные пейзажи	14%	«Осень на Алтае», «Малиновый закат на Телецком»
1	Экономика туризма	18%	«Как россияне экономят на турах», «Строительство трассы М-12 от Москвы до Казани...»
2	Горы/водопады	28%	«Долина гейзеров», «Салтинский водопад»
3	Архитектура	18%	«Бранденбургские ворота», «Замок Гарибальди»
4	Водные объекты	22%	«Софийские озёра в Архызе», «Сплав по реке Ай»



Недостатки:

- Есть дисбаланс (разброс размеров: 13.3%)
- Тематическое перекрытие (горы ↔ вода: 0↔2, 2↔4)

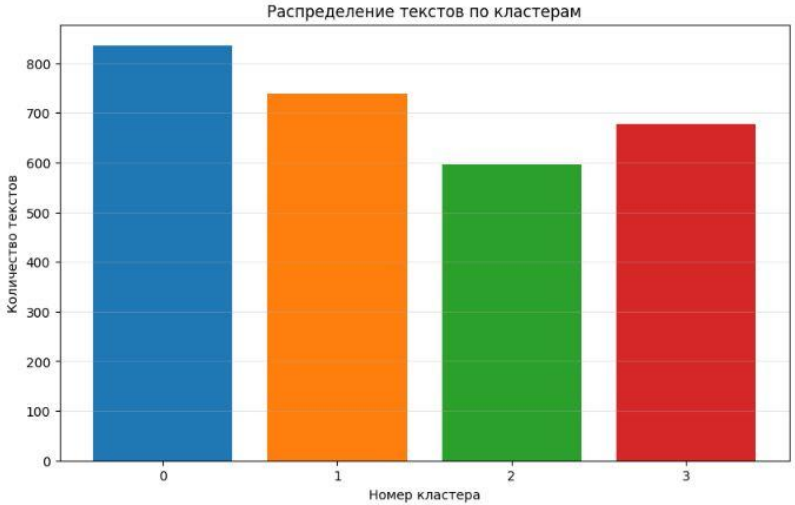
Слайд 6. Эмоциональный интеллект данных. Кластер 0 Модели 5К

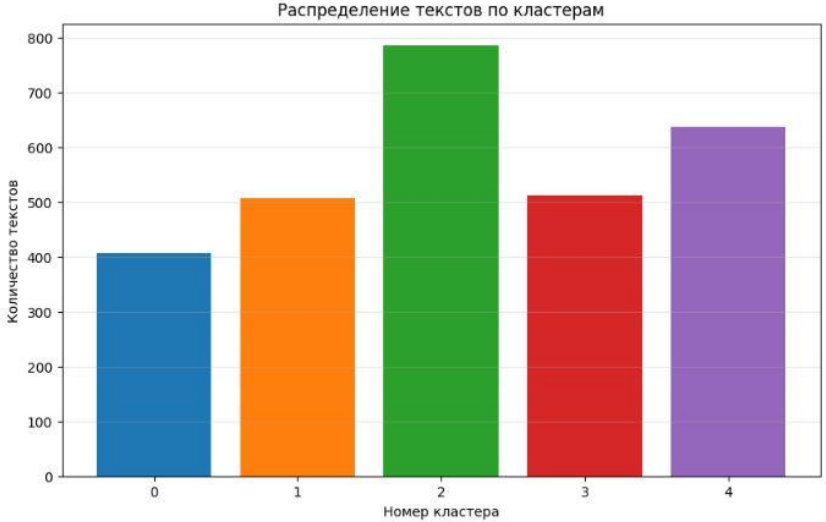
Аспект	Характеристики	Примеры
Суть кластера	• Фокус на *моментах* и *эмоциях* • Восприятие природы через эстетику времени (рассветы, сезоны) • Сильный оценочный компонент	*"Как же красиво в Адыгее"*, *"Ловим волну уходящего лета"*
Ключевые слова	• **Явления/время:** рассвет, закат, утро, вечер • **Сезоны:** осень, зима, снег • **Эмоции:** красивый, золотой, малиновый (в заголовках)	Топ-слова: *зима, рассвет, осень, закат, небо*
Уникальные черты	• Акцент на *визуальной трансформации* природы • Личный опыт (обращение к читателю) • Метафоры и поэтические сравнения	*"Осенняя сказка Домбая"*, *"Цветочные открытки"*, *"Лесная терапия"*
Тематическая роль	• "Поэзия места" • Контент для вдохновения • Фотогеничные состояния природы	**Интерпретация:** *визуальные впечатления и эмоциональные моменты путешествий*
Бизнес-применение	• Фототуры и мастер-классы • Эмоциональный маркетинг • Контент для Instagram/Pinterest • Премиум-позиционирование отелей "с видом"	**Пример:** *реклама тура "Золотые рассветы Алтая" для фотографов*

Слайд 7: Сравнительный анализ

Критерии сравнения:

Параметр	4K	5K
Чёткость тем	✓ Высокая	✗ Средняя
Сбалансированность	✓ 21-29%	✗ 14-28%
Отсутствие аномалий	✓ Нет смешения	✗ Есть смешение
Интерпретируемость	✓ Прозрачная	✗ Сложная





Вывод: 4K обеспечивает более логичное и чистое разделение тем, кластеры лучше сбалансированы.

- **Модель 4K**: природно-урбанистическая кластеризация
- **Модель 5K**: детализированная тематическая сегментация

Слайд 8: Применение модели 4К в Travel-бизнесе

Сфера	Бизнес-задачи	Конкретные примеры применения	Источники данных
Туроператоры	• Сегментация клиентов • Персонализация предложений • Разработка новых туров • Партнерские программы	• Рекомендация похода на Казбек (К0) любителям гор • Рассылка туров по Золотому кольцу (К2) ценителям истории • Тур "Алтай: горы (К0) + Телецкое озеро (К3)« • Сотрудничество с историческими гидами для гостей, интересующихся усадьбами (К2)	Ключевые слова К0: *гора, водопад* Примеры К2: "Суздаль", "Замок Гарибальди« Примеры К2: "Середниково", "Покровская пустынь"
Онлайн-платформы бронирования (Booking, Avito Путешествия, отели и глэмпинги)	• Персонализация поиска • Рубрикация контента	• Показ отелей в Сочи (К3) при запросе "пляжный отдых" • Раздел "Активный отдых" для статей о водопадах (К0)	Заголовки К3: "Дикий пляж Сосновка", "Бухта Ежовая" Ключевые слова К0: *скала, каньон, пляж, водопад*
Контент-платформы (Journal Travel, Яндекс.Путешествия)	• Анализ интересов аудитории • Таргетированная реклама • Тематическое позиционирование	• Статья "Топ-10 горных маршрутов" (К0) для 29.3% аудитории • Баннер от отеля в разделе "Морские курорты" (К3) • Слоган "ЭкоЛодж у водопада" (К0)	Доля кластеров: К0 (29.3%), К3 (23.8%) Пример К3: "Рай на краю света"

Слайд 9: Важность выбора количества кластеров (K)

Критерий влияния	Последствия выбора K
Глубина анализа	Слишком мало K → обобщенные кластеры (потеря нюансов) Слишком много K → искусственные группы
Маркетинг и персонализация	Неточный K → нерелевантные предложения Оптимальный K → точный таргетинг
Разработка продуктов	Ошибка K → пропуск нишевых возможностей Правильный K → создание востребованных услуг
Качество данных	Неверный K → искажение структуры интересов Оптимальный K → точные входные данные для ML-систем

Спасибо за внимание!