# Voroge Jeskela
# Использование ИИ в продукте

# Проблематика

При проведении опросов часто встречаются невалидные или бессмысленные ответы, что затрудняет выявление ключевых результатов из пользовательских данных. Такие ответы могут исказить общую картину и отвлечь внимание от действительно значимых тенденций. Кроме того, важно получать обработанные ответы в удобном формате, который будет легко воспринимать и анализировать. Эффективная визуализация данных может значительно облегчить процесс интерпретации результатов. В конечном итоге, качественный анализ ответов способствует более глубокому пониманию потребностей и предпочтений целевой аудитории.

# Задача

Разработать систему для анализа ответов и визуализации результатов анализа

# Цели

- Изучить входные данные и конвертировать их в удобный формат
- Разработать систему предобработки пользовательских ответов
- Реализовать алгоритм выделения смысловых групп

# Процесс реализации

- Форматирование файла ответов пользователей
- Препроцессинг данных
- Реализация модели машинного обучения
- Формирование облака слов
- Создание телеграм-бота

# Форматирование входных данных

3 типа принимаемых файлов (аналогично Yandex Forms):

1) .csv
2) .xlsx
3) .json

В качестве .json файла было выбрано специально представление:

```
[
[["key_1", "val_1"], ["key_2", "val_2"], ...], // item1
[["key_1", "val_1"], ["key_2", "val_2"], ...], // item2
...
]
```

# Препроцессинг данных

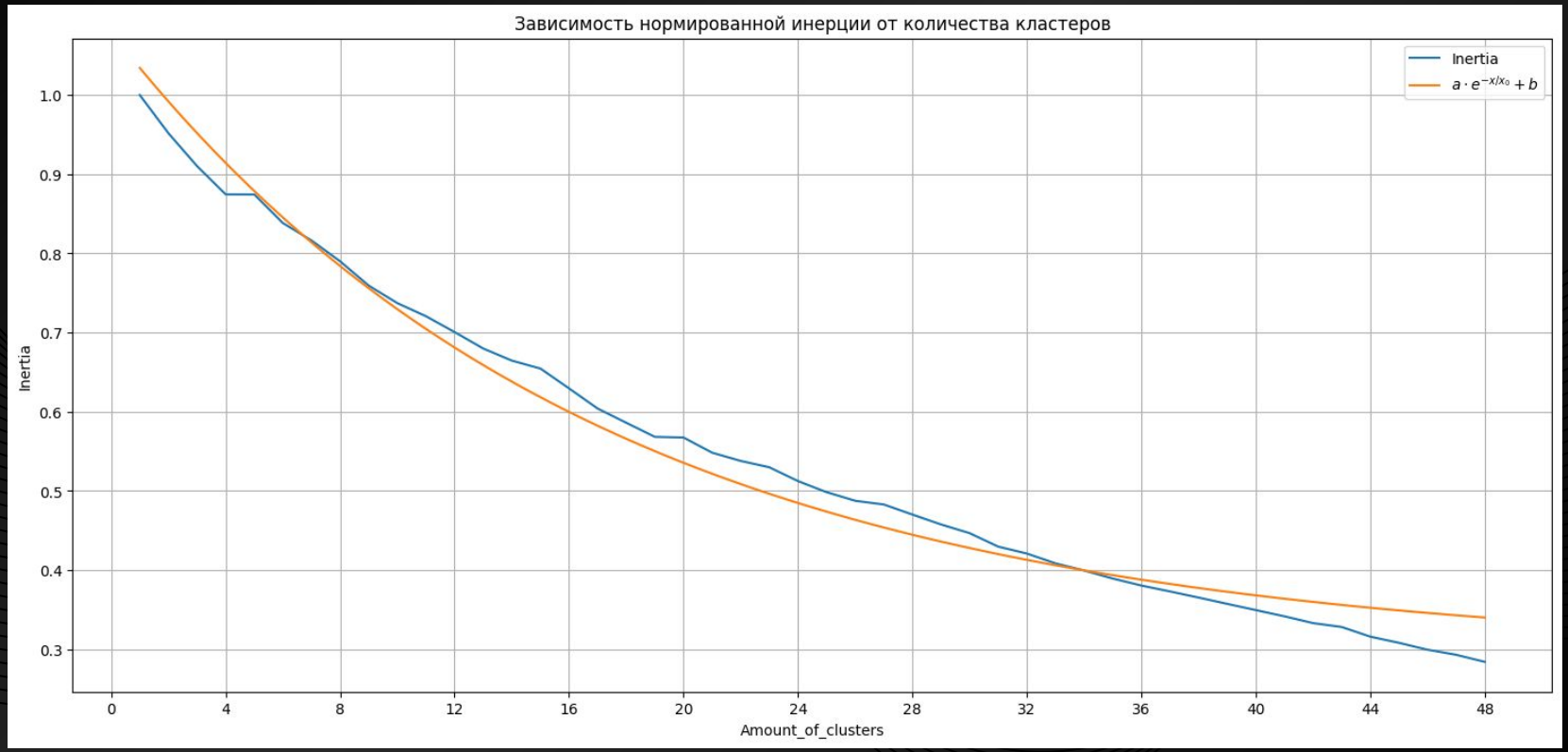
1) Черновое удаление стоп-слов и слов негативного окраса с помощью регулярных выражений
2) Лемматизация с использованием pymorphy2
3) Векторизация с использованием SBERT

Были попытки формализации слов и фраз с использованием LLM (API YandexGPT и другие), но доступ к ним платный и долгий.

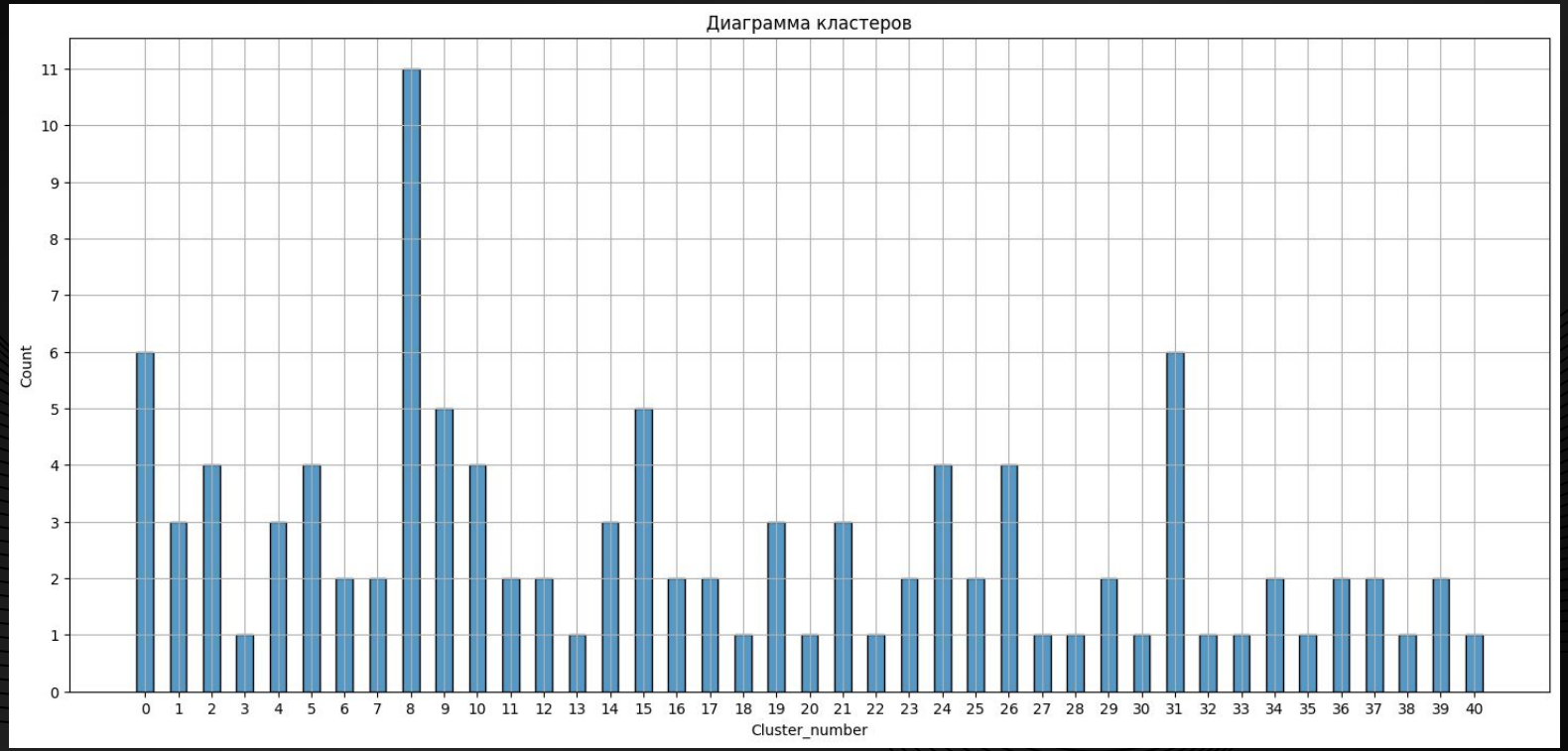
# Реализация модели машинного обучения

1) Кластеризация с использованием Kmeans и kmeans_plusplus
2) Интерполяция инерции гладкой экспоненциальной функцией
3) Подбор параметров интерполяционной функции методом градиентного спуска
4) Автоматический выбор оптимального количества кластеров с использованием физического (Дебаевского) формализма

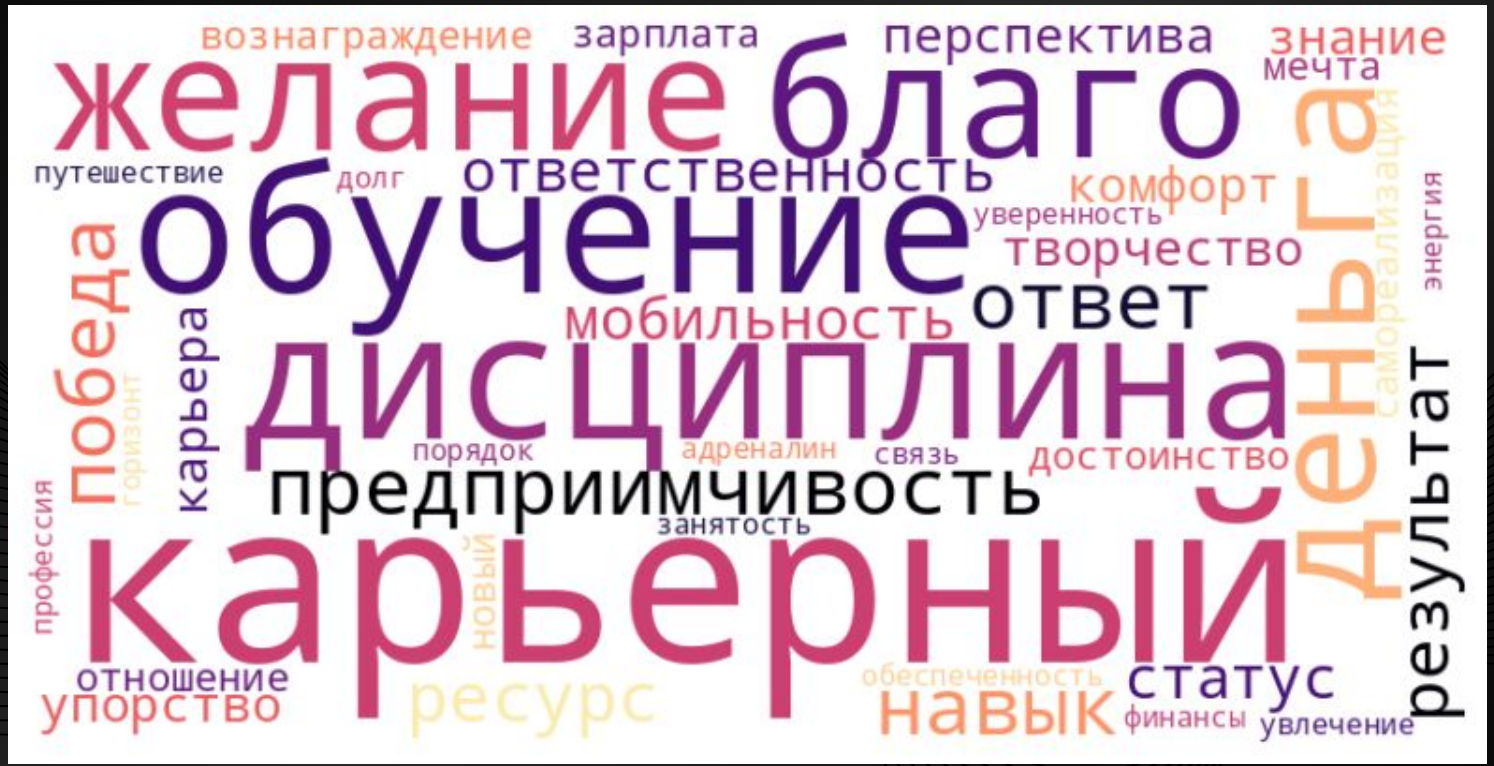
# Выбор количества кластеров



# Визуализация кластеров для 100



# Формирование облака слов



# Взаимодействие с программой

1) Начинаем диалог с телеграм ботом
2) Присылаем файл с входными данными в формате .csv, .xlsx или .json
3) На выходе получаем облако слов и json файл с посчитанными словами в кластерах
4) В зависимости от количества столбцов в .csv файле бот может вернуть несколько изображений / файлов - ответов на каждый вопрос

# Команда Voroge Jeskela

**Сулимов Александр**

Капитан, ML-разработчик

**Маньшин Тимур**

MLOps

**Усков Даниил**

Аналитик-разработчик

**Монастырный Максим**

ML-разработчик