

# “Nuclear IT Huck” МТС Линк

---

**РАЗРАБОТКА СИСТЕМЫ НА ОСНОВЕ ИИ,  
АНАЛИЗИРУЮЩЕЙ ПОЛЬЗОВАТЕЛЬСКИЕ  
ОТВЕТЫ И ВОЗВРАЩАЮЩЕЙ ПОНЯТНОЕ  
ОБЛАКО СЛОВ**

---

Дайте работу

## Цель

Разработка системы на основе ИИ, способной анализировать и кластеризовать пользовательские ответы, а также предоставлять понятное и наглядное облако слов, отображающее ключевые темы и понятия.

## Задачи

- Выбор модели для получения эмбедингов ответов
- Выбор алгоритма ИИ и его оптимизация
- Выбор алгоритма для обобщения мысли в каждой категории ответов
- Создание приложения для визуализации облака слов

# Получение данных

Google- форма

Имя Фамилия Отчество \*

Мой ответ

Наименование Организации \*

☐ МТС link

☐ МИФИ

☐ Другое

Что мотивирует вас работать больше? \*

Мой ответ

Спасибо за ответ!

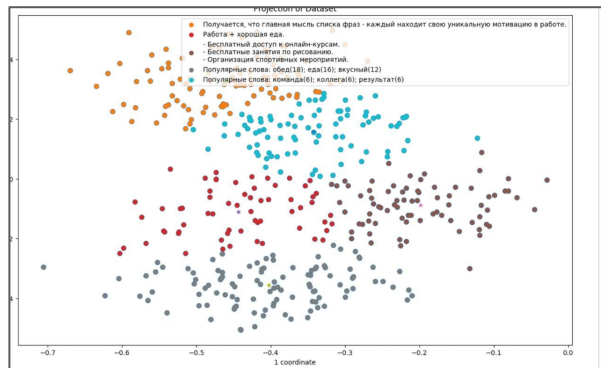
csv-файл

Отметка времени	Что мотивирует вас работать больше?	Имя Фамилия Отч	Наименование Орг
26.09.2024 0:06:33	Премия	Махнев Кирилл Алекс	МИФИ
26.09.2024 0:06:33	Хорошая зарплата	Иванов Иван Иванович	МИФИ
26.09.2024 0:06:33	Друзья на работе	Петров Петр Петрович	МТС link
26.09.2024 0:06:33	Возможность карьерного роста	Смирнова Анна Сергее	Другое
26.09.2024 0:06:33	Нормальная обстановка	Кузнецов Алексей Ник	МИФИ
26.09.2024 0:06:33	Да просто бабки нужны!	Васильева Светлана А	МТС link
26.09.2024 0:06:33	Получить опыт	Лебедев Сергей Викто	Другое
26.09.2024 0:06:33	Prosto zhal' ne uspet'	Сидорова Мария Влад	МИФИ
26.09.2024 0:06:33	Все ради команды	Григорьев Николай Юр	МТС link
26.09.2024 0:06:33	Крутые проекты!	Федорова Ольга Викто	Другое
26.09.2024 0:06:33	Заказов много, надо делать	Егорова Татьяна Дмит	МИФИ
26.09.2024 0:06:33	Работа ради удовольствия	Новиков Андрей Серге	МТС link
26.09.2024 0:06:33	Уважение коллег	Ковальчук Виктория А	Другое
26.09.2024 0:06:33	Хорошая команда	Зайцева Марина Анат	МИФИ
26.09.2024 0:06:33	Премии и бонусы	Семенов Игорь Валент	МТС link

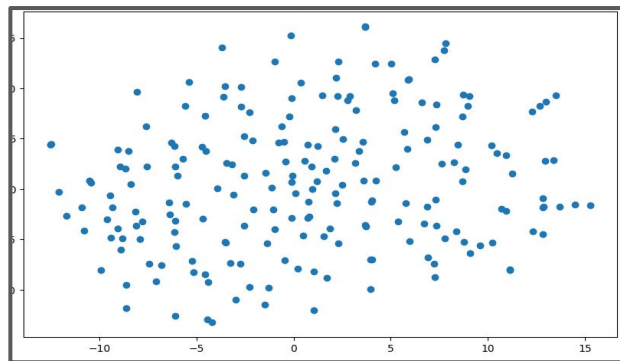
# Была поставлена задача кластеризации эмбеддингов ОТВЕТОВ



USE



RuBERT



Word2Vec



## Embeddings

Конвертируем текст в эмбединги, используя либо RuBERT, либо USE

## K-Means

Кластеризуем, используя K-Means. Вычисляем оптимальное количество кластеров с помощью силуэтного коэффициента

## DBSCAN

Кластеризуем, используя DBSCAN. Вычисляем оптимальное количество кластеров с помощью силуэтного коэффициента

## The Best

Сравнение метрик двух кластеризаторов и выбор наилучшего алгоритма

## GPT

При наличии сети отправляем запрос к GPT для суммаризации текста в кластерах

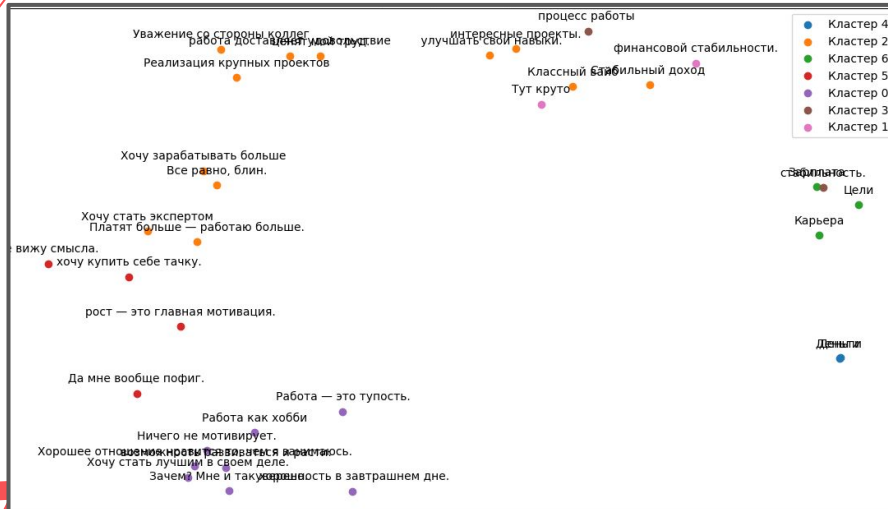
## Lemming + Bag Of Words

Если сети нет, то используем наивный подход, считая частоту слов в кластере

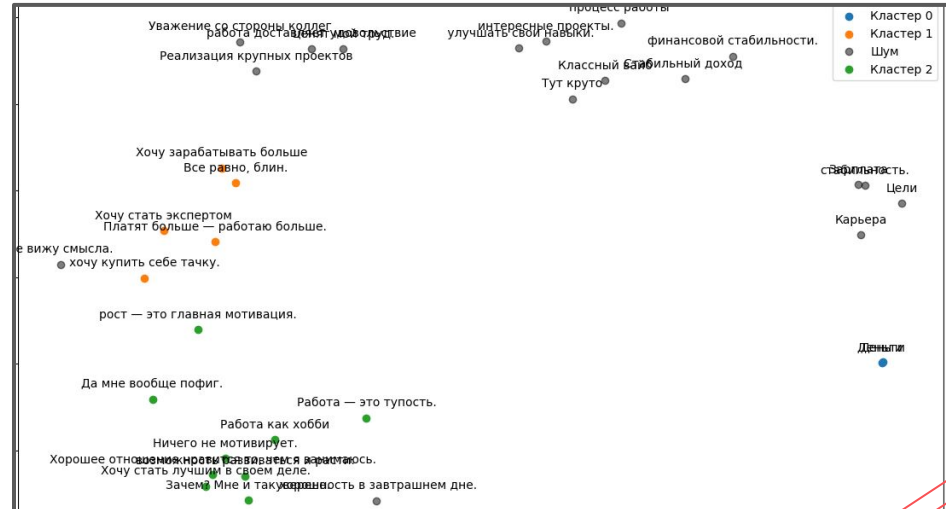


# Выбор модели кластеризации

## K-Means

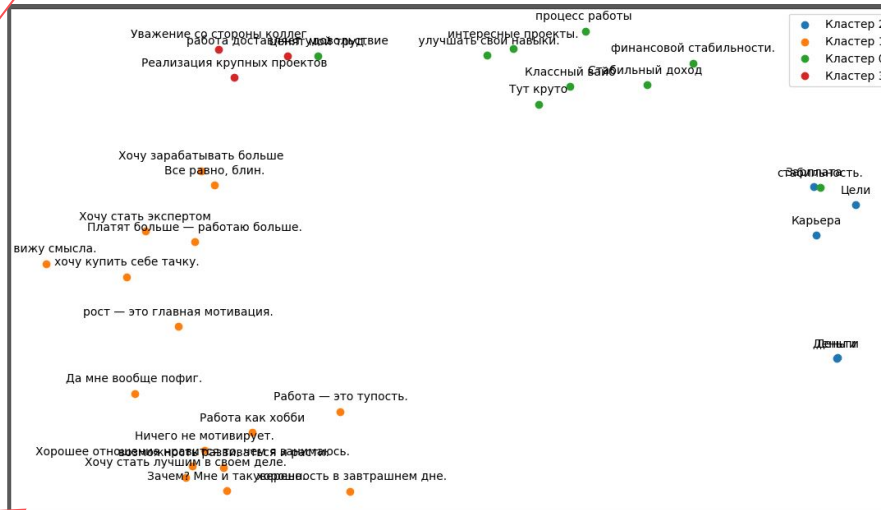


## DBSCAN



# Выбор модели кластеризации

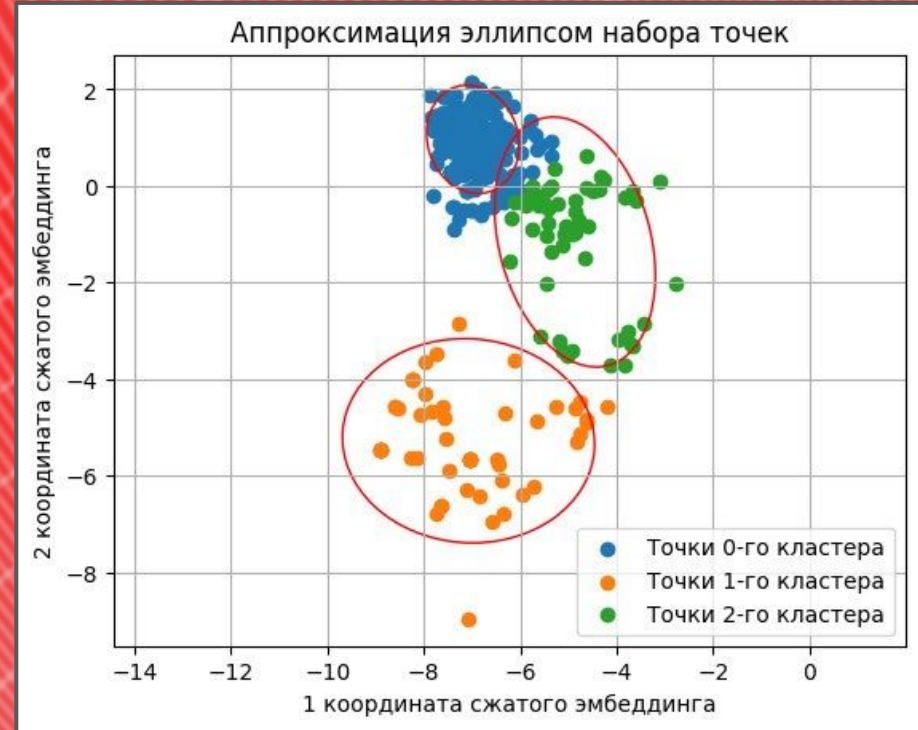
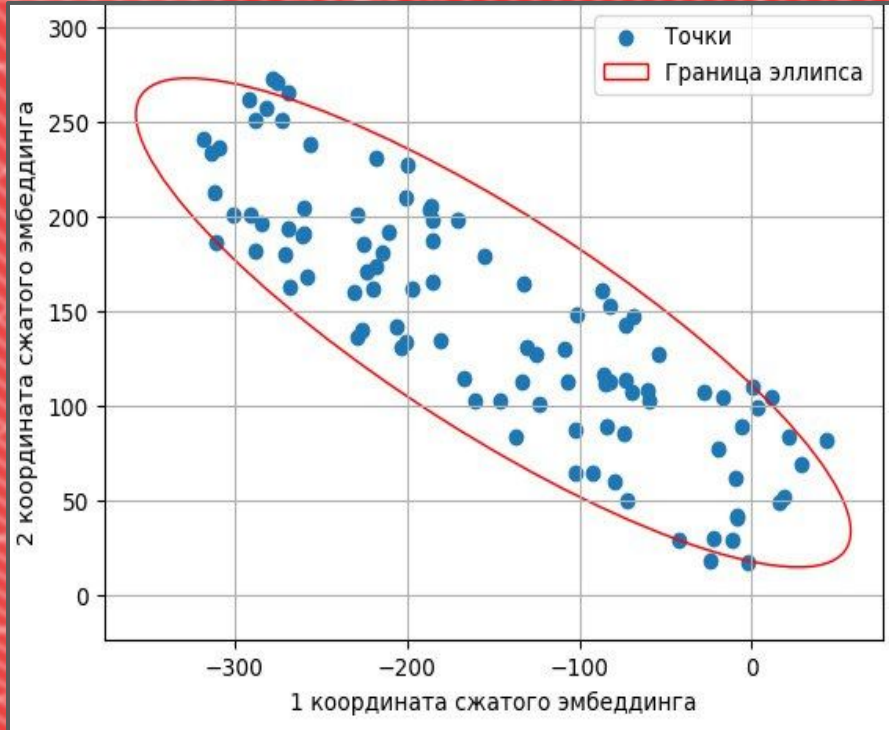
● Birch



● Diameter-clustering



# Отрисовка кластеров





# Суммаризация

GPT  
при наличии сети



Lemming + Bag of Words  
нет сети



# Приложение PyQT

### Аналитика ответов

Загружайте и анализируйте данные, чтобы получить информацию о мнениях, предпочтениях и мотивах

Данная программа использует искусственный интеллект для обработки текстовых данных. Конечным результатом работы алгоритма является построение на двумерной плоскости смысловых кластеров, содержащих близкие по смыслу выражения. Название кластеров отражает основную мысль выражений, попавших в этот кластер. Для более подробного ознакомления с алгоритмом работы рекомендуем ознакомиться с файлом README.md в репозитории

Руководство пользователя:  
 1) Выберите файл с данными, отвечающими вашему опросу в google forms  
 2) Выберите модель (RuBERT или USE) для образования эмбедингов  
 3) Нажмите кнопку "Запустить модель"

Если все прошло успешно, вы увидите, что на графике справа появилось изображение кластеров. При нажатии курсором на кластер, под графиком выведутся фразы, попавшие в этот кластер

### Выберите модель для создания эмбедингов

Загружена модель RuBERT

Загрузить RuBERT

Загрузить Universal sentence encoder

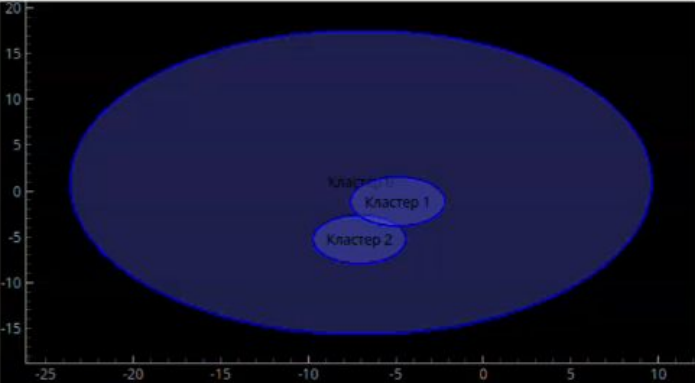
Выберите файл для извлечения данных

Выбранный файл: /home/drozdoymk/Projects/NLP\_hakaton/Hakaton-mts-link/notebooks/Answers.csv

Получить данные из Google forms

Получить данные из csv

Запустить модель



Кластер 1

Кластер 2

Да просто бабки нужны!  
 Крутые проекты!  
 Заказов много, надо делать  
 Да просто хрень какая-то  
 Нужен движ, иначе жопа  
 Лаве, тачки, отдых  
 Чтобы не заскучать  
 Бабки, чтобы поехать на море  
 Фичи и бабки — вот для чего!  
 За движуху, а не за хрень какую-то  
 Блин, не хочу скучать на работе!  
 Телки и нормальный кеш — важнее всего  
 Работа, чтобы не быть как все  
 Фарт, а не просто деньги  
 Хочу крутую атмосферу, иначе жопа!  
 Иногда просто надо делать, блин  
 За креатив, иначе просто жопа!  
 Проблемы решать — это отвал башки!  
 Все ради кайфа, бабки не главное



**СПАСИБО ЗА ВНИМАНИЕ!**