

Использование ИИ в продукте

хихитеam

Стародуб Анастасия
Феофилактова Юлия
Майорова Анна

Задача от МТС Линк на хакатоне
IT Nuclear Hackathon 2024

Проблемы

Сотни отзывов,
получаемые от
клиентов, обрабатывать
вручную трудоемко и
неэффективно

Быстрое реагирование
на обратную связь
клиентов критически
важно для поддержания
высокого уровня
сервиса

Необходимо извлекать
ценные инсайты из
массива отзывов для
улучшения продукта

Задачи

Предобработка текста

- включает очистку предложений от пунктуации и лишних символов, а также лемматизацию (приведение слов к их начальной форме) с использованием библиотеки `rumorphy2`.

Кластеризация текста

- предложения из CSV-файла преобразуются в эмбединги с использованием модели `Sentence Transformers`, а затем кластеризуются с помощью метода `KMeans`.

Визуализация результатов

- для каждого кластера создаются облака слов на основе наиболее часто встречающихся существительных, а также гистограммы частот этих слов.

Использованные библиотеки

```
import csv
import pymorphy2
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
import numpy as np
from sentence_transformers import SentenceTransformer
from collections import Counter
from wordcloud import WordCloud
```

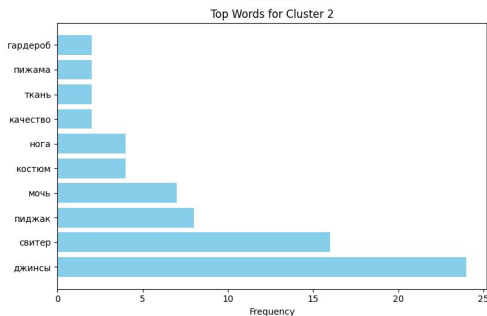
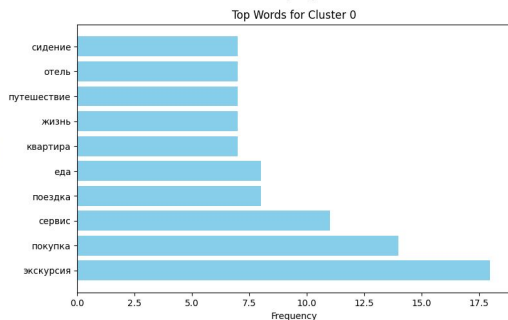
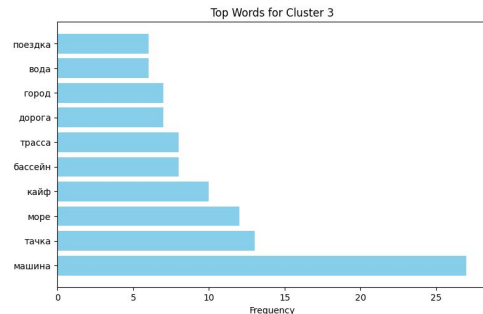
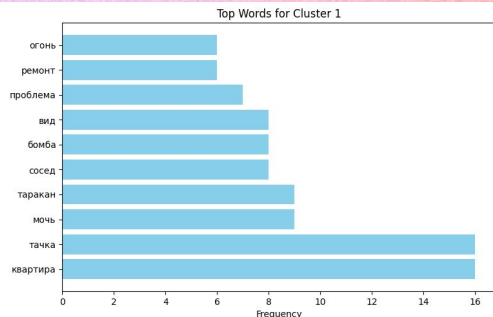
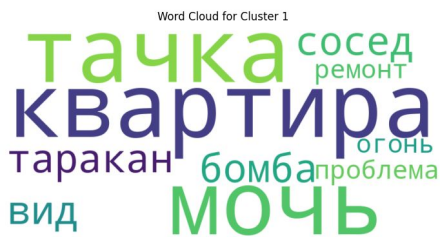
Использованная предобученная языковая модель

paraphrase-multilingual-MiniLM-L12-v2

Пример данных, для проверки работы программы

205	Пиджак супер, но ожидал немного другую текстуру.
18	Не всегда работает отопление, пришлось вызывать...
57	Кто строил этот дом? Все ужасно, даже не буду ...
6	Качественная отделка, но проблемы с коммуникац...
416	5. Просто бомба, рекомендую всем друзьям!
76	Каждый раз, когда прихожу домой, чувствую себя...
168	Почему они такие узкие? Я в шоке был!
169	Купила джинсы, но они сидят как мешок, нууу!
502	15. Купил и не знаю, как раньше жил без неё!

Визуализация



Итоги работы



Создана основа для сервиса по сбору и аналитике данных



Создана гибкая программа, подходящая для обработки данных из различных тематических опросов, которая основана на методе кластеризации



Создана структурированная и понятная для пользователя визуализация ключевых слов из ответов в опросе



Требуется доработка для получения полноценного сервиса

Вопросы