Разработка системы анализа и кластеризации отзывов пользователей

Выполнили:
Похачевский Всеволод
и
Пономаренко Алексей

Цель и задачи

Цель:

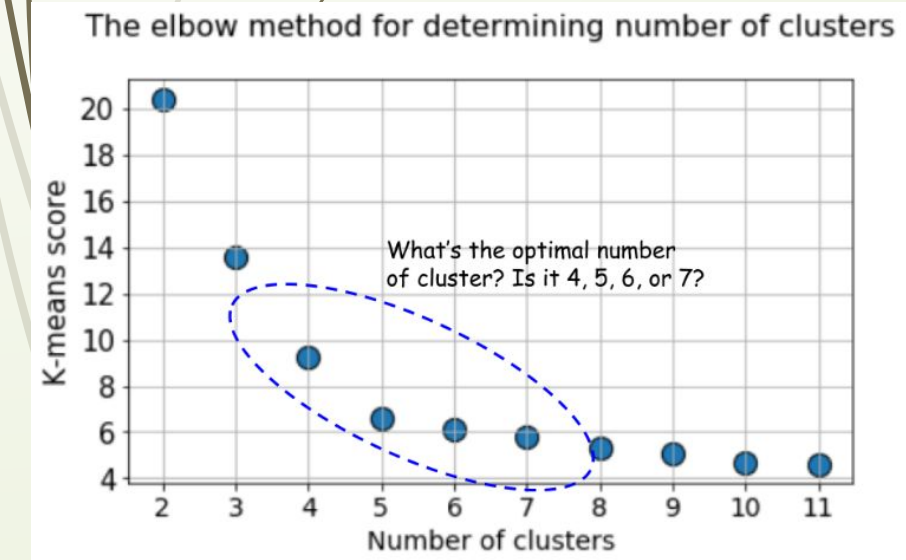
Создать модель кластеризации отзывов на отели

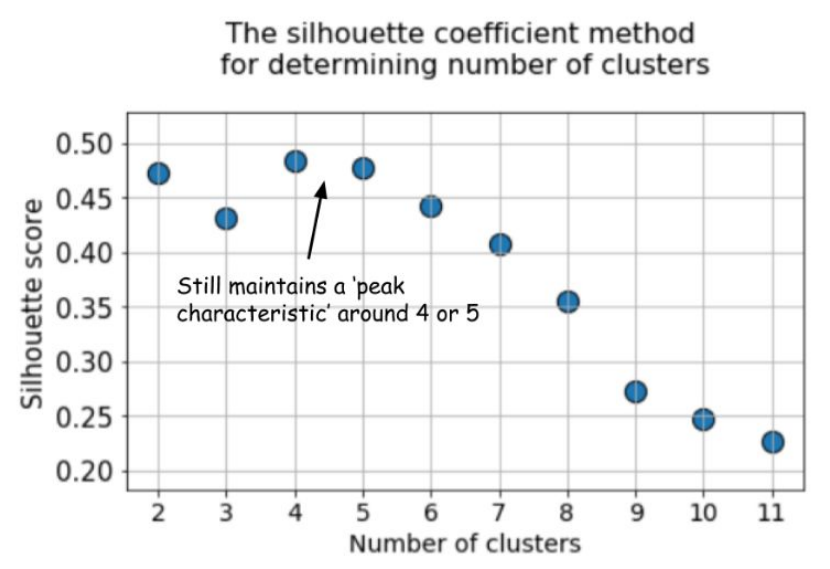
Задачи:

- Сбор и предобработка данных
- Grid-search подходов векторизации и кластеризации текстов
- Оценка качества кластеризации, интерпретация результатов
- Формирование датасета и обучение модели классификации

Инструменты и технологии

- Python, Jupyter
- Библиотеки: Scikit-learn, Matplotlib, Seaborn, Pandas, NLTK, SpaCy
- Модели: TF-IDF, BERT (Hugging Face Transformers)
- Метод «силуэта» (вместо классического метода «локтя») для оценки качества кластеризации





Проверенные пайплайны

- 1. TF-IDF + K-means
- 2. TF-IDF + Agglomerative Clustering
- 3. TF-IDF + DBSCAN
- 4. BERT + Agglomerative Clustering
- 5. BERT + K-means
- 6. BERT + DBSCAN

P.S. Для сравнения метрики альтернативных лучших моделей кластеризации

3 кластера (Bert + AgglomerativeClustering)

- Силуэтный коэффициент: 0.148
- Коэффициент WCSS: 9805.120107891566
- Коэффициент Дависа-Болдина: 2.9256600501151273
- Коэффициент Калински-Харабаса: 812.4585974655664

2 кластера (Bert + AgglomerativeClustering)

- Силуэтный коэффициент: 0.164
- Коэффициент WCSS: 9805.120107891566
- Коэффициент Дависа-Болдина: 2.2528058633859533
- Коэффициент Калински-Харабаса: 1260.6858206839922

4 кластера (TFIDF + KMeans)

- Силуэтный коэффициент: 0.028
- Коэффициент WCSS: 5172.948703535489
- Коэффициент Дависа-Болдина: 5.195330828109466
- Коэффициент Калински-Харабаса: 142.99585431263847

Выбор оптимального подхода

Оптимальный подход:

- BERT + Agglomerative Clustering
- Количество кластеров: 2
- Причины: Учет контекста
- метрика Silhouette

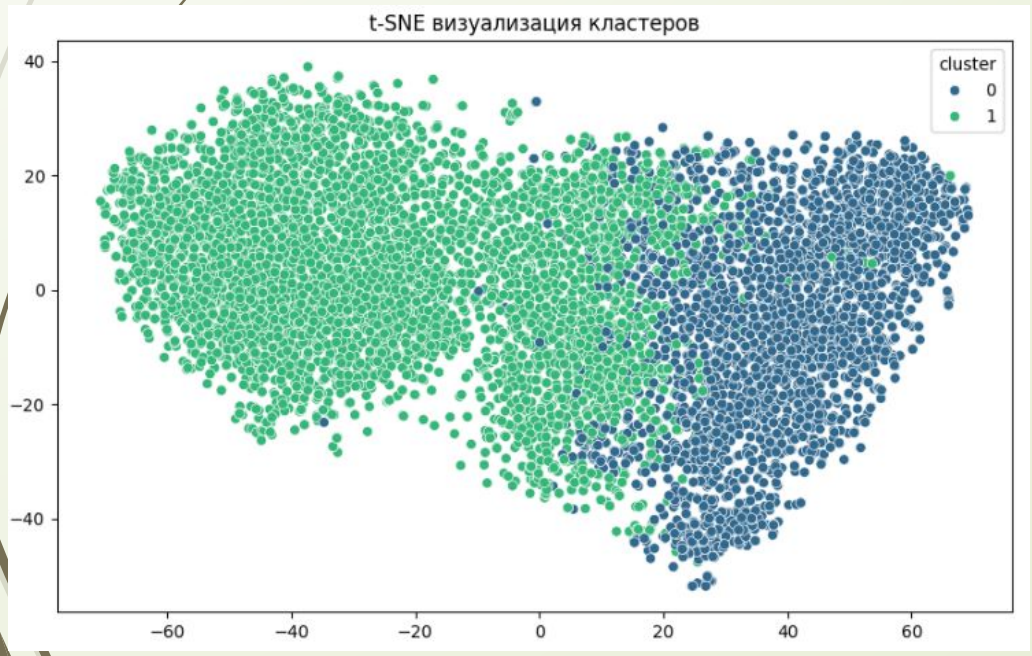
```
Силуэтный коэффициент: 0.198
Коэффициент WCSS: 8486.438759716468
Коэффициент Дависа-Болдина: 2.061500887538999
Коэффициент Калински-Харабаса: 1445.2920038183147
```

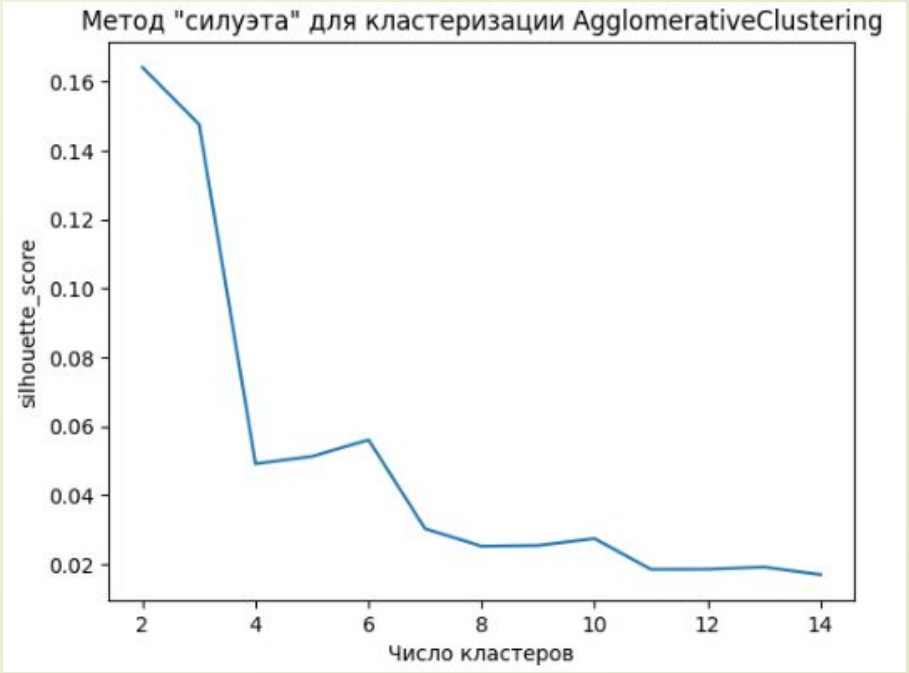
Результаты кластеризации

Описание кластеров на основе ключевых слов:

- Кластер 0: акцент на комфорте и удобствах отеля
- Кластер 1: общие впечатления от пребывания

На графиках t-SNE визуализация кластеров и метрика Silhouette (0.198)

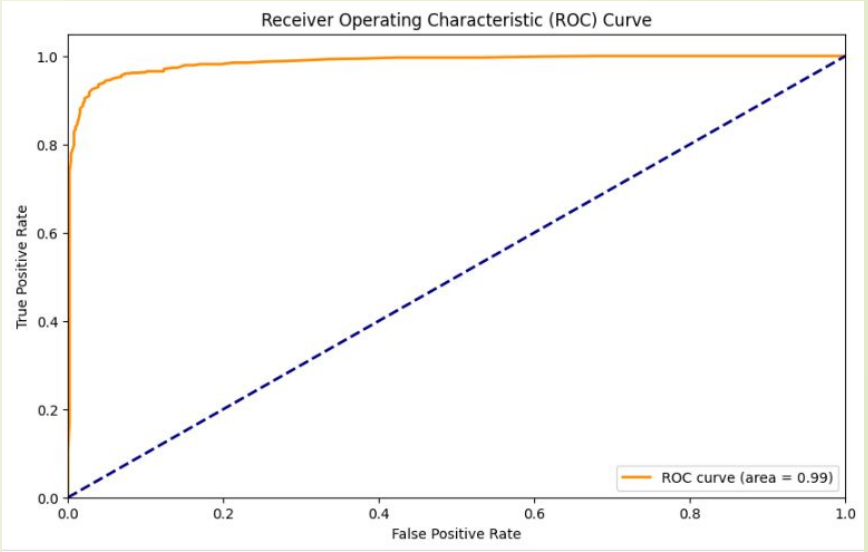


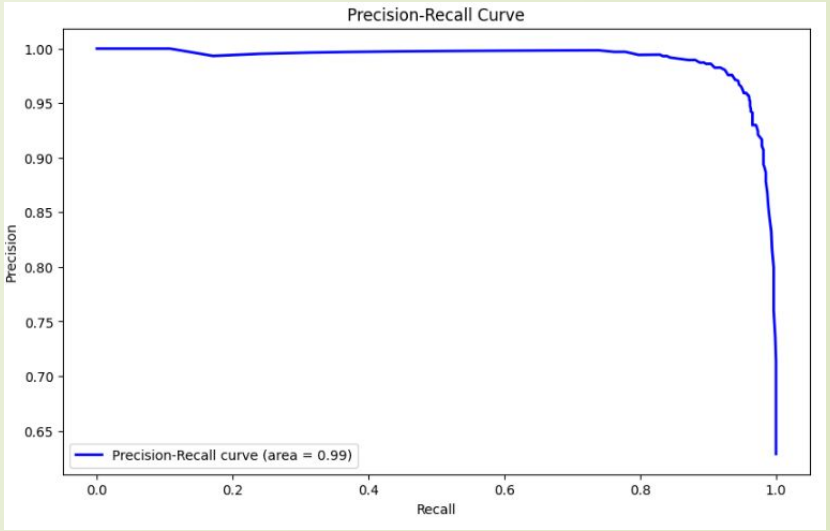


Результаты автоматической категоризации

- Модель: Random Forest
- Метрики:
  - F1 Score: 0.94

На графиках ROC-кривая и Precision-Recall кривая.





Спасибо за внимание!