Инструмент для анализа клиентских отзывов

Задание на второй этап отбора

СИРИУС.ИИ

Сбор данных

Для этого задания мы решили взять данные с сайта https://otzovik.com/.

С помощью веб-скрейпинга мы проходимся по каждой из 422 страниц с отзывами, на каждой странице открываем полностью каждый отзыв и с этой страницы берём информацию: текст отзыва (вместе с темой, положительными и отрицательными чертами), дату написания отзыва и оценку пользователя. Эту информацию сортируем по спискам.

СИРИУС.ИИ

Извлечение признаков

Список стоп-слов взят из библиотеки nltk и дополнен некоторыми словами вручную. Из текстов отзывов убираются все стоп-слова, цифры и лишние символы, текст приводится в нижней регистр.

После обработки, тексты переводятся в векторное представление используя Word2Vec.

СИРИУС.ИИ

Анализ и интерпретация отзывов

С помощью Kmeans тексты распределяются по кластерам. Далее, для определения тем кластеров используется TfidfVectorizer. С его помощью мы определяем топ-10 самых частых слов для каждого кластера, по которым можно определить тему.

Находим такую статистику как количество положительных, отрицательных и нейтральных отзывов.

Формируем круговую диаграмму с помощью matplotlib, что бы увидеть отзывов из каких кластеров больше всего. Далее для каждого кластера выводим топ-10 слов, что бы определить тему.

СИРИУС.ИИ

Анализ и интерпретация отзывов

Формируем график, на котором видно сколько положительных, отрицательных и нейтральных отзывов было за каждый проанализированный месяц. Это поможет выявить аномалии и узнать качество нововведений. Пример работы ниже:

Результат

К сожалению, веб-скрейпинг без использования API перегружает сайт, и тот начинает блокировать запросы с айпи адреса пользователя. Так что пока что удалось только собрать статистику по количеству отзывов:

Из 8424 проанализированных отзывов:

Положительных отзывов 6422, это 76,2% от общего количества.

Отрицательных отзывов 1806, это 21,5% от общего количества.

Нейтральных отзывов 196, это 2,3% от общего количества.

СИРИУС.ИИ

Результат

Однако, в финальном решении поставленной задачи, мы будем использовать API, что позволит обойти эти ограничения, и успешно проанализировать отзывы.

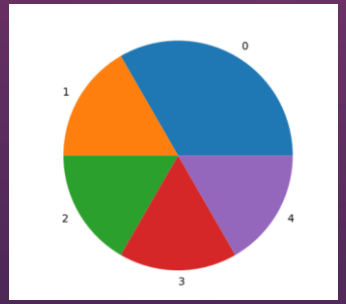
Примеры

Примеры работы диаграммы и кластеризации если взять несколько случайных отзывов (на маленьких наборах данных могут быть неточности. Примеры приведены для демонстрации функционирования кода в целом):

```
response_texts = [
    'Отзыв: Тинькофф банк - Отличный банк Достоинства: Отличный Недостатки: Нет Вообще сначала не хотела даже дебитовую карту заводить так как очень уж плохого слышала о ,
    'Отзыв: Тинькофф банк - Просто волшебно! Достоинства: Скорость Выгода Простота Недостатки: Нету ! Произошло ДТП, была оформлена страховка осаго Тинькофф(номер 0318788
    'Отзыв: Тинькофф банк - Хороший банк Достоинства: Удобное обслуживание Красивый интерфейс Легко переводить на разные нужды Недостатки: Плата за уведомления об операци
    'Отзыв: Тинькофф банк - Удобный мобильный банк , хороший кэшбек Достоинства: Удобный мобильный банк Недостатки: Мало офлайн офисов Хороший кэшбек, удобный мобильный б
    'Отзыв: Тинькофф банк - Бесконечные пустые отписки, нарушение законных прав вкладчиков, при требовании предоставить установленные законодательством документы, отключа
    'Отзыв: Тинькофф банк - Возьмите кредит до 30 млн., только мы вам их никогда не одобрим. Как обманывает Тинькофф Достоинства: Когда то была очень клиентоориентированн
]
response_values = ['5', '5', '4', '4', '1', '1']
response_dates = ['30 мар 2024', '29 мар 2024', '26 мар 2024', '22 мар 2024', '21 мар 2024', '15 мар 2024']
```

```
K-means Clusters: [4 1 0 2 0 3]
```

```
Кластер 0, самые частые слова: ['маленький', 'произойти', 'потрясать', 'страховка', 'помочь', 'скорректировать', 'уровень', 'дтп', 'выгода', 'оперативно'].
Кластер 1, самые частые слова: ['легко', 'уведомление', 'обслуживание', 'вклад', 'нужда', 'красивый', 'банковский', 'который', 'интерфейс', 'договор'].
Кластер 2, самые частые слова: ['доставка', 'быстрый', 'звонок', 'отправить', 'реально', 'удобный', 'хороший', 'мобильный', 'кэшбек', 'карта'].
Кластер 3, самые частые слова: ['покупка', 'снова', 'писать', 'очень', 'тыс', 'сумма', 'одобрить', 'клиент', 'млн', 'кредит'].
Кластер 4, самые частые слова: ['ужасно', 'мочь', 'вообще', 'пользоваться', 'период', 'год', 'очень', 'отличный', 'карта', 'кредитный'].
```

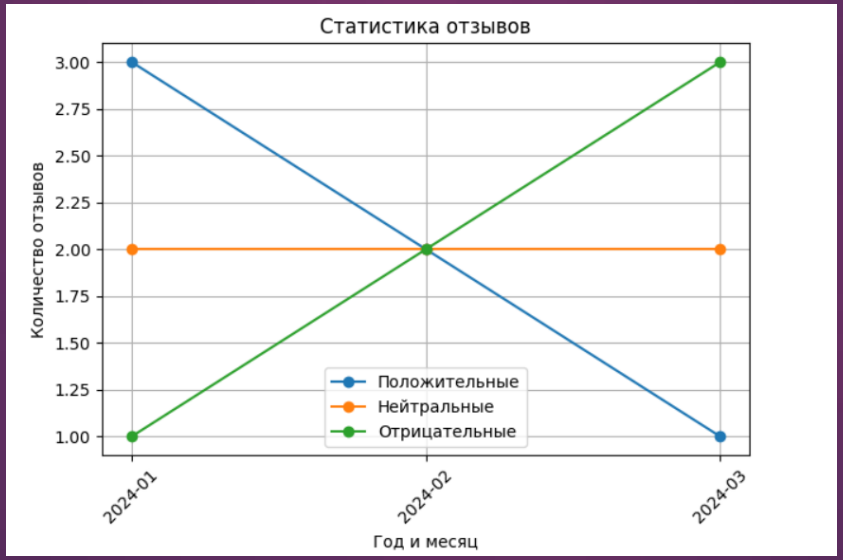


СИРИУС.ИИ

Примеры

Пример визуализации статистики отзывов:

```
response_dates = ['28 янв 2024', '29 янв 2024', '30 янв 2024', '28 янв 2024', '29 янв 2024', '30 янв 2024',
                  "16 фев 2024", "17 фев 2024", "18 фев 2024", "1 фев 2024", "2 фев 2024", "20 фев 2024",
                  "9 мар 2024", "10 мар 2024", "11 мар 2024", "4 мар 2024", "10 мар 2024", "29 мар 2024"]
response_values = ['5', '5', '4', '3', '3', '1', '4', '4', '3', '3', '1', '2', '5', '3', '3', '1', '1',' 2']
```



Ссылка

Ссылка на репозиторий github с решением:

https://github.com/egorshalin08/siriusii_tinkoff_Shalin_EV

СИРИУС.ИИ