Дополнительное задание

анализ отзывов из открытых источников

Цель

Выявить общие темы и тенденции в отзывах о компании Тинькофф

Ход работы

1. Сбор данных
2. Обработка
3. Анализ

Сбор данных

Разумеется, отзывов на просторе интернета очень много. Наша команда работает над сервисом парсинга в режиме реального времени актуальных отзывов из наиболее релевантных источников по ключевому слову.

Пока в данной задаче мы будем собирать отзывы с платформы Banki.ru

Подготовка к сбору данных

Моменты, которые были выявлены во время изучения платформы:

1) На тематической странице компании содержатся не сами отзывы, а лишь линки на них
2) Сами отзывы хранятся отдельно

Поэтому было решено спарсить все линки с тематической страницы, а затем по ним собирать сами отзывы

Сбор данных

Используя инструменты языка Python: requests, BeautifulSoup4, pandas, мы собрали 4000 отзывов за 2 часа парсинга(весь код парсеров в репозитории)

ПРИМЕР полученных данных

	Unnamed: 0	header	created_date	rating	message
0	0	Поддержка	01.03.2024 22:39	5	Хочу выразить огромную благодарность сотрудник...
1	1	Персональный Менеджер	26.02.2024 14:48	5	Добрый день.
2	2	Самый лучший банк	09.03.2024 10:39	5	Очень редко пишу отзывы, но сейчас то самое вр...
3	3	Обращение по отправке годовой декларации в нал...	05.03.2024 18:15	5	Общалась с Ф-вой Натальей несколько дней. Снач...
4	4	Спасибо большое сотрудникам Тинькофф поддержки...	26.02.2024 16:01	5	26,02 обратились в Тиньков по спорному моменту...

Обработка данных

Хороший современный pipeline NLP предобработки включает: удаление мусора из текстов(стоп-слов, знаков препинания и т.д). Также вдобавок к этому, мы решили заменить эмоджи словесным описанием, так как они описывают эмоциональную окраску, особенно в отзывах. Так как количество отзывов небольшое, было принято решение лемматизировать, нежели использовать стемминг, потому что лемматизация более точный метод, хотя обладает большей вычислительной сложностью. На этом базовая предобработка завершена. Использованы следующие инструменты: NLTK, pymorphy2, emoji - библиотеки Python.

До/после предобработки

ДО: `Очень оперативно и с результатом решили мои вопросы в чат поддержке: недавно у меня сняли деньги за непонятную подписку🙂`

ПОСЛЕ: `очень оперативно результат решить мой вопрос чат поддержка недавно снять деньга непонятный подписка улыбается`

Векторизация

Мы использовали сразу два базовых способа векторизации: CountVectorize(мешок слов) и TD-IDF, min_df = 2|3, чтобы избавиться от аномалий

Анализ

1) Первый способ для анализа - Латентное размещение Дирихле
2) Второй способ - изучение значений TF-IDF

LDA(Latent Dirichlet allocation)

Разделили на 10 тем, из которых можем выделить общие темы отзывов, например: topic 1 - тема - банк, деньги; topic 0 - кэшбэк тинькофф и т.д.

topic 0	topic 1	topic 2	topic 3	topic 4
кэшбэк	банк	банк	деньга	оператор
покупка	счёт	вопрос	карта	поддержка
начислить	тинькофф	очень	вернуть	курс
категория	перевод	тинькофф	день	который
кэшбек	комиссия	решить	банк	это
повышенный	деньга	проблема	подписка	приложение
вопрос	банкомат	быстро	поддержка	бек
чат	день	поддержка	средство	кеш
начисление	средство	спасибо	тинькофф	рубль
кешбек	поддержка	хороший	списать	валюта

topic 5	topic 6	topic 7	topic 8	topic 9
банк	вопрос	банк	банк	тинькофф
карта	очень	сотрудник	сотрудник	хотеть
это	помочь	мой	тинькофф	банк
тинькофф	быстро	счёт	благодарность	всё
мой	банк	вопрос	хотеть	работа
год	менеджер	который	вопрос	свой
кредитный	спасибо	документ	спасибо	поделиться
день	который	линия	выразить	здравствуйте
который	работа	горячий	огромный	декларация
деньга	помощь	горячий линия	работа	ип

TF-IDF анализ

Токены с низкими значениями **TF-IDF** встречаются в отзывах часто,

Исходя из этого, можно сделать вывод, что пользователи больше всего пишут в своих отзывах на темы договоров, документов, оферт и т.п.

```
Признаки с наименьшими td-idf:
['оферта должный' 'инициатива иной' 'должный признаваться' 'анкета банк'
 'установить договор' 'ухудшить' 'акцептант который' 'документ выразить'
 'акцептант правопорядок' 'акцептант такой' 'договор условие'
 'отписка скрипт' 'договор оферент' 'договор норма' 'такой бремя'
 'иной грубо' 'правопорядок допускать' 'реагирование самый'
 'признаваться ничтожный' 'интерес адресат' 'ухудшить положение' 'ст гк'
 'исключительно собственный' 'кодекс российский'
 'обращение регистрировать' 'неудобный положение' 'давать глупый'
 'грубо посягать' 'положение вынудить' 'положение гражданский']
Признаки с наибольшими td-idf:
['заблокировать карта' 'описание ситуация' 'клуб' 'читать'
 'уважаемый тинькофф' 'обратиться поддержка' 'банк высота' 'опыт тинькофф'
 'премиум клиент' 'обращение продукт' 'отличный карта' 'день добрый'
 'заказывать' 'моментальный' 'душа' 'ноябрь' 'банк' 'здравствуйте'
 'февраль' 'иван' 'добрый' 'dmitry' 'год' 'отзыв' 'претензия'
 'приветствовать' 'оценка' 'благодарить' 'сотрудник' 'nan']
```

Вывод

Нам удалось с помощью общедоступных инструментов выделить наиболее информативные слова и словосочетания в отзывах клиентов, и что более ценно, без «ручного» прочтения каждого отзыва в отдельности выявить из всего массива информации наиболее популярные темы, которые волнуют пользователей.

Анализ отзывов по мере TF-IDF может помочь узнать, какие слова и выражения преобладают в клиентских отзывах. И если там преобладают такие выражения, как «проблема» или «ужасно», стоит об этом знать.