Эксперименты по выявлению ключевых аспектов из текстовых отзывов

на основе отзывов с сервиса Яндекс.Карты

посмотрим на исходные данные

	0	1	2	3	4
0	address=Екатеринбург, ул. Московская / ул. Вол...	name_ru=Московский квартал	rating=3.	rubrics=Жилой комплекс	text=Московский квартал 2.\nШумно : летом по н...
1	address=Московская область, Электросталь, прос...	name_ru=Продукты Ермолино	rating=5.	rubrics=Магазин продуктов;Продукты глубокой за...	text=Замечательная сеть магазинов в общем, хор...
2	address=Краснодар, Прикубанский внутригородско...	name_ru=LimeFit	rating=1.	rubrics=Фитнес-клуб	text=Не знаю смутят ли кого-то данные правила,...

В этом проекте мы будем работать с 10 531 отзывом о кофейнях.

Источник данных: https://github.com/yandex/geo-reviews-dataset-2023/blob/master/README.md

сентимент-анализ с помощью RuBERT for Sentiment Analysis

```
def predict(text):
  inputs = tokenizer(text, max_length=512, padding=True, truncation=True, return_tensors='pt').to(device)
  outputs = sentiment_model(**inputs)
  predicted = torch.nn.functional.softmax(outputs.logits, dim=1)
  if (predicted > 0.8).any():  # отбираем отзывы, в которых нейросеть уверена на более, чем 80%
    predicted_label = torch.argmax(predicted, dim=1).cpu().numpy()
  else:
    predicted_label = np.array([0])  # остальные относим к нейтральным
  return predicted_label
```

В результате мы получили 8931 позитивный отзыв, 966 негативных.

частотный анализ и извлечение ключевых n-грамм

1. Нормализация данных:
   a. лемматизация
   b. кастомный список стоп-слов
   c. список общих фраз
2. CountVectorizer
3. TF-IDF
4. RAKE
5. KeyBERT + Counter

ключевые n-граммы позитивных отзывов

	CV биграммы	частота	CV триграммы	частота.1	TF-IDF биграммы	показатель	RAKE n-граммы	KeyBERT триграммы	частота.2
0	вкусный кофе	1370	вкусный кофе приятный	77	вкусный кофе	143.396732	уютный стильный кофейня	вкусный кофе	939
1	кофе вкусный	419	вкусный кофе десерт	67	кофе вкусный	55.358189	приемлемый цена рекомендовать	кофе вкусный	372
2	вежливый персонал	367	большой выбор напиток	52	вежливый персонал	54.923807	поесть обслуживание высота	приятный атмосфера	307
3	большой выбор	334	место вкусный кофе	51	приветливый персонал	48.839478	пекарня выпечка советовать	хороший кофе	206
4	приветливый персонал	327	персонал вкусный кофе	43	приятный атмосфера	48.323662	отличный уютный кофейня	уютный атмосфера	178
5	приятный атмосфера	323	кофе вкусный десерт	39	вкусный еда	43.714007	отличный напиток десерт	отличный кофе	163
6	вкусный еда	312	вкусный кофе вкусный	36	большой выбор	43.384899	обслуживание кухня высота	кофе приятный	139
7	хороший кофе	255	кофе приветливый персонал	36	хороший кофе	37.201893	добродушный интересный выбор	уютный кафе	129
8	уютный место	251	соотношение цена качество	35	уютный место	35.612992	потрясать кофе процветание	кофе десерт	125
9	персонал вежливый	232	большой выбор десерт	33	персонал вежливый	33.327539	отличный пирожное кофе	кофе хороший	123
10	вкусный десерт	216	вкусный кофе приветливый	33	отличный кофе	32.591938	скидка приятный фишка	вкусный десерт	118
11	отличный кофе	187	вкусный кофе хороший	32	вкусный десерт	31.752239	отличный атмосферный место	уютный кофейня	103
12	уютный атмосфера	185	кофейня вкусный кофе	32	приятный персонал	31.469157	атмосфера официант молодец	атмосферный место	102
13	приятный персонал	173	обслуживание высокий уровень	32	кофе приятный	27.991520	доступный вкусный меню	кофе отличный	99
14	персонал приветливый	166	кофе большой выбор	31	уютный атмосфера	27.821394	вкусный фо бо	приятный персонал	93

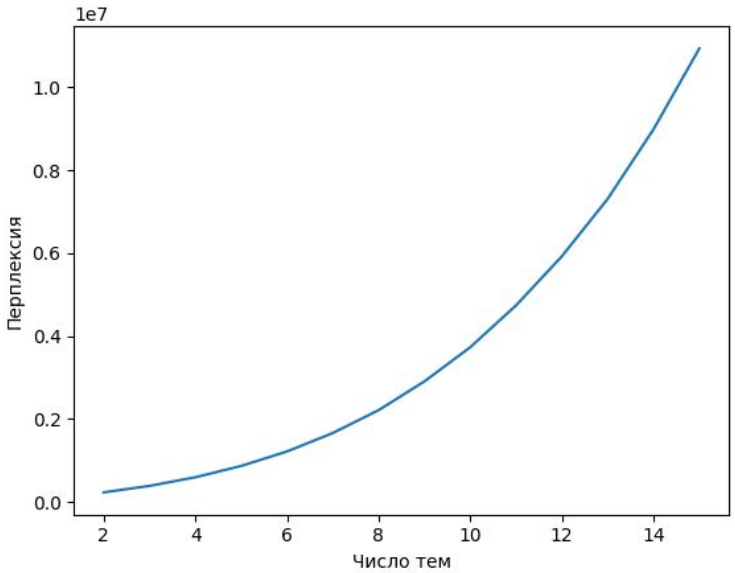
ключевые n-граммы негативных отзывов

	CV триграммы	частота	TF-IDF триграммы	показатель	RAKE n-граммы	KeyBERT триграммы	частота.1
0	не обращать внимание	7	вкусный кофе пироженое	1.000000	вкусный кофе пироженое	официант не подходить	5.0
1	кофе не вкусный	6	кофе не вкусный	0.979330	душевный кофе	официант не знать	3.0
2	не первый свежесть	6	не первый свежесть	0.857663	выливать отвратительный	кофе не вкусный	3.0
3	не начинать готовить	5	не понравиться не	0.772993	восхитительный сотрудник	приносить холодный кофе	2.0
4	не понравиться не	5	десерт не свежий	0.735327	вернуть скоун	кофе пирожное не	2.0
5	официант не знать	5	вкусно душно проблема	0.707107	средненький	вкусный кофе барист	2.0
6	официант не подходить	5	душно проблема кондиционер	0.707107	любить	не обращать внимание	2.0
7	принимать заказ не	5	зашкаливать не город	0.707107	NaN	официантка не предупреждать	2.0
8	стол не убирать	5	не уютно прохладно	0.707107	NaN	руководство обращать внимание	2.0
9	заказ ждать долго	4	отличный чайная тхегванинь	0.707107	NaN	цена не соответствовать	2.0
10	не испортить настроение	4	уютно прохладно временно	0.707107	NaN	кофе не понравиться	2.0
11	не официант не	4	цена зашкаливать не	0.707107	NaN	испортить праздник не	2.0
12	цена не соответствовать	4	чайная тхегванинь недорогой	0.707107	NaN	минута выходить официантка	2.0
13	внимание не обращать	3	официант не подходить	0.704588	NaN	чай не завариваться	2.0
14	грязный стол не	3	не начинать готовить	0.678953	NaN	кофе холодный стол	2.0

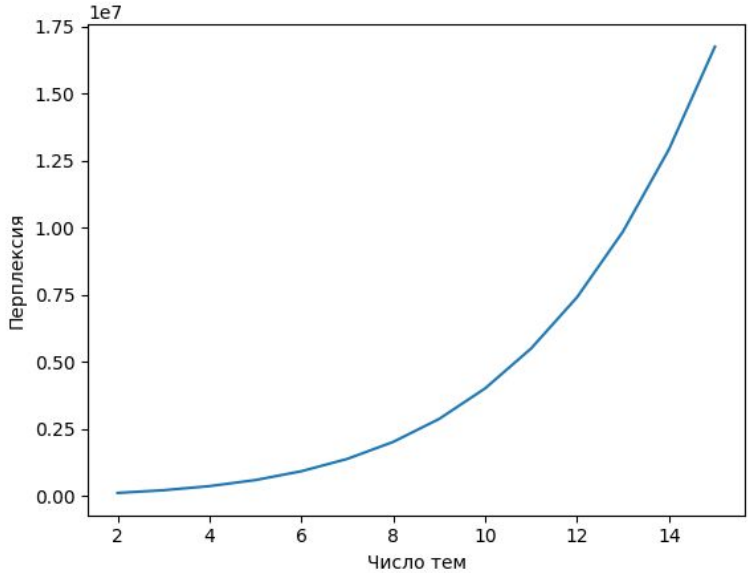
тематическое моделирование с помощью латентного размещения Дерихле (LDA)

графики зависимости перплексии от числа топиков:

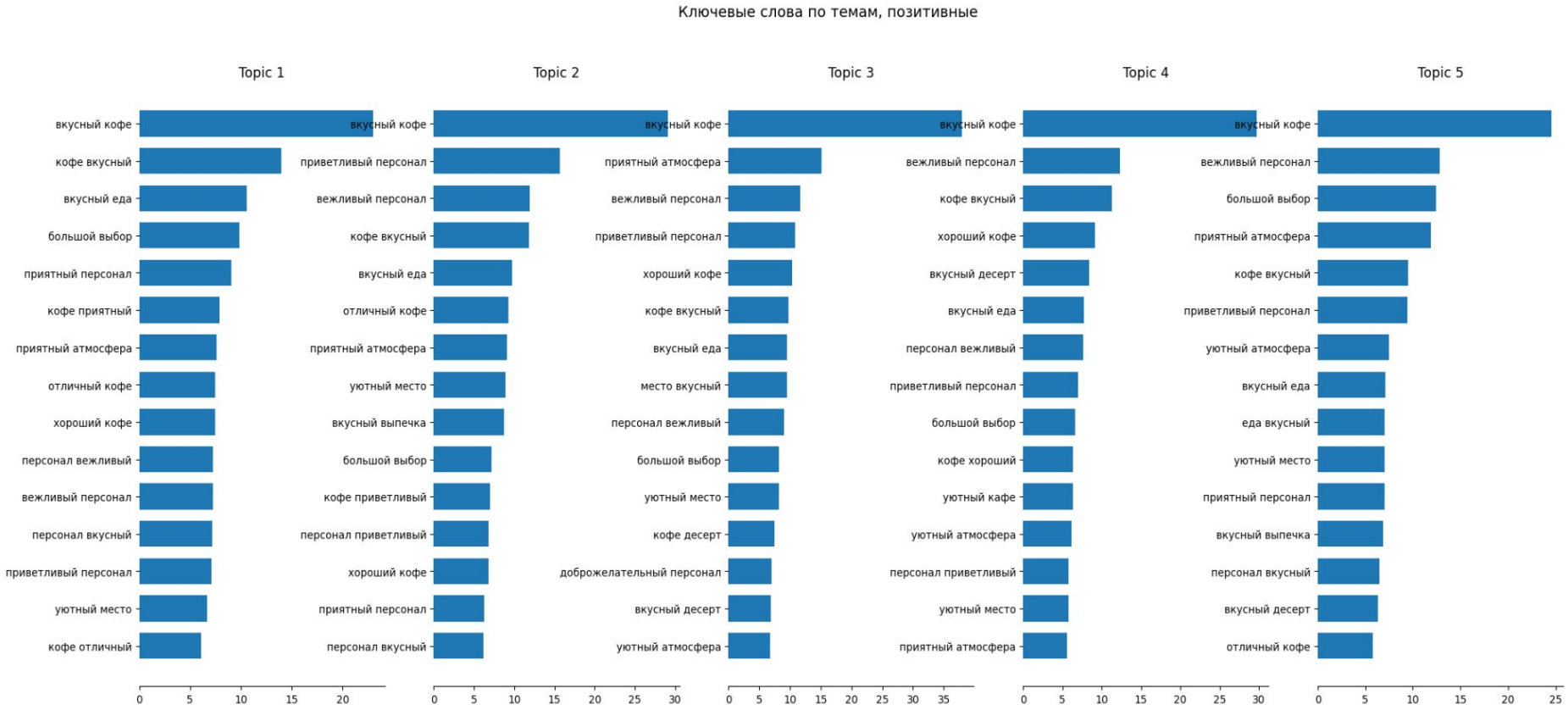
для позитивных отзывов:



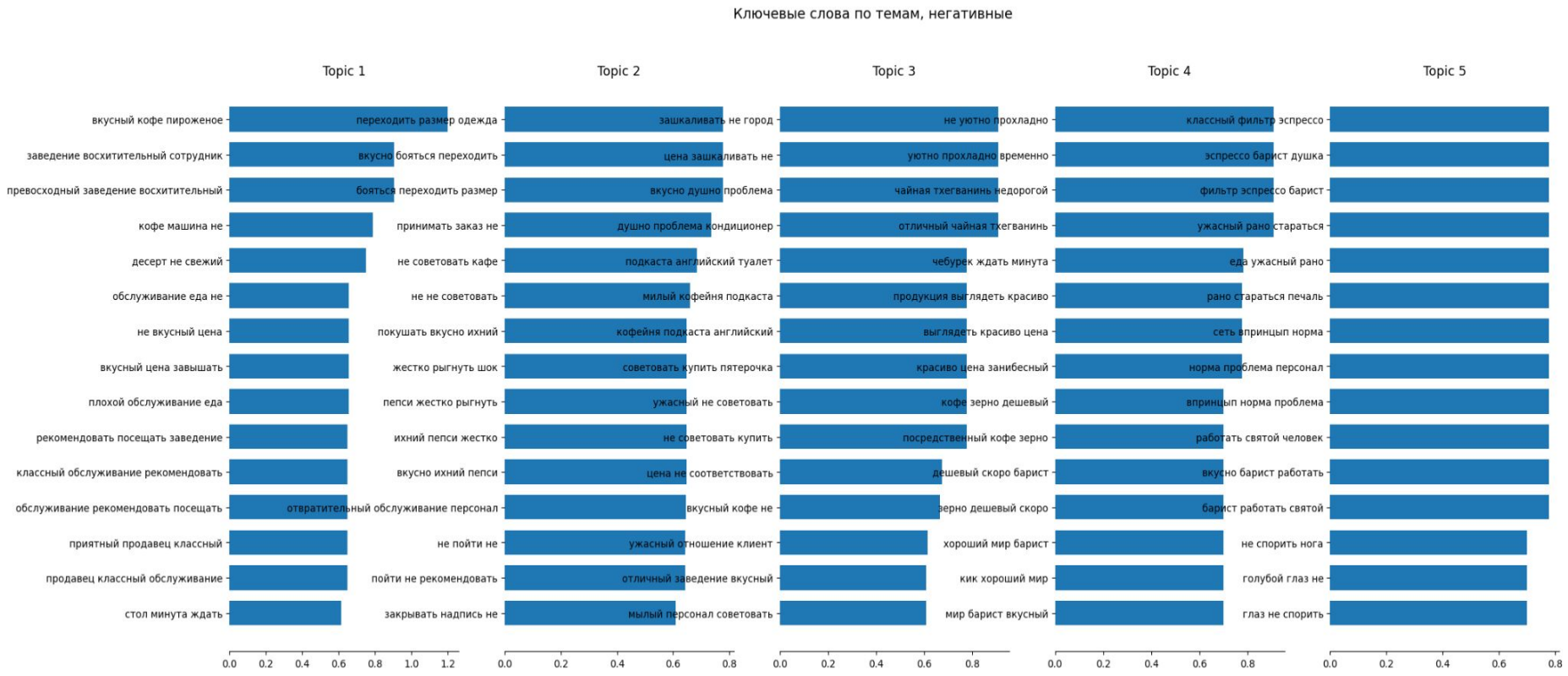
для негативных отзывов:



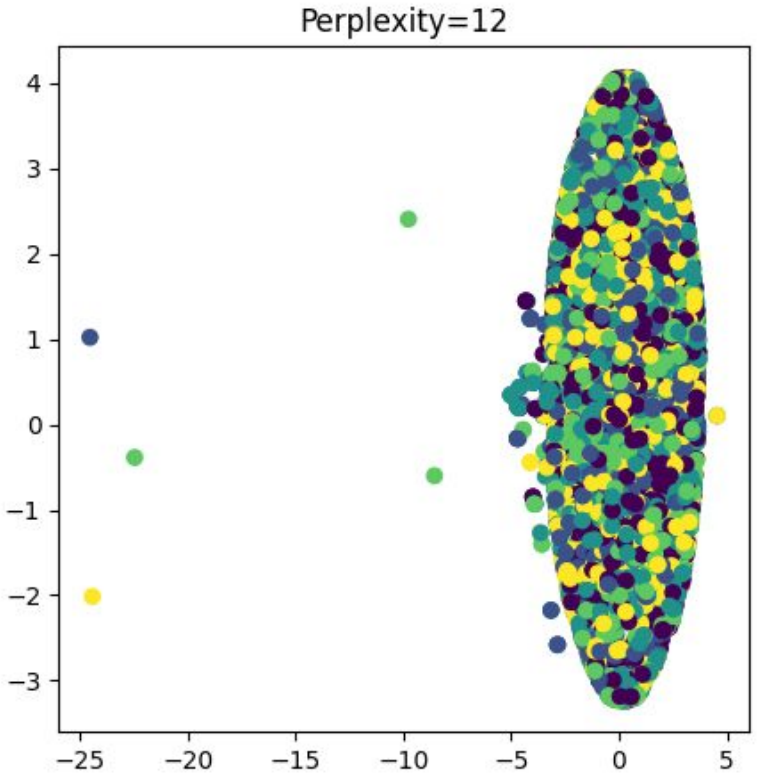
ключевые биграммы позитивных отзывов по топикам (LDA)

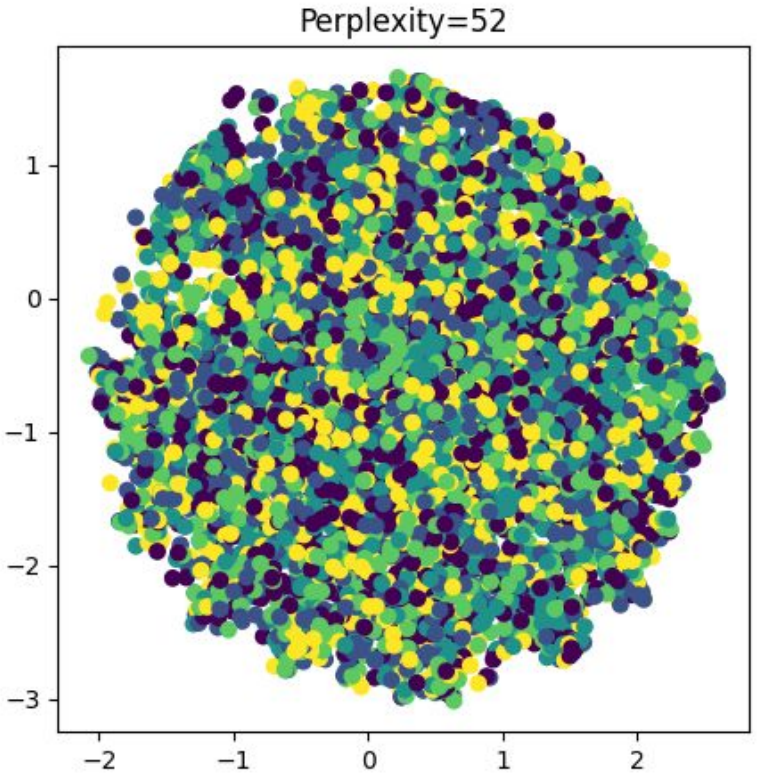


ключевые триграммы негативных отзывов по топикам (LDA)

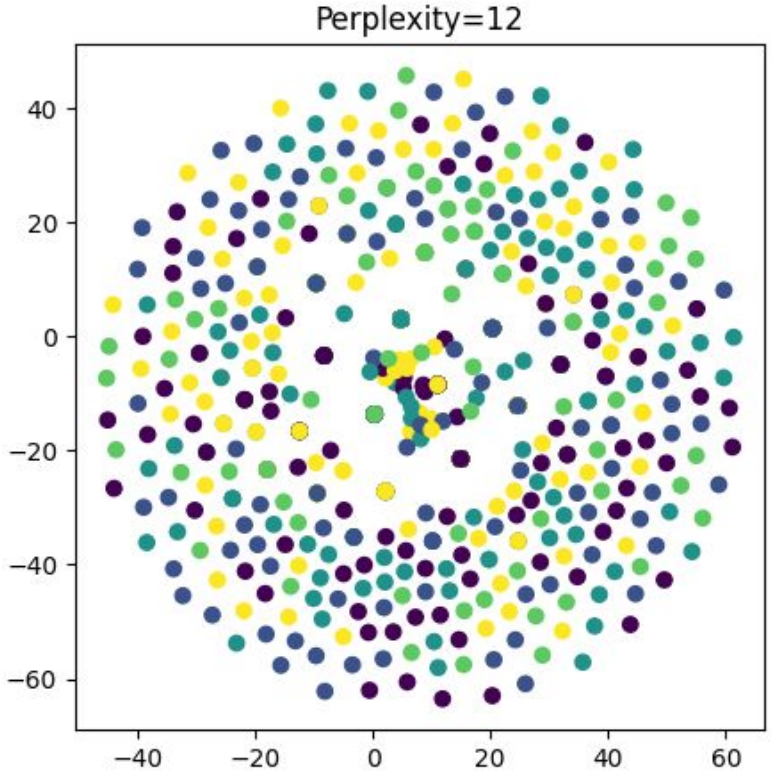


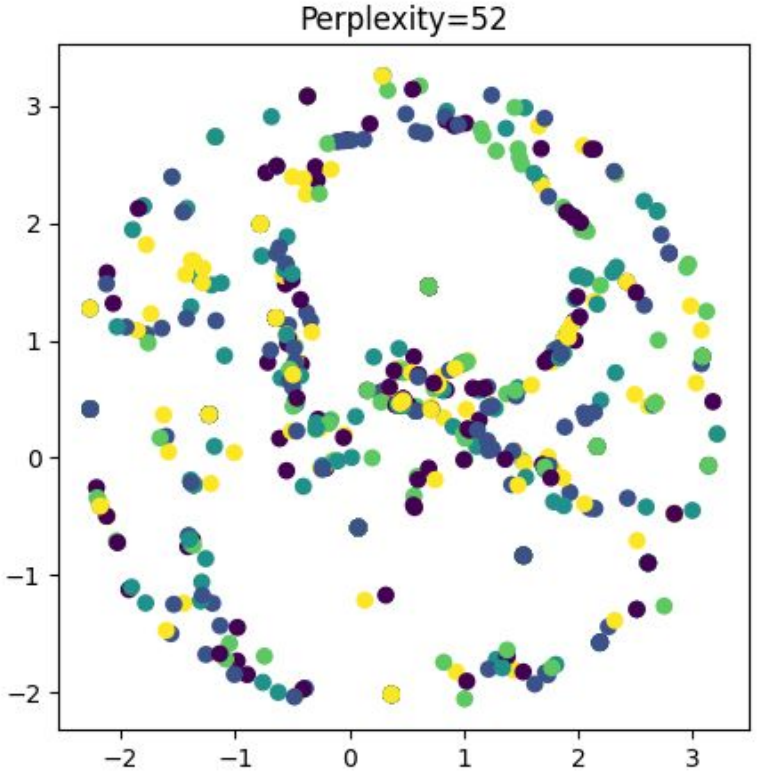
визуализация отзывов с помощью TSNE: позитивные отзывы





визуализация отзывов с помощью TSNE: негативные отзывы





суммаризация отзывов с помощью T5

```
from transformers import GPT2Tokenizer, T5ForConditionalGeneration
sum_tokenizer = GPT2Tokenizer.from_pretrained('RussianNLP/FRED-T5-Summarizer',eos_token='</s>')
device = torch.device('cuda' if torch.cuda.is_available() else 'cpu')
sum_model = T5ForConditionalGeneration.from_pretrained('RussianNLP/FRED-T5-Summarizer').to(device)
```

```
# определим функцию суммаризации, по которой будет работать модель:
def summarize_texts(prompt, text_list):
    input_text = prompt + " ".join(text_list)
    input_ids=torch.tensor([sum_tokenizer.encode(input_text, max_length=512, truncation=True)]).to(device)
    outputs = sum_model.generate(input_ids,
                                 eos_token_id=sum_tokenizer.eos_token_id,
                                 num_beams=4,
                                 min_new_tokens=10,
                                 max_new_tokens=100,
                                 do_sample=False,
                                 no_repeat_ngram_size=2,
                                 early_stopping=True)
    text_list_summary = sum_tokenizer.decode(outputs[0][1:], skip_special_tokens=True)
    return(text_list_summary)
```

примеры суммаризации

“<LM> Определи, что именно понравилось автору отзыва, игнорируй негативные оценки.

5 из 5🖤 Пил кофе и в Риме, и в Париже, но вкуснее, чем капуч на фундучном молоке с фирменными сливками Джинжу, не пробовал ничего! Десерты тоже очень необычные. Ребята - бариста большие молодцы! Что можно улучшить? Маловато места, с посадкой можно что-то придумать?”

Output: “Автор отзыва оценил кафе на 5 из 5, отметив, что капуччино на фундучном молоке с фирменными сливками Джинжу был вкуснее, чем в других местах. Он также отметил необычные десерты и персонал, который был внимательным и отзывчивым.”

“<LM> Определи, на что именно жалуется автор. Игнорируй нейтральные или положительные оценки.

Не очень удобное расположение, от метро идти мин 20 быстрым шагом через промзону. В самом кофе мест очень мало, а желающих очень много((пирожные очень вкусные, кофе…бывает вкуснее. Второй раз именно туда на пойду.”

Output: “Автор жалуется на неудобное расположение и малое количество мест в кафе. Он также отмечает, что пирожные вкусные, но кофе бывает вкуснее.”

оценка суммаризации по ключевым n-граммам

- обучение модели Word2Vec на корпусе нормализованных отзывов;
- получение векторов ТОП-100 n-грамм CountVectorizer оригинальных и суммаризованных отзывов, их усреднение по спискам;
- сравнение векторов с применением косинусного сходства.

Word2Vec позволяет оценить семантическую близость списков n-грамм.

Косинусное сходство для позитивных векторов составило 1, для негативных - 0,94.

подводим итоги

Хорошо работают с отзывами:

- RuBERT for Sentiment Analysis
- CountVectorizer
- TF-IDF
- KeyBERT
- T5

Плохо работают с отзывами:

- RAKE
- LDA

проект подготовила студентка ДПО ВШЭ Компьютерная лингвистика
Черных Мария