Машинное обучение в задах классификации текстов

Волненко Александр
ПМ19-1

Постановка задачи и выбор датасета

Задача: классифицировать текст с помощью машинного обучения

Датасет IMDb

kaggle

Датасет с Kaggle

Предварительный обзор датасета

	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive

Пример данных

```
positive    25000
negative    25000
Name: sentiment, dtype: int64
```

Распределение меток классов

Кол-во дубликатов: 418

Предобработка данных

Предобработка данных:

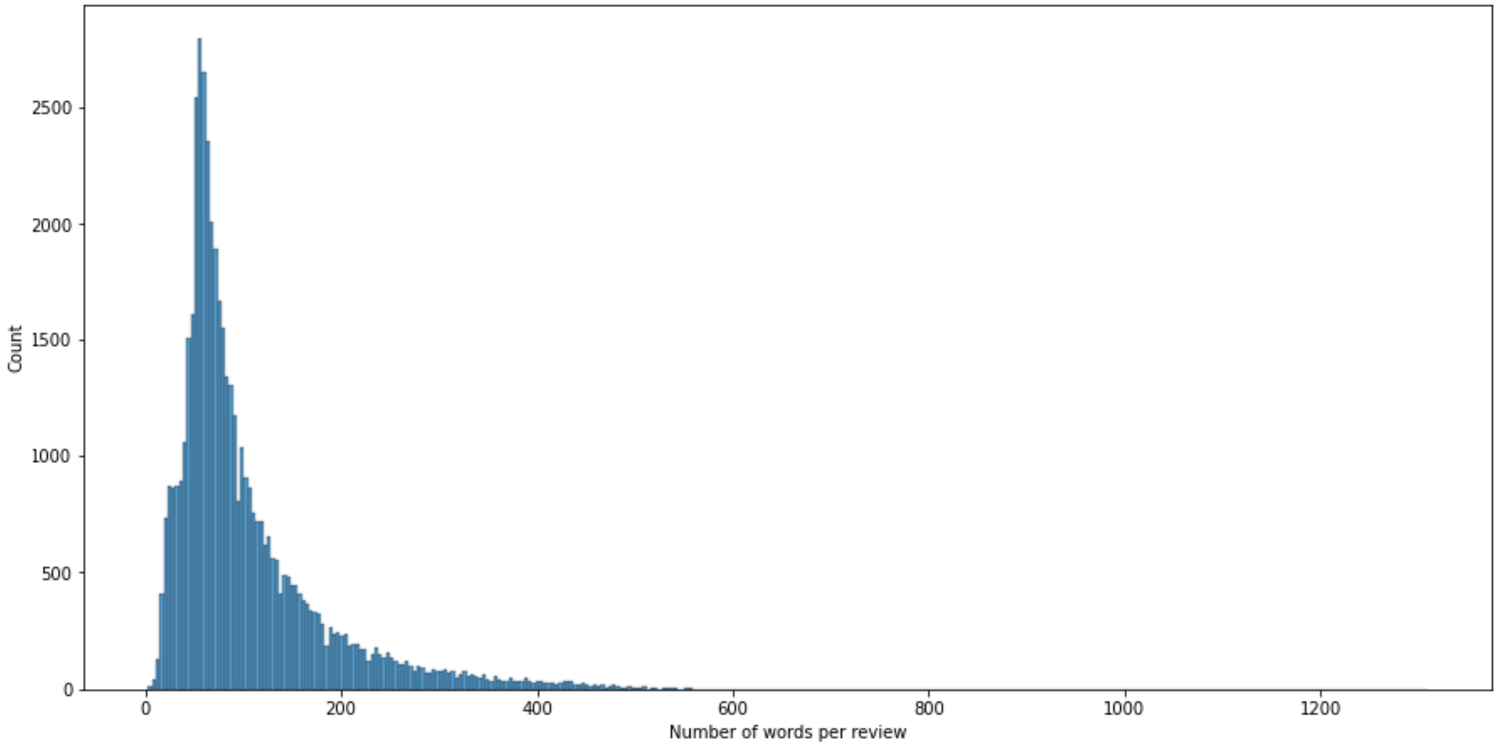
- Удаление дубликатов
- Очистка каждого отзыва:
  - Удаление html тэгов
  - Перевод символов в нижний регистр
  - Замена разговорных выражений
  - Удалены все символы, кроме букв
  - Удаление стоп слов
  - Взята лемма каждого слова

```
mapping = {"ain't": "is not", "aren't": "are not","can't": "cannot",
           "'cause": "because", "could've": "could have", "couldn't": "could not",
           "didn't": "did not",  "doesn't": "does not", "don't": "do not", "hadn't": "had not",
           "hasn't": "has not", "haven't": "have not", "he'd": "he would","he'll": "he will",
           "he's": "he is", "how'd": "how did", "how'd'y": "how do you", "how'll": "how will",
           "how's": "how is",  "I'd": "I would", "I'd've": "I would have", "I'll": "I will",
           "I'll've": "I will have","I'm": "I am", "I've": "I have", "i'd": "i would",
           "i'd've": "i would have", "i'll": "i will",  "i'll've": "i will have",
           "i'm": "i am", "i've": "i have", "isn't": "is not", "it'd": "it would",
           "it'd've": "it would have", "it'll": "it will", "it'll've": "it will have",
           "it's": "it is", "let's": "let us", "ma'am": "madam", "mayn't": "may not",
           "might've": "might have","mightn't": "might not","mightn't've": "might not have",
           "must've": "must have", "mustn't": "must not", "mustn't've": "must not have",
           "needn't": "need not", "needn't've": "need not have","o'clock": "of the clock",
           "oughtn't": "ought not", "oughtn't've": "ought not have", "shan't": "shall not",
           "sha'n't": "shall not", "shan't've": "shall not have", "she'd": "she would",
           "she'd've": "she would have", "she'll": "she will", "she'll've": "she will have",
           "she's": "she is", "should've": "should have", "shouldn't": "should not",
           "shouldn't've": "should not have", "so've": "so have","so's": "so as", "this's": "this is",
           "that'd": "that would", "that'd've": "that would have", "that's": "that is",
           "there'd": "there would", "there'd've": "there would have", "there's": "there is",
           "here's": "here is","they'd": "they would", "they'd've": "they would have",
           "they'll": "they will", "they'll've": "they will have", "they're": "they are",
           "they've": "they have", "to've": "to have", "wasn't": "was not", "we'd": "we would",
           "we'd've": "we would have", "we'll": "we will", "we'll've": "we will have",
           "we're": "we are", "we've": "we have", "weren't": "were not",
           "what'll": "what will", "what'll've": "what will have","what're": "what are",
           "what's": "what is", "what've": "what have", "when's": "when is", "when've": "when have",
           "where'd": "where did", "where's": "where is", "where've": "where have", "who'll": "who will",
           "who'll've": "who will have", "who's": "who is", "who've": "who have", "why's": "why is",
           "why've": "why have", "will've": "will have", "won't": "will not", "won't've": "will not have",
           "would've": "would have", "wouldn't": "would not", "wouldn't've": "would not have",
           "y'all": "you all", "y'all'd": "you all would","y'all'd've": "you all would have",
           "y'all're": "you all are","y'all've": "you all have","you'd": "you would",
           "you'd've": "you would have", "you'll": "you will", "you'll've": "you will have",
           "you're": "you are", "you've": "you have" }
```

wonderful little production film technique unassuming old time bbc fashion give comforting discomforte sense realism entire piece actor extremely choose michael sheen get polari voice pat truly seamless editing guide reference williams diary entry worth watching terrificly write perform piece masterful production great master s comedy life realism come home little thing fantasy guard use traditional dream technique remain solid disappear play knowledge sense particularly scene concern orton halliwell set particularly flat halliwell s mural decorate surface terribly

Пример отзыва после предобработки данных

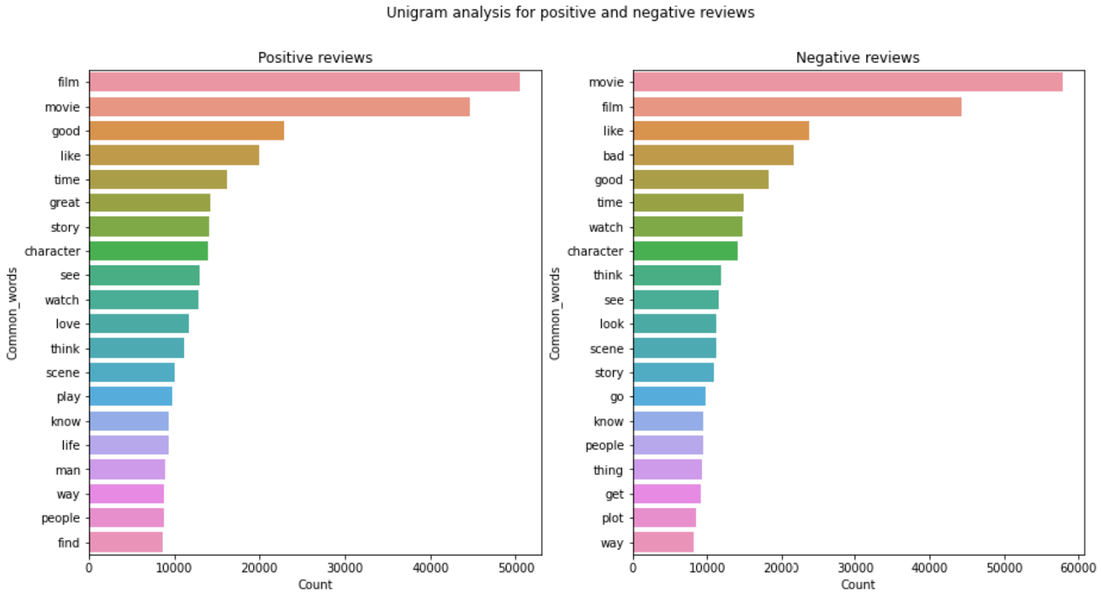
Последующий анализ данных



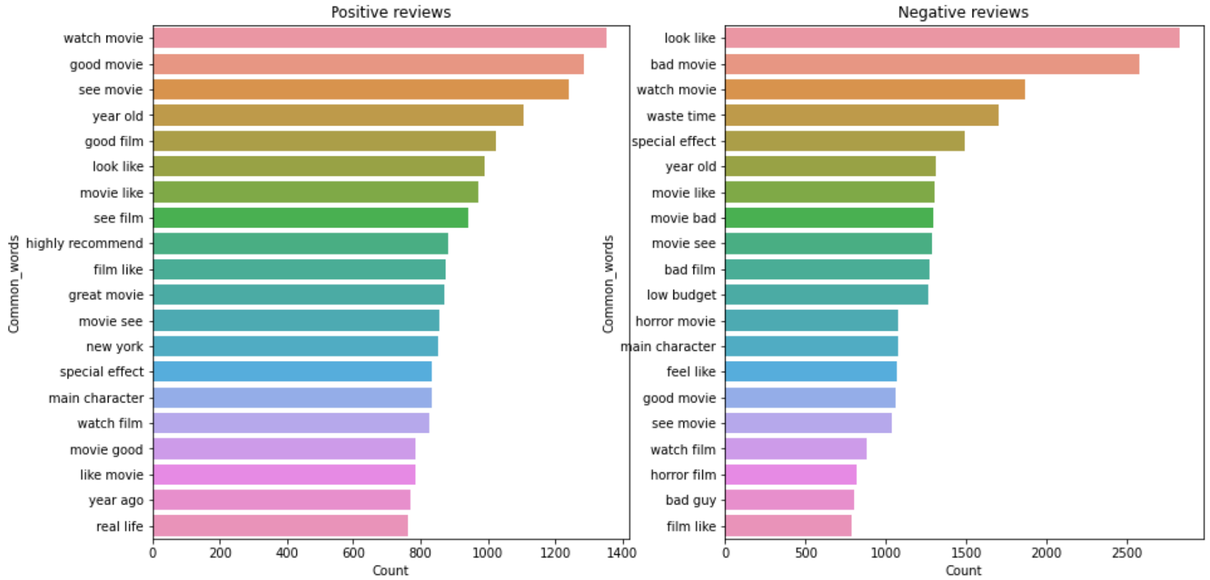
Распределение слов в отзывах

```
count    49582.000000
mean       104.595519
std         80.630538
min          3.000000
25%         55.000000
50%         77.000000
75%        127.000000
max       1310.000000
Name: clean_review, dtype: float64
```

Распределение униграмм и биграмм



Распределение униграм



Распределение биграмм

Классическое машинное обучение

Алгоритмы классического машинного обучения:

- Наивный Байес
- Логистическая регрессия (TF-IDF)

Model	Accuracy
Log Regression	88.97%
Naive Bayes	85.28%

Результаты работы

Предобработка данных для нейронных сетей

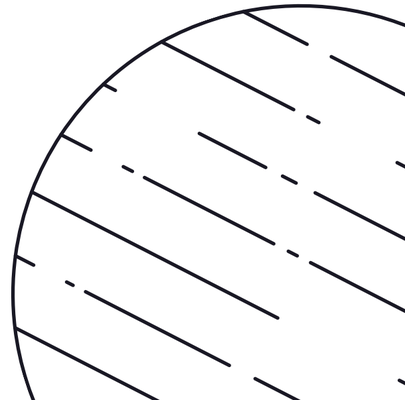
Предобработка данных:

- Определение максимальное длины строки
- Создание словаря

```
count    34707.000000
mean       104.692252
std         80.691193
min          3.000000
25%         55.000000
50%         77.000000
75%        127.000000
max       1007.000000
Name: clean_review, dtype: float64
```

Максимальная длина строки: 266 слов

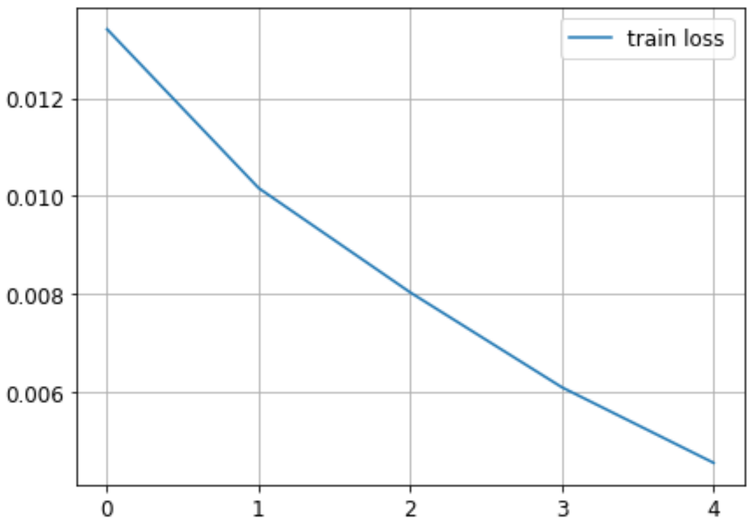
Длина словаря: 73569 уникальный токенов

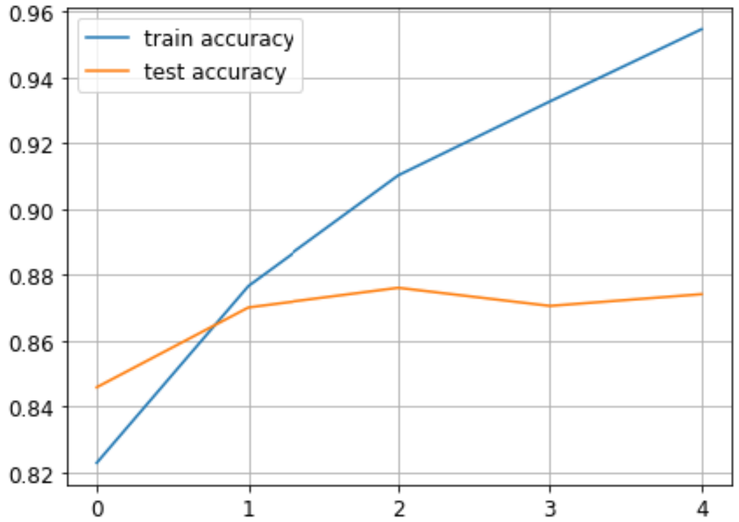


Рекуррентные нейронные сети

Результаты:

- Двухслойная рекуррентная сеть: 87.42%
- Двухслойная рекуррентная сеть с предобученными эмбедингами: 88.40%





Процесс обучения рекуррентной сети

Заключение

В целом задача классификации текстов хорошо решается в данный момент и с появлением все новых методов машинного обучения, а в частности новых архитектур нейронных сетей, качество будет только становится все лучше и лучше. Безусловно, с развитием сферы машинного обучения и нахождением большего количества возможностей для его применений в бизнесе, задача классификации текстов будет становится только более востребованней.