In [5]:

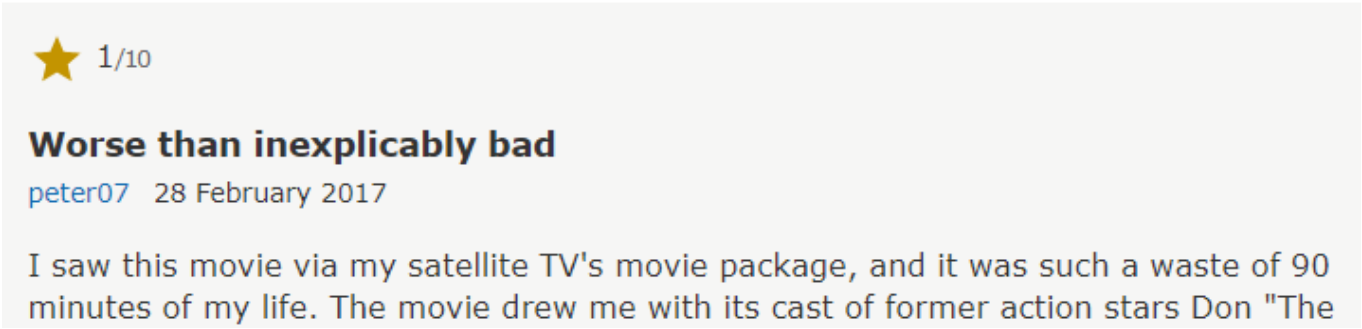
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

%matplotlib inline

plt.style.use('ggplot')
plt.rcParams['figure.figsize'] = (12,5)

# Для кириллицы на графиках
font = {'family': 'Arial'}
plt.rc('font', **font)

try:
    from ipywidgets import interact, IntSlider, fixed, FloatSlider
except ImportError:
    print u'Так надо'
```

ТЕХНОСФЕРА

Предсказание оценки пользователя фильму по тексту отзыва

Докшина Елизавета
Жолковский Евгений

Постановка задачи

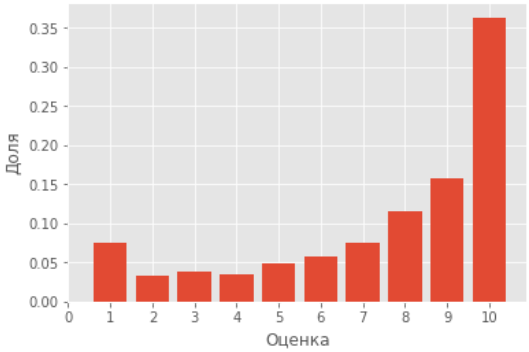


Получение и предобработка данных

- Парсинг - BeautifulSoup
- Стемминг - Porter Stemmer

- 400 тысяч отзывов
- 1500 фильмов
- более 1 Гб данных

Распределение числа оценок



Выбор критерия качества

Mean Absolute Error:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|x_{pred}^{(i)} - x_{true}^{(i)}\right|$$

Предобработка текста

- Предобработка текста и токенизация: `CountVectorizer`
- tf-idf трансформация: `TfidfTransformer`

Baseline

Константное предсказание: `y_pred = 8`

`MAE = 2.37`

Рассматриваемые модели

- Линейная регрессия
- Multinomial Naive Bayes
- Логистическая регрессия

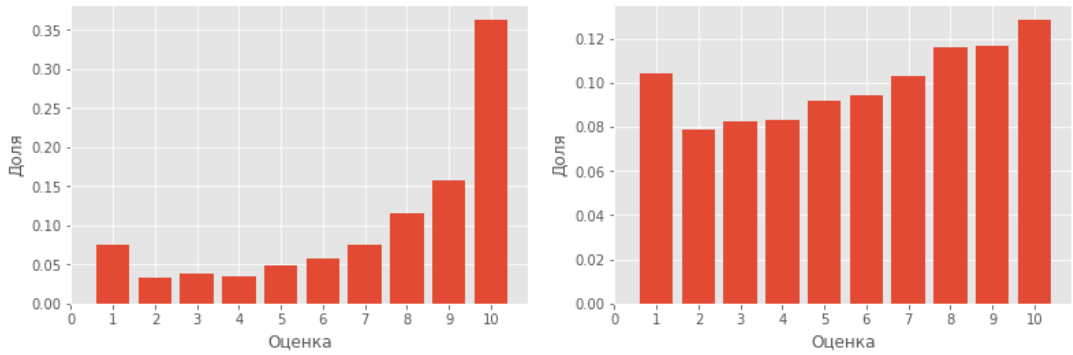
Выбор гиперпараметров для моделей

- Длина n-граммы
- Минимальное и максимальное значения document frequency
- Использование idf
- для NaiveBayes:
  - сглаживание
- для логистической регрессии
  - веса классов
  - penalty

Нормировка обучающей выборки

- LinearRegression
  - normed train mean MAE 1.3439019177
  - not normed train mean MAE 1.1931498474
- MultinomialNB
  - normed train mean MAE 0.976856316297
  - not normed train mean MAE 1.13093131456
- LogisticRegression
  - normed train mean MAE 0.93764228693
  - not normed train mean MAE 0.922247959518

Нормировка обучающей выборки



Результаты

Линейная регрессия

- Ненормированная выборка
  - Кросс-валидация: `<MAE> = 1.21`
  - На тестовых данных: `MAE = 1.20`
- Нормированная выборка
  - Кросс-валидация: `<MAE> = 1.29`
  - На тестовых данных: `MAE = 1.30`

Самые весомые слова

- ### Позитивные: hooked, perfect, favorite, best, superb, brilliant, excellent, amazing, rare, perfectly, incredible, awesome, masterpiece, intense, outstanding, hilarious, great, fantastic, greatest, wonderful, unique, notch, twists, beautifully, loved, fascinating
- ### Негативные: worst, awful, waste, garbage, torture, crap, lacks, poorly, boring, avoid, terrible, overrated, fails, worse, disappointment, wasted, ridiculous, horrible, bland, disappointing, supposedly, laughable, nothing, whatsoever, thin, supposed, mess, sucks, unfortunately, sorry, badly, instead, substance, stupid, poor, nonsense, dull, painful, lame, bad, disappointed, shame

Multinomial Naive Bayes

- Ненормированная выборка
  - Кросс-валидация: `<MAE> = 1.15`
  - На тестовых данных: `MAE = 1.16`
- Нормированная выборка
  - Кросс-валидация: `<MAE> = 1.00`
  - На тестовых данных: `MAE = 1.02`

Логистическая регрессия

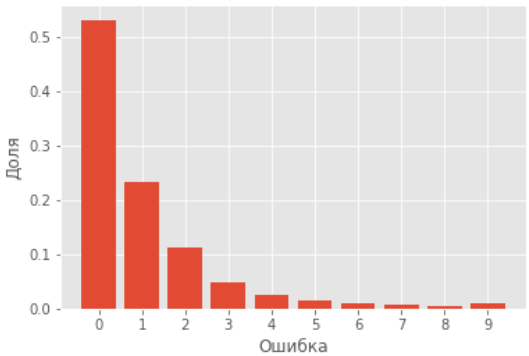
- Ненормированная выборка
  - Кросс-валидация: `<MAE> = 0.92`
  - На тестовых данных: `MAE = 0.94`
- Нормированная выборка
  - Кросс-валидация: `<MAE> = 1.02`
  - На тестовых данных: `MAE = 1.03`

Результаты на чисто тестовой выборке

</br> </br>

	LinReg	LogReg	MNbayes
ненормированная	1.21	0.94	1.16
нормированная	1.29	1.02	1.02

Распределение ошибки



- более 50% - точное попадание
- более 90% - ошибка не более трех

Спасибо за внимание!