

# Разметка наборов данных

## Описание

В данной работе рассматривается оригинальный необработанный набор Large Movie Review Dataset, а именно самая первая его версия, составленная 13 лет назад в 2011 году.

Основные причины выбора данного набора:

- оригинальный датасет
- необходимость скрейпинга и парсинга данных
- содержит 100 тысяч отзывов, 50 тысяч из которых вообще не имеют оценок
- практика работы с большими данными в label studio

Архивный файл находится на [странице](#), автором которой является [Эндрю Маас](#), профессор Стенфордского университета.

В архивном файле, кроме сотни тысячи файлов с отзывами, присутствуют служебные файлы, представляющие научный интерес:

- README - содержит описание от автора
- \*.vocab - словарь, содержащий 89 526 английских слов
- \*.feat - файлы в формате LIBSVM, содержащие матрицу векторов для каждого маркированного слова
- url.txt - файлы со ссылками на оригинальный отзыв

Один файл это отзыв без маркировки (без оценки). Мне нужно обработать все сто тысяч, извлечь из каждого отзыв, присвоить ему нужный лейбл и объединить их, в зависимости от контекста, с нужным датасетом, в формате пригодным для импорта в jupyter-ноутбук.

Наборы данных имеют вложенную иерархию папок со следующей структурой:

```
.
├── test
│   ├── neg
│   └── pos
└── train
    ├── neg
    ├── pos
    └── unsup
```

19865600 bytes used in 7 directories

В папке **test** находятся тестовые данные, которые разбиты еще на две подпапки **neg** и **pos**: негативные и позитивные отзывы.

В папке **train**, набор таких же папок, как и в test: **neg** и **pos**, там же расположена еще одна, представляющая для меня интерес, папка **unsup** с неразмеченными отзывами.

Всю работу от получения данных их обработки, разметки до описания выводов по машинному обучению, можно разбить на 15 этапов (задач)

# Задачи

1. установка и подключение необходимых библиотек
2. скрейпинг архива с оригинального сайта
3. распаковка архива
4. парсинг 100 тысяч маленьких файлов: чистка от мусора и лишних символов
5. rule-based labeling - создание 5 файлов: test\_neg.txt, test\_pos.txt, train\_neg.txt, train\_pos.txt, unsup.txt
6. загрузка файлов в jupyter-блокнот с авторазметкой столбца label
7. перемешивание данных
8. обучение модели на тренировочном датасете
9. проверка на тестовом датасете
10. расчет эффективности модели
11. предсказание большого (неразмеченного) датасета
12. сохранение большого датасета в csv файл
13. загрузка большого датасета в label-studio
14. ручная проверка выборочных отзывов из большого датасета
15. выводы

## 1. установка и подключение необходимых библиотек

Раскомментируйте строку ниже в случае необходимости

```
In [1]: # pip install scikit-learn pandas numpy matplotlib requests tqdm
```

```
In [2]: from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import f1_score
from sklearn.utils import shuffle
from tqdm import tqdm
import pandas as pd
import numpy as np
import requests
import tarfile
import glob
import os
import re
import warnings
warnings.filterwarnings('ignore')
```

## 2. скрейпинг архива

```
In [3]: url = 'https://ai.stanford.edu/~amaas/data/sentiment/aclImdb_v1.tar.gz'
response = requests.get(url, stream=True)
total_size = int(response.headers.get('content-length', 0))
block_size = 1024

with open('archive.tar.gz', 'wb') as file:
    with tqdm(total=total_size, unit='B', unit_scale=True, ncols=100) as pbar:
```

```
for data in response.iter_content(block_size):
    file.write(data)
    pbar.update(len(data))
```

100% |██| 84.1M/84.1M [00:31<00:00, 2.65MB/s]

### 3. распаковка архива

```
In [4]: with tarfile.open("archive.tar.gz", 'r:gz', errorlevel=0) as tar:
        members = tar.getmembers()
        for member in tqdm(members, desc='Extracting', unit='files', ncols=100):
            tar.extract(member)
```

Extracting: 100% |██| 100019/100019 [01:13<00:00, 1369.78files/s]

### 4. функция парсинга файлов базы

```
In [5]: def process_text_files(file_pattern, output_file):
        all_text = ""
        file_list = glob.glob(file_pattern)
        with tqdm(total=len(file_list), unit='file') as pbar:
            for filename in file_list:
                with open(filename, 'r', encoding='utf-8') as file:
                    text = file.read()
                    text = re.sub(r'<br />|\t|\n|;', ' ', text) # Замена указанных подстрок
                    text += ';' # Добавление символа ';' в конец строки
                    text += '\n' # Добавление символа '\n' в конец строки
                    all_text += text
                pbar.update(1)

        with open(output_file, 'w', encoding='utf-8') as outfile:
            outfile.write(all_text)
```

### 5. rule-based labeling

```
In [6]: # Обработка файлов в папке aclImdb/train/pos
        process_text_files('aclImdb/train/pos/*.txt', 'train_pos.txt')
```

0% |██| 10/12500 [00:00<02:05, 99.37file/s]100% |██| 12500/12500 [02:19<00:00, 89.81file/s]

```
In [7]: # Обработка файлов в папке aclImdb/train/neg
        process_text_files('aclImdb/train/neg/*.txt', 'train_neg.txt')
```

100% |██| 12500/12500 [02:13<00:00, 93.50file/s]

```
In [8]: # Обработка файлов в папке aclImdb/test/pos
        process_text_files('aclImdb/test/pos/*.txt', 'test_pos.txt')
```

100% |██| 12500/12500 [02:12<00:00, 94.16file/s]

```
In [9]: # Обработка файлов в папке aclImdb/test/neg
        process_text_files('aclImdb/train/neg/*.txt', 'test_neg.txt')
```

100% |██| 12500/12500 [00:43<00:00, 284.54file/s]

```
In [10]: # Обработка файлов в папке aclImdb/train/unsup
process_text_files('aclImdb/train/unsup/*.txt', 'unsup.txt')
```

100%|██████████| 50000/50000 [17:43<00:00, 47.01file/s]

## 6. загрузка файлов в jupyter-блокнот с авторазметкой столбца label

```
In [11]: test_neg = pd.read_csv('test_neg.txt', engine='python',
                                on_bad_lines='warn', header=None, names=['text', 'label'], sep=';', encoding='utf-8')
test_neg["label"] = 0
test_neg
```

```
Out[11]:
```

	text	label
0	Story of a man who has unnatural feelings for ...	0
1	Airport '77 starts as a brand new luxury 747 p...	0
2	This film lacked something I couldn't put my f...	0
3	Sorry everyone,,, I know this is supposed to b...	0
4	When I was little my parents took me along to ...	0
...	...	...
12495	Towards the end of the movie, I felt it was to...	0
12496	This is the kind of movie that my enemies cont...	0
12497	I saw 'Descent' last night at the Stockholm Fi...	0
12498	Some films that you pick up for a pound turn o...	0
12499	This is one of the dumbest films, I've ever se...	0

12500 rows × 2 columns

```
In [12]: test_pos = pd.read_csv('test_pos.txt', engine='python',
                                on_bad_lines='warn', header=None, names=['text', 'label'], sep=';', encoding='utf-8')
test_pos["label"] = 1
test_pos
```

Out[12]:

	text	label
0	I went and saw this movie last night after bei...	1
1	Actor turned director Bill Paxton follows up h...	1
2	As a recreational golfer with some knowledge o...	1
3	I saw this film in a sneak preview, and it is ...	1
4	Bill Paxton has taken the true story of the 19...	1
...	...	...
12495	I was extraordinarily impressed by this film. ...	1
12496	Although I'm not a golf fan, I attended a snea...	1
12497	From the start of The Edge Of Love , the view...	1
12498	This movie, with all its complexity and subtle...	1
12499	I've seen this story before but my kids haven'...	1

12500 rows × 2 columns

```
In [13]: train_neg = pd.read_csv('train_neg.txt', engine='python',
                                on_bad_lines='warn', header=None, names=['text', 'label'], sep=';', encoding='utf-8')
train_neg["label"] = 0
train_neg
```

Out[13]:

	text	label
0	Story of a man who has unnatural feelings for ...	0
1	Airport '77 starts as a brand new luxury 747 p...	0
2	This film lacked something I couldn't put my f...	0
3	Sorry everyone,,, I know this is supposed to b...	0
4	When I was little my parents took me along to ...	0
...	...	...
12495	Towards the end of the movie, I felt it was to...	0
12496	This is the kind of movie that my enemies cont...	0
12497	I saw 'Descent' last night at the Stockholm Fi...	0
12498	Some films that you pick up for a pound turn o...	0
12499	This is one of the dumbest films, I've ever se...	0

12500 rows × 2 columns

```
In [14]: train_pos = pd.read_csv('train_pos.txt', engine='python',
                                   on_bad_lines='warn', header=None, names=['text', 'label'], sep=';', encoding='utf-8')
train_pos["label"] = 1
train_pos
```

```
Out[14]:
```

	text	label
<b>0</b>	Bromwell High is a cartoon comedy. It ran at t...	1
<b>1</b>	Homelessness (or Houselessness as George Carli...	1
<b>2</b>	Brilliant over-acting by Lesley Ann Warren. Be...	1
<b>3</b>	This is easily the most underrated film inn th...	1
<b>4</b>	This is not the typical Mel Brooks film. It wa...	1
...	...	...
<b>12495</b>	Seeing as the vote average was pretty low, and...	1
<b>12496</b>	The plot had some wretched, unbelievable twist...	1
<b>12497</b>	I am amazed at how this movie(and most others ...	1
<b>12498</b>	A Christmas Together actually came before my t...	1
<b>12499</b>	Working-class romantic drama from director Mar...	1

12500 rows × 2 columns

```
In [15]: unsup = pd.read_csv('unsup.txt', engine='python',
                        on_bad_lines='warn', header=None, names=['text', 'label'], sep=';', encoding='utf-8')
unsup
```

```
Out[15]:
```

	text	label
<b>0</b>	I admit, the great majority of films released ...	NaN
<b>1</b>	Take a low budget, inexperienced actors doubli...	NaN
<b>2</b>	Everybody has seen 'Back To The Future,' right...	NaN
<b>3</b>	Doris Day was an icon of beauty in singing and...	NaN
<b>4</b>	After a series of silly, fun-loving movies, 19...	NaN
...	...	...
<b>49995</b>	Delightfully awful! Made by David Giancola, a ...	NaN
<b>49996</b>	Watching Time Chasers, it obvious that it was ...	NaN
<b>49997</b>	At the beginning we can see members of Troma t...	NaN
<b>49998</b>	The movie was incredible, ever since I saw it ...	NaN
<b>49999</b>	TCM came through by acquiring this wonderful, ...	NaN

50000 rows × 2 columns

## 7. перемешивание данных

```
In [16]: test = pd.concat([test_neg, test_pos], ignore_index=True)
test = shuffle(test)
test
```

Out[16]:

	text	label
<b>21349</b>	Lifeforce starts in outer space where the HMS ...	1
<b>2697</b>	I loved the first two movies, but this movie w...	0
<b>18517</b>	All the characters in this cartoon were hilari...	1
<b>1940</b>	Although it got some favorable press after pla...	0
<b>2974</b>	The British claymation series putting witty ...	0
...	...	...
<b>10330</b>	Suppose you've been on a deserted island the l...	0
<b>9918</b>	This is a badly made, poor remake of Bimalda's...	0
<b>23950</b>	I usually steer clear of Film Festivals and do...	1
<b>10027</b>	What a load of rubbish.. I can't even begin to...	0
<b>12055</b>	Very strange but occasionally elegant exploita...	0

25000 rows × 2 columns

```
In [17]: train = pd.concat([train_neg, train_pos], ignore_index=True)
train = shuffle(train)
train
```

Out[17]:

	text	label
<b>4762</b>	This movie was absolutely ghastly! I cannot fa...	0
<b>22652</b>	The Movie Freddy's dead the final nightmare is...	1
<b>18761</b>	This is one powerful film. The first time I sa...	1
<b>14205</b>	Bill Crain's rarer than rare 'slasher' movie c...	1
<b>16841</b>	Kubrick again puts on display his stunning abi...	1
...	...	...
<b>2850</b>	Oliver Hardy awakens with a hangover and soon ...	0
<b>12838</b>	Bruce Almighty looks and sounds incredibly s...	1
<b>12378</b>	Why is it that any film about Cleopatra, the l...	0
<b>7541</b>	This is not really a zombie film, if we're def...	0
<b>19188</b>	The only show I have watched since 90210! Why ...	1

25000 rows × 2 columns

## 8. обучение модели на тренировочном датасете

```
In [18]: def train_model(label):
vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(train['text'])
y = label['label']
```

```
model = LogisticRegression()  
model.fit(X, y)  
return model, vectorizer
```

```
In [19]: model, vectorizer = train_model(train)
```

## 9. проверка на тестовом датасете

```
In [20]: x_test = vectorizer.transform(test['text'])  
y_test = model.predict(x_test)
```

```
In [21]: test['predicted'] = y_test  
test.head()
```

Out[21]:

	text	label	predicted
21349	Lifeforce starts in outer space where the HMS ...	1	0
2697	I loved the first two movies, but this movie w...	0	0
18517	All the characters in this cartoon were hilari...	1	1
1940	Although it got some favorable press after pla...	0	0
2974	The British claymation series putting witty ...	0	0

## 10. эффективность модели

```
In [22]: test['loss'] = test['label'] ^ test['predicted']  
test
```

Out[22]:

	text	label	predicted	loss
21349	Lifeforce starts in outer space where the HMS ...	1	0	1
2697	I loved the first two movies, but this movie w...	0	0	0
18517	All the characters in this cartoon were hilari...	1	1	0
1940	Although it got some favorable press after pla...	0	0	0
2974	The British claymation series putting witty ...	0	0	0
...	...	...	...	...
10330	Suppose you've been on a deserted island the l...	0	0	0
9918	This is a badly made, poor remake of Bimalda's...	0	0	0
23950	I usually steer clear of Film Festivals and do...	1	1	0
10027	What a load of rubbish.. I can't even begin to...	0	0	0
12055	Very strange but occasionally elegant exploita...	0	0	0

25000 rows × 4 columns

```
In [23]: guess, loss = test['loss'].value_counts()
```



```
In [24]: print(f"Предсказано: {guess / (loss+guess) * 100} %")
```

Предсказано: 90.572 %

```
In [25]: f1 = f1_score(test["label"], y_test)
```

```
In [26]: print(f"Метрика эффективности F1: {f1}")
```

Метрика эффективности F1: 0.9033976802327964

## 11. предсказание большого (неразмеченного) датасета

```
In [27]: x_unsup = vectorizer.transform(unsup['text'])
y_unsup = model.predict(x_unsup)
```

```
In [28]: unsup['predicted'] = y_unsup
unsup
```

```
Out[28]:
```

	text	label	predicted
0	I admit, the great majority of films released ...	NaN	1
1	Take a low budget, inexperienced actors doubli...	NaN	0
2	Everybody has seen 'Back To The Future,' right...	NaN	0
3	Doris Day was an icon of beauty in singing and...	NaN	1
4	After a series of silly, fun-loving movies, 19...	NaN	1
...	...	...	...
49995	Delightfully awful! Made by David Giancola, a ...	NaN	1
49996	Watching Time Chasers, it obvious that it was ...	NaN	0
49997	At the beginning we can see members of Troma t...	NaN	0
49998	The movie was incredible, ever since I saw it ...	NaN	1
49999	TCM came through by acquiring this wonderful, ...	NaN	1

50000 rows × 3 columns

## 12. сохранение большого датасета в табулированный csv файл

```
In [29]: unsup.to_csv('unsup.tsv', index=False, encoding='utf-8', sep='\t')
```

## 13. загрузка большого датасета в label-studio

Загрузка огромного CSV файла 68 Мегабайт, с 50 тысячами строк, оказалось не простым делом для label studio (LS).

Постоянно вываливались ошибки, о большом количестве SQL данных. Т.к. под капотом LS переводит полученную информацию в json-формат и умеет работать с SQL, то в огромном массиве английских слов встречаются их комбинации со спецсимволами, которые LS воспринимает как служебные инструкции для баз данных.

Удалось победить данную проблему с помощью табулированного формата CSV или как его обозначают TSV. Ошибки все равно появлялись, но данные полностью загрузились и с ними можно работать.

## Create Project

Project Name

Data Import

Labeling Setup

Delete

Save

Project Name

GeekBrains

Description

Разметка наборов данных

Workspace 🚀 Enterprise

Select an option



Simplify project management by organizing projects into workspaces. [Learn more](#)

После создания проекта LS, необходимо выбрать тип данных для маркировки:

# Create Project

Project Name

Data Import

Labeling Setup

Delete

Save

Computer Vision

Natural Language  
Processing

Audio/Speech  
Processing

Conversational AI

Ranking & Scoring

Structured Data  
Parsing

Time Series Analysis

Videos

Generative AI

Custom template

Please read the passage

The boundary of the region from which no escape is possible is called the event horizon. Although the event horizon has an enormous effect on the fate and circumstances of an object crossing it, according to general relativity it has no locally detectable feature. In many ways, a black hole acts like an ideal black body, as it reflects no light. [5][6] A quantum field theory in curved spacetime predicts that event horizons emit Hawking radiation, with the same spectrum as a black body of a temperature inversely proportional to its mass. [7][8] This temperature is on the order of billionths of a kelvin for black holes of stellar mass, making it essentially impossible to observe directly.

Select a text span answering the following question:

How black holes could be detected?

Question Answering

To have faith is to trust yourself to the water

Opossum x Extraterrestrial x

- ☐ Archaea
- ☐ Bacteria
- ☐ Eukarya
  - ☐ Human
  - ☒ Opossum
  - ☒ Extraterrestrial

This is a very right on case movie that delivers everything almost right

Choose text sentiment

- ☐ Positive[1]
- ☐ Negative[2]
- ☒ Neutral[3]

Text Classification

Organization[1] Person[2] Datetime[3]

Microsoft Organization was founded by Bill Gates Person and  
to develop and sell BASIC interpreters for the AI  
Datetime

PER[1] ORG[2] LOC[3] MISC[4]

A Florida LOC restaurant paid 10,925 pounds  
which Hendrix PER penned on a piece of Lor

Named Entity Recognition

Read the sentence in English

The Sun is the star at the center of the Solar System. It is a nearly perfect sphere of hot plasma, [18][19] heated to incandescence by nuclear fusion reactions in its core, radiating the energy mainly as visible light and infrared radiation. It is by far the most important source of energy for life on Earth. Its diameter is about 1.39 million kilometres (864,000 miles), or 109 times that of Earth. Its mass is about 330,000 times that of Earth, and accounts for about 99.86% of the total mass of the Solar System. [20] Roughly three quarters of the Sun's mass consists of hydrogen (~73%); the rest is mostly helium (~25%), with much smaller quantities of heavier elements, including oxygen, carbon, neon,

Provide translation in Spanish

El Sol es la estrella en el centro del Sistema Solar. Es una esfera casi perfecta de plasma caliente, [18][19] calentado hasta la incandescencia por reacciones de fusión nuclear en su núcleo, irradiando la energía principalmente como luz visible y radiación infrarroja. Es, con mucho, la fuente de energía más importante para la vida en la Tierra. Su diámetro es de aproximadamente 1,39 millones de kilómetros (864,000 millas), o 109 veces el de la Tierra. Su masa es aproximadamente 330,000 veces la de la Tierra y representa aproximadamente el 99,86% de la masa total del Sistema

Далее, задаем два лейбла для маркировки: Positive, Negative

# Create Project

Project Name

Data Import

Labeling Setup

Delete

Save

Browse Templates

Code

Visual

## Configure data

Use text from <set manually> \$text

## Add choices

Use new line as a separator to add multiple labels

## Choices (2)

Positive

Negative

Add

## UI Preview

This is a great 3D movie that delivers everything almost right in your face.

## Choose text sentiment

☐ Positive<sup>[1]</sup> ☐ Negative<sup>[2]</sup>

Regions

History

Relations

Info

Comments

Manual

By Time

Regions not added

Выбираем TSV файл и импортируем его в проект LS:

## Create Project

Project Name

Data Import

Labeling Setup

Delete

Save

Dataset URL

Add URL


or

 Upload Files

Drag & drop files here  
or click to browse



Text	txt
Audio	wav, mp3, flac, m4a, ogg
Video	mpeg4/H.264 webp, webm*
Images	Jpg, jpeg, png, gif, bmp, svg, webp
HTML	html, htm, xml
Time Series	csv, tsv

 See the documentation to [import preannotated data](#) or to [sync data from a database or cloud storage](#).

## 14. ручная проверка выборочных отзывов

Загруженные данные в LS представлены в табличном виде

**Label Studio**

Projects / GeekBrains

Settings

AL

Default

Tasks: 50000 / 50000 Annotations: 0 Predictions: 0

Actions

Columns

Filters

Order

not set

Label All Tasks

Import



Export

List


Grid



<input type="checkbox"/>	ID	Completed				Annotated by	text
<input type="checkbox"/>	1		0	0	0		I admit, the great majority of films released before say 1933 are just not for me. Of
<input type="checkbox"/>	2		0	0	0		Take a low budget, inexperienced actors doubling as production
<input type="checkbox"/>	3		0	0	0		Everybody has seen 'Back To The Future,' right? Whether you LIKE that
<input type="checkbox"/>	4		0	0	0		Doris Day was an icon of beauty in singing and acting by her warm voice and
<input type="checkbox"/>	5		0	0	0		After a series of silly, fun-loving movies, 1955 was a big year for Doris Day. That
<input type="checkbox"/>	6		0	0	0		This isn't exactly a musical, but it almost seems like one because there more singing
<input type="checkbox"/>	7		0	0	0		After seven years and seventeen pictures at Warner Brothers Doris Dav

После нажатия на кнопку **Label All Tasks** запустится процесс ручной разметки с помощью "карточек" с отзывами и возможность присвоить нужный лейбл:

 **Label Studio** 

Projects / GeekBrains / Labeling





Settings 

#1  
1 of 1  


I admit, the great majority of films released before say 1933 are just not for me. Of the dozen or so major silents I have viewed, one I loved (The Crowd), and two were very good (The Last Command and City Lights, that latter Chaplin circa 1931). So I was apprehensive about this one, and humor is often difficult to appreciate (uh, enjoy) decades later. I did like the lead actors, but thought little of the film. One intriguing sequence. Early on, the guys are supposed to get de-loused and for about three minutes, fully dressed, do some schtick. In the background, perhaps three dozen men pass by, all naked, white and black (WWI ?), and for most, their butts, part or full backside, are shown. Was this an early variation of beefcake courtesy of Howard Hughes?




### Choose text sentiment

☐ Positive<sup>[1]</sup> ☐ Negative<sup>[2]</sup>

Skip




Submit 

Info





History

Selection Details

Regions

Relations

 Manual  By Time  

Regions not added

Благодаря развитию нейронных сетей, машинный перевод с иностранных языков позволяет получить приемлемый русский текст за доли секунд. Встроенная функция перевода в интернет браузере:

Это не совсем мюзикл, но он кажется таковым, потому что в нем больше пения, чем во многом другом, но это, очевидно, один из крючков. Однажды в баре, где она работала, Рут Эттинг (Дорис Дэй) познакомилась с продюсером Мартином Снайдером (номинант на премию «Оскар» Джеймс Кэгни), который мог пообещать новую и более высокооплачиваемую карьеру на сцене. Она не сразу приступает к делу, так как она включена в течение нескольких секунд, но вскоре у нее появляется шанс на более длительные шоу. Она наслаждается собой, а критики и продюсеры забрасывают ее предложениями. Позже, после свадьбы, Рут и Мартин становятся еще больше, и она становится одной из самых успешных женщин на сцене, а ее карьера вскоре приходит в Голливуд. Только она обнаруживает, что над фильмом работает старый друг (почти любовь), Джонни Олдермен (Кэмерон Митчелл), что для нее и брака/отношений с Мартином все идет немного не так. Также в главных ролях Роберт Кит в роли Бернарда В. Лумиса, Том Талли в роли Фробишера и Гарри Беллавер в роли Джорджи. Он получил «Оскар» за лучший сценарий к фильму, а также был номинирован на «Лучшую песню» за «I'll Never Stop Loving You», «Лучшую музыку» для Перси Фейта и Джорджа Э. Столла, «Лучший звук» и «Лучший сценарий». Дорис Дэй заняла 84-е место в списке 100 величайших кинозвезд, а Джеймс Кэгни занял 8-е место в списке 100 лет, 100 звезд. Очень хорошо!

### Choose text sentiment

☐ Positive<sup>[1]</sup> ☐ Negative<sup>[2]</sup>

После окончания ручной выборочной разметки получаем таблицу, в которой виден наш лейб и лейбл расставленный с помощью машинного обучения.



Действия		Столбцы	Фильтры	Порядок		Не задано	Пометить все задачи				
<input type="checkbox"/>	ИДЕ	Завершённый				Аннот	СМС	txt	предсказе	ярлык	Ул
возможности, и вы не можете ожидать большего, чем дает вам											
<input type="checkbox"/>	3	06 Apr 2024, 00:0	1	0	0	Аль	Все смотрели «Назад в будущее», верно? Независимо от того, нравится вам этот фильм или нет, вы видели пример того, как сделать так, чтобы фильм о путешествии во времени работал. В этом		0		
<input type="checkbox"/>	4	06 Apr 2024, 00:1	1	0	0	Аль	Дорис Дэй была иконой красоты в пении и актерской игре благодаря своему теплему голосу и гениальной игре в различных фильмах, получивших этот фильм благодаря своим легендарным песням,		1		
<input type="checkbox"/>	5	06 Apr 2024, 00:1	1	0	0	Аль	После серии глупых, веселых фильмов, 1955 год стал большим годом для Дорис Дэй. В том же году ей досталась роль певицы Рут Эттинг, которая добилась успеха, но чья личная жизнь пострадала из-за ее		1		
<input type="checkbox"/>	6	06 Apr 2024, 00:1	1	0	0	Аль	Это не совсем мюзикл, но он кажется таковым, потому что в нем больше пения, чем во многом другом, но это, очевидно, один из крючков. Однажды в баре, где она работала, Рут Эттинг (Дорис Дэй)		1		
<input type="checkbox"/>	7	06 Apr 2024, 00:1	1	0	0	Аль	После семи лет и семнадцати картин в Warner Brothers, Дорис Дэй перешла в MGM, чтобы сыграть главную роль в драматическом мюзикле, основанном на жизни поющей звезды 30-х годов Рут		1		
<input type="checkbox"/>	8	06 Apr 2024, 00:1	1	0	0	Аль	В 1950-х годах было много фильмов с фильмами: «Для меня запомнились 4»: «Прерванная мелодия» с Элеанор Паркер / «Великий Карузо» с Марио Ланца / «С песней в моем сердце» со Сьюзен		1		
<input type="checkbox"/>	9	06 Apr 2024, 00:1	1	0	0	Аль	МОЯ ОЦЕНКА- 7.3 Это любопытный ход, сделанный, когда Джеймс Кэгни уже оставил свои криминальные ходы. В нем также снялась популярная Дорис Дэй в роли певицы, которую он спонсирует. Кэгни		1		
<input type="checkbox"/>	10	06 Apr 2024, 00:1	1	0	0	Аль	Дорис Дэй и Джеймс Кэгни великолепны в этой пышной истории Technicolor, в которой подробно рассказывается о карьере певицы Рут Эттинг. Яркая, жизнерадостная, оптимистичная, счастливая		1		
<input type="checkbox"/>	11	06 Apr 2024, 00:1	1	0	0	Аль	«Люби меня или оставь меня» — это не типичный мюзикл. То есть никто не разражается песнями и плясками. Цифры используются в контексте истории либо на репетиции, либо в спектакле. Это		1		
<input type="checkbox"/>	12	06 Apr 2024, 00:2	1	0	0	Аль	Должен сказать, что когда я увидел такое название, как Zombiageddon и атомную бомбу на обложке, я ожидал увидеть откровенный фанг-ку, но вместо этого я получил комедию. В общем,		0		
<input type="checkbox"/>	13	06 Apr 2024, 00:2	1	0	0	Аль	Этот фильм рассказывает историю певицы 20-х годов Рут Эттинг и ее брака с Марти Снайдером. Когда Дорис поет старые стандарты, это здорово, но MGM добавили два номера, которые звучали нормально,		1		

Результат можно экспортировать в json, csv или tsv.

## 15. выводы

Прежде всего, хочу отметить монотонную сложность выставления лейблов вручную.

Каждый отзыв нужно прочитать, осмыслить и вынести свой вердикт. Может уйти несколько минут на обработку одного отзыва.

В результате, некоторые мои лейблы, отличались от лейблов выставленных машиной.

После изучения таких разногласий, перечитывал еще раз конкретный отзыв и к своему ужасу осознавал, что машина права! Т.е. человек хуже справляется с задачами разметки эмоциональной окраски, в данном случае, не говоря уже о скорости.

По грубым расчетам, для маркировки 50 тысяч записей (по 5 минут на каждую) может понадобиться 4167 человеко-часов работы или 173 суток. Если учесть восьми часовой рабочий день, два выходных дня в недели и отпуск раз в году, то специалист за год сможет осилить 2080 отзывов. Т.е. на конкретный большой датасет из нашего примера у человека уйдет два года. Машина это сделала за несколько секунд.

В результате те, кто сделал ставку на машинное обучение и нейросети 15 лет назад, стали очень богатыми людьми. Например, как автор данного набора [Эндрю Маас](#), попавший в журнал Форбс.