Для начала мы объединим данные для их очистки.

```
Python 3.8.10 (default, Nov 26 2021, 20:14:08)
[GCC 9.3.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> import numpy as np
>>> import pandas as pd
>>> df1 = pd.read_csv('train_tweets.csv')
>>> df2 = pd.read_csv('test_tweets.csv')
>>> df1.head(), df2.head()
(   id  label                                              tweet
0   1      0   @user when a father is dysfunctional and is s...
1   2      0  @user @user thanks for #lyft credit i can't us...
2   3      0                                bihday your majesty
3   4      0  #model   i love u take with u all the time in ...
4   5      0             factsguide: society now    #motivation,       id                                              tweet
0  31963  #studiolife #aislife #requires #passion #dedic...
1  31964   @user #white #supremacists want everyone to s...
2  31965  safe ways to heal your #acne!!    #altwaystohe...
3  31966  is the hp and the cursed child book up for res...
4  31967    3rd #bihday to my amazing, hilarious #nephew...)
>>> df1.dtypes, df2.dtypes
(id        int64
label     int64
tweet    object
dtype: object, id        int64
tweet    object
dtype: object)
>>> df1.info(), df2.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 31962 entries, 0 to 31961
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   id      31962 non-null  int64
 1   label   31962 non-null  int64
 2   tweet   31962 non-null  object
dtypes: int64(2), object(1)
memory usage: 749.2+ KB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17197 entries, 0 to 17196
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   id      17197 non-null  int64
 1   tweet   17197 non-null  object
dtypes: int64(1), object(1)
memory usage: 268.8+ KB
(None, None)
>>> df = pd.concat([df1, df2], ignore_index=True)
>>> df.to_csv('data.csv')
>>>
```

Задание 1.
Исходные данные:
Заменим html-сущности (к примеру: < > &). "<" заменим на “<” и "&" заменим на “&”)""".
Сделаем это с помощью HTMLParser.unescape(). Всю предобработку делаем в новом столбце 'clean_tweet'.
Решение:

```
Python 3.8.10 (default, Nov 26 2021, 20:14:08)
[GCC 9.3.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> import re
>>> import html
>>> import numpy as np
>>> import pandas as pd
>>> from html.parser import HTMLParser
>>> from html import unescape
>>> df = pd.read_csv('data1.csv')
>>> df['clean_tweet']
0         @user when a father is dysfunctional and is s...
1        @user @user thanks for #lyft credit i can't us...
2                                      bihday your majesty
3        #model   i love u take with u all the time in ...
4                   factsguide: society now    #motivation
                               ...
49154    thought factory: left-right polarisation! #tru...
4915
>>> my_string = df['clean_tweet']
>>> def html_decode(s):
...     htmlCodes = (
...             ("'", '&#39;'),
...             ('"', '&quot;'),
...             ('>', '&gt;'),
...             ('<', '&lt;'),
...             ('&', '&amp;')
...     for code in htmlCodes:
...         s = s.replace(code[1], code[0])
...     return s
...
>>> unescaped = html_decode(my_string)
>>> print(unescaped)
0         @user when a father is dysfunctional and is s...
1        @user @user thanks for #lyft credit i can't us...
2                                      bihday your majesty
3        #model   i love u take with u all the time in ...
4                   factsguide: society now    #motivation
                               ...
49154    thought factory: left-right polarisation! #tru...
4915
>>> df.to_csv('data1.csv')
>>>
```