—

November 17, 2025

[1]:
```
# !pip install librosa
```

[2]:
```
from google.colab import drive
import pandas as pd
import json


#
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call
drive.mount("/content/drive", force_remount=True).
```

[3]:
```
path_audio="/content/drive/MyDrive/     /2025_  _     1/data/TS_in_1"
```

[4]:
```
path_audio
```

[4]:
```
'/content/drive/MyDrive/     /2025_  _     1/data/TS_in_1'
```

[5]:
```
path_text="/content/drive/MyDrive/     /2025_  _     1/data/TL_out"
```

[6]:
```
label=pd.read_csv('/content/drive/MyDrive/     /2025_  _     1/data/emotions.
↪csv')
```

0.1 text(json)

[ ]:
```

```

[7]:
```
import os

all_files_json=os.listdir(path_text)
```

[8]:
```
len(all_files_json)
```

[8]:
```
2876
```

[9]:
```
#
# type A
#   start
#   last


import json

all_data=[]

for i in range(len(all_files_json)):

    file_name=os.path.join(path_text,all_files_json[i])

    with open(file_name, 'r', encoding='utf-8') as f:
        data=json.load(f)

        # print(data['info']['ID'])
        # print(data['list'])
        # print(len(data['list']))



        temp={}
        text=[]
        start=[]
        end=[]

        for p in range(len(data['list'])):
            for q in range(len(data['list'][p]['list'])):


                data_lst=data['list'][p]['list'][q]


                if 'audio' in data_lst:

                    for i in range(len(data_lst['audio'])):
                        if data_lst['audio'][i]['type']=='A':
                            start.append(data_lst['audio'][i]['start'])
                            end.append(data_lst['audio'][i]['end'])
                            text.append(data_lst['audio'][i]['text'])


                else:
                    text.append('')
```

```
        temp['id']=data['info']['ID']
        temp['start']=start
        temp['end']=end
        temp['text']=text

    all_data.append(temp)
```

[10]:
```
text_data=pd.DataFrame(all_data)
```

[11]:
```
text_data['text']=text_data['text'].apply(lambda x: ' '.join(x))
```

[12]:
```
text_data[text_data['id']=='0002']['text']
```

[12]:
```
2067    .   .    .    .    . …
Name: text, dtype: object
```

0.2 audio

[13]:
```
import os

all_files=os.listdir(path_audio)

audio_data={}
mp3_files=[]
id_lst=[]

for f in all_files:
    if f.endswith('.mp3'):
        mp3_files.append(f)

        id=f.replace('.mp3',"")
        id_lst.append(id)

audio_data['id']=id_lst
audio_data['mp3_files']=mp3_files

audio_data=pd.DataFrame(audio_data)
```

0.3 label

[14]:
```
label['emotion'] = label['emotion'].replace({
    '  ': ' ',
    '  ': ' ',
    '   ': ' ',
    '  ': ' ',
    '  ': ' '
})
```

[15]:
```
label['emotion'].value_counts()
```

[15]:
```
emotion
     1450
      659
      561
       94
       89
       23
Name: count, dtype: int64
```

[16]:
```
label.head()
```

[16]:

	filename	emotion
0	0002.json	
1	0004.json	
2	0005.json	
3	0006.json	
4	0007.json	

[17]:
```
label['id']=label['filename'].apply(lambda x: x.replace('.json',''))
```

[18]:
```
label=label[['emotion','id']]
```

[19]:
```
label
```

[19]:

	emotion	id
0		0002
1		0004
2		0005
3		0006
4		0007
…	…	…
2871		5194
2872		5196
2873		5247
2874		5248
2875		5250

[2876 rows x 2 columns]

0.3.1 join// concate stt

[20]:
```
concat_data=pd.merge(text_data, audio_data, on='id', how='right')
```

[21]:
```
concat_data=pd.merge(label, concat_data, on='id', how='right')
```

[22]:
```
concat_data
```

[22]:

	emotion	id	start	\
0		0058	[00:03.130, 00:09.240, 00:14.730, 00:21.780, 0…	
1		0017	[00:08.142, 00:22.553, 00:35.999, 00:52.070, 0…	
2		0028	[00:04.270, 00:09.810, 00:15.210, 00:22.710, 0…	
3		0034	[00:02.128, 00:06.029, 00:10.803, 00:18.112, 0…	
4		0067	[00:04.647, 00:15.760, 00:23.129, 00:29.440, 0…	
..	…	…	…	
715		1117	[00:12.328, 00:29.638, 00:46.652, 01:05.778, 0…	
716		1118	[00:11.107, 00:28.367, 00:45.852, 01:07.765, 0…	
717		1116	[00:12.842, 00:26.632, 00:43.474, 01:00.737, 0…	
718		1112	[00:10.207, 00:18.175, 00:25.544, 00:34.495, 0…	
719		1105	[00:05.688, 00:13.635, 00:20.743, 00:27.380, 0…	

	end	\
0	[00:05.330, 00:11.600, 00:18.340, 00:24.450, 0…	
1	[00:12.321, 00:25.874, 00:41.517, 00:58.070, 0…	
2	[00:06.120, 00:11.750, 00:17.500, 00:25.670, 0…	
3	[00:03.655, 00:07.913, 00:13.889, 00:22.205, 0…	
4	[00:11.281, 00:18.440, 00:25.456, 00:33.495, 0…	
..	…	
715	[00:14.566, 00:32.551, 00:53.070, 01:08.818, 0…	
716	[00:15.159, 00:31.819, 00:52.606, 01:16.545, 0…	
717	[00:17.579, 00:34.948, 00:51.685, 01:04.737, 0…	
718	[00:13.372, 00:20.631, 00:29.419, 00:45.684, 0…	
719	[00:06.672, 00:14.631, 00:21.401, 00:28.218, 0…	

	text	mp3_files
0	…	0058.mp3
1	…	0017.mp3
2	…	0028.mp3
3	…	0034.mp3
4	…	0067.mp3
..	…	…
715	…	1117.mp3
716	…	1118.mp3
717	…	1116.mp3
718	…	1112.mp3

```
719   .   .      .   .     .         …  1105.mp3

[720 rows x 6 columns]
```

[23]:
```
concat_data['text'][0]
```

[23]:
```
'  .      .         .       .          .          .
     .    .             .   .    ,      .    .     .    .
          .     ,   .       .        .     .          . .
                          .          ,       .    .   .
     .   .  .          .      .      .   .      .
   .      .   .      .   .      .   .   .   .   .   .       .
   .      .   .'
```

[24]:
```
concat_data=concat_data.drop(['id'], axis=1)
```

[25]:
```
concat_data['label']=concat_data['emotion']
```

[26]:
```
concat_data['mp3_files'][0]
```

[26]:
```
'0058.mp3'
```

[27]:
```
import librosa
import numpy as np
from tqdm import tqdm
import pandas as pd


n_mfcc = 30 #         ? #      30~40     .
```

[28]:
```
concat_data['mp3_files']
```

[28]:
```
0      0058.mp3
1      0017.mp3
2      0028.mp3
3      0034.mp3
4      0067.mp3
         …
715    1117.mp3
716    1118.mp3
717    1116.mp3
718    1112.mp3
719    1105.mp3
Name: mp3_files, Length: 720, dtype: object
```

[29]:
```
mfcc_lst=[]
id_lst=[]
```

```
for idx in range(len(concat_data)):


    id=concat_data['mp3_files'][idx].replace(".mp3","")

    start_list = concat_data['start'][idx]
    end_list = concat_data['end'][idx]


    voc_path = os.path.join(path_audio, concat_data['mp3_files'][idx])

    data, sr = librosa.load(voc_path, sr = 16000) #        16000
 ↪      .


    #
    audio_segments = []

    for start_time, end_time in zip(start_list, end_list):

        if '--' in str(start_time) or '--' in str(end_time):
            continue

        try:

            start_min, start_sec = start_time.split(":")
            start_int=float(start_min)*60 + float(start_sec)

            end_min, end_sec = end_time.split(":")
            end_int=float(end_min)*60 + float(end_sec)



            # # "start" ~ "end"
            start_sr=int(start_int*sr)
            end_sr=int(end_int*sr)

            segment =data[start_sr:end_sr]

            if len(segment) > 0:
                audio_segments.append(segment)

        except Exception as e:
            print(f"  idx={idx}   : {e}")
            continue
```

```
    #
    if len(audio_segments) == 0:
        print(f"  idx={idx} ({id})    -       ")
        skipped_count += 1
        continue


    #    :
    data = np.concatenate(audio_segments)


    #    (    )
    data_trim,temp=librosa.effects.trim(data, top_db=20)
    #print(data_trim)


    #        (     (trim) / (  RMS +epsilon) ) *   RMS
    rms=librosa.feature.rms(y=data_trim)

    mean_rms=np.mean(rms)
    epsilon=1e-10 # rms  0    0
    target_rms=0.05  # 0.01 ~ 0.1
    data_normalized=(data_trim/(mean_rms+epsilon) * target_rms)


    # mfcc
    mfcc = librosa.feature.mfcc(y=data_normalized, sr=sr, n_mfcc=n_mfcc)

    mfcc_lst.append(mfcc)
    id_lst.append(id)



###     id    join      ? ###
concat_data['mfcc']=mfcc_lst
concat_data['id']=id_lst
```

[30]:
```
concat_data['mfcc'][1]
```

[30]:
```
array([[-2.57824860e+02, -1.63186356e+02, -1.31970291e+02, …,
        -2.23640320e+02, -2.90793884e+02, -3.93042725e+02],
       [ 8.60681763e+01,  1.16179764e+02,  1.15459427e+02, …,
         1.90170593e+02,  1.79603119e+02,  1.24511749e+02],
       [-6.54679260e+01, -6.71483307e+01, -6.86114960e+01, …,
        -1.18421650e+01, -1.67629218e+00,  9.33983803e+00],
       …,
```

```
[ 1.15865841e+01,  1.26029072e+01, -4.86166716e+00, …,
 -1.26275480e+00,  9.80686128e-01,  1.79550266e+00],
[-3.97589755e+00, -9.26762962e+00, -1.00750065e+01, …,
  4.33649063e+00,  1.45739985e+00,  4.10671425e+00],
[ 6.73126316e+00, -2.64116377e-01, -2.82701969e+00, …,
 -5.44367611e-01,  1.10771580e+01,  5.86241627e+00]], dtype=float32)
```

[31]:
```
concat_data_1=concat_data
```

[32]:
```
# id    DataFrame
concat_data_1 = concat_data_1.sort_values('id').reset_index(drop=True)
```

[33]:
```
#concat_data_1.to_csv('    _1')
```