dataset_exploration

July 3, 2020

0.1 LibriSpeech Structure

The LibriSpeech corpus is a standard benchmark for recent speech related neural network and machine learning research << CITE >>. It contains various sets of speech samples (`urls`) of readers reading from texts contained within the Project Gutenberg corpus of books. The speech samples are saved within speaker specific directories in the `.flac` file format. The follow `urls` are contained within it: * `dev-clean` * `dev-other` * `test-clean` * `test-other` * `train-clean-100` (100 hours) * `train-clean-360` (360 hours) * `train-other-500` (500 hours)

All of the audio has a sampling rate of 16kHz.

```
In [ ]: import os
```

```
In [54]: # change this to wherever it lives on your machine
         DATA_DIR = '/home/thomas/Dir/ccny/ccny-masters-thesis/raw-data/LibriSpeech'
```

The LibriSpeech dataset (thankfully) has a standardized structure of its directories:

```
|- <url>
   |-- <speaker_id>
       |-- <chapter_id>
           |-- <speaker_id>-<chapter_id>-<snippet_id>.flac
           .
           .
           .
           |-- <speaker_id>-<chapter_id>.trans.txt
```

This makes it useful for most speech-related tasks, be it speech recognition or speaker recognition / verfication.

```
In [55]: os.listdir(DATA_DIR)
```

```
Out[55]: ['CHAPTERS.TXT',
          'dev-clean',
          'BOOKS.TXT',
          'README.TXT',
          'LICENSE.TXT',
          'test-clean',
          'SPEAKERS.TXT']
```

Let's investigate on sample file in particular to just to get a sense of how we would interact with it.

```
In [56]: url = 'dev-clean'

In [57]: speakers = os.listdir(os.path.join(DATA_DIR, url))
         speakers[0]

Out[57]: '6345'

In [58]: chapters = os.listdir(os.path.join(DATA_DIR, url, speakers[0]))
         chapters

Out[58]: ['93302', '64257', '93306']

In [59]: files = os.listdir(os.path.join(DATA_DIR, url, speakers[0], chapters[0]))
         files[0]

Out[59]: '6345-93302-0013.flac'
```

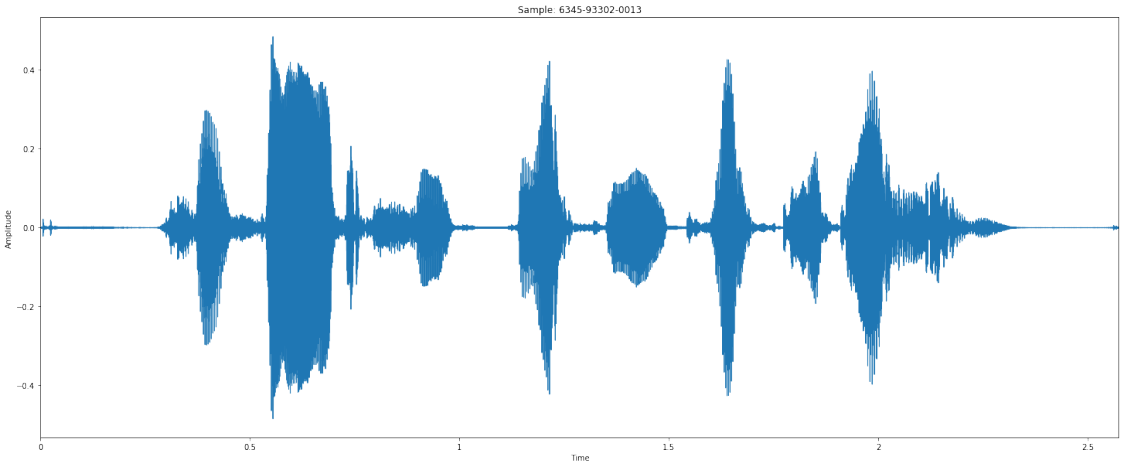
0.2 Audio Sample

Here we'll load and display the audio file using the `librosa` library which provides many helpful utilities for dealing with audio files. Something worth noting here is that the default sample rate for `librosa` is typically 22050Hz.

```
In [60]: import librosa
         import librosa.display
         import matplotlib.pyplot as plt
         import IPython

In [61]: # pick an example file
         sample_file_path = os.path.join(DATA_DIR, url, speakers[0], chapters[0], files[0])
         sample_file_path

Out[61]: '/home/thomas/Dir/ccny/ccny-masters-thesis/raw-data/LibriSpeech/dev-clean/6345/93302/6

In [62]: sample = os.path.basename(sample_file_path).strip('.flac')
         sample

Out[62]: '6345-93302-0013'

In [63]: librispeech_sr = 16000
         y, _ = librosa.core.load(sample_file_path, sr=librispeech_sr)
```

Now given that raw audio in waveform, we can visualize the amplitude (?) of the wave over the time of the sample.

```
In [80]: # TODO: add axis label
         plt.figure(figsize=(25,10))
         plt.title(f'Sample: {sample}')
         plt.ylabel('Amplitude')
         _ = librosa.display.waveplot(y, sr=librispeech_sr)
```



What does the sample actually sound like?

```
In [68]: IPython.display.Audio(y, rate=librispeech_sr)
```

```
Out[68]: <IPython.lib.display.Audio object>
```

And, as mentioned, LibriSpeech also provides the transcript of the audio file, which will come in handy for performing speech recognition.

```
In [69]: transcript = [
             f for f in os.listdir(os.path.join(DATA_DIR, url, speakers[0], chapters[0])) if f
         ][0]
         transcript
```

```
Out[69]: '6345-93302.trans.txt'
```

```
In [46]: # all the snippet transcripts are in the same file with their labels at the beginning
         # of each line
         with open(os.path.join(DATA_DIR, url, speakers[0], chapters[0], transcript), 'r') as
             for line in f:
                 if sample in line:
                     print(line)
```

```
6345-93302-0013 SHE SAID HOW FRIGHTFULLY COLD IT IS
```