TUH Seizure Corpus (TUSZ) Preprocessing Report

Some statistics:

Number of seizures in validation set- 673
Number of seizures in training set- 2377
Total number of seizures- 3050

Seizure duration in the validation set- 58445 seconds
Seizure duration in the training set- 169794 seconds
Total seizure duration in seconds- 228239 seconds

Number of unique patients with seizures in validation set- 40
Number of unique patients with seizures in training set- 202
Total number of unique patients with seizures- 242

TUH_Preprocessing.py builds and preprocesses the raw TUH seizure dataset (TUSZ version 1.5.2). The TUSZ dataset follows a large hierarchical Unix-style file tree structure.

Notes on downloading the TUH seizure dataset:
https://www.isip.piconepress.com/projects/tuh_eeg/html/downloads.shtml

Example rsync command to download seizure corpus: rsync -auxvL nedc@www.isip.piconepress.com:~/data/tuh_eeg_seizure/v1.5.2/ /home/dplatnick/TUHData

Note: The space is important between: /v1.5.2/ /home/dplatnick/TUHData

TUH_Preprocessing.py is a modification of IBM's python code to build and preprocess the TUH seizure corpus dataset.

IBM code can be found here: https://github.com/IBM/seizure-type-classification-tuh/blob/master/data_preparation/build_data.py

Getting TUH_Preprocessing.py to run:

This code expects the TUH dataset to have a slightly different file-tree structure compared to the one which is downloaded by the command:
rsync -auxvLnedc@www.isip.piconepress.com:~/data/tuh_eeg_seizure/v1.5.2/ /home/dplatnick/TUHData

To make this code work, you first must create a directory called v1.5.2 inside of the TUHData folder and move the _DOCS and edf folders inside, so it looks like this:

```
(dan_test_1) dplatnick@pilot:~/TUHData/v1.5.2$ pwd
/home/dplatnick/TUHData/v1.5.2
(dan_test_1) dplatnick@pilot:~/TUHData/v1.5.2$ l
_DOCS/  edf/
(dan_test_1) dplatnick@pilot:~/TUHData/v1.5.2$
```

Next, TUH_Preprocessing.py depends on the files seizures_v36r.csv, and parameters.csv. Make sure parameters.csv is in the same working directory as TUH_Preprocessing.py.

The file parameters.csv is found in the LIA GitHub repository. The file seizures_v36r.csv is located in the TUH Corpus in the _DOCS directory.

After finishing the above steps, adjust lines 290-298 of TUH_Preprocessing.py to include the appropriate file paths for the TUH Seizure Corpus, and an output directory to export the .pkl and .lbl contents.

Example:

```
289     if platform.system() == 'Linux':
290         parser.add_argument('--base_dir', default='/home/dplatnick/TUHData',
291                             help='path to raw seizure dataset')
292         parser.add_argument('--save_data_dir', default='/home/dplatnick/TUH_Output_test',
293                             help='path to save processed data')
294     elif platform.system() == 'Darwin':
295         parser.add_argument('--base_dir', default='/home/dplatnick/TUHData',
296                             help='path to raw seizure dataset')
297         parser.add_argument('--save_data_dir',
298                             default='/home/dplatnick/TUH_Output_test',
299                             help='path to save processed data')
```

Information about Signal Processing:

TUH_Preprocessing.py uses seizures_v36r.csv to parse and process the TUH dataset.

For every distinct seizure, TUH_Preprocessing.py outputs 8 pieces of information into a **.pkl** file, and also outputs a corresponding **.lbl** annotation file.

A given **.pkl** output file contains a named tuple which holds 8 pieces of information:
1. Patient ID
2. Seizure class label
3. Seizure start time (raw signal)
4. Seizure end time (raw signal)
5. Processed signal (10x seizure duration before and after)
6. Seizure start time (processed signal)
7. Seizure end time (processed signal)
8. Original sample frequency.

Patient ID is the unique identifier of each patient.

Seizure class label refers to the annotated type of seizure, from the set of seizure types:
1. Focal Non-Specific Seizure (FNSZ)
2. Generalized Non-Specific Seizure (GNSZ)
3. Complex Partial Seizure (CPSZ)
4. Absence Seizure (ABSZ)
5. Tonic Seizure (TNSZ)
6. Tonic Clonic Seizure (TCSZ)
7. Simple Partial Seizure (SPSZ)
8. Myoclonic Seizure (MYSZ)

Seizure start time (raw signal) refers to the seizure start time (in seconds) of the raw EEG signal.
Seizure end time (raw signal) refers to the seizure end time (in seconds) of the raw EEG signal.

Processed signal is the new EEG signal which is processed to contain a signal with length equivalent to 10 times the seizure duration before and after the seizure event (If enough signal is available).

Seizure start time (processed signal) refers to the seizure start time (in seconds) of the processed EEG signal.
Seizure end time (processed signal) refers to the seizure end time (in seconds) of the processed EEG signal.
Original sample frequency is the original sampling rate of the raw EEG signal.

Reading the .pkl and .lbl File Contents:

A script called Unpickle_TUHData.py can be found on the LIA GitHub page. This script gives some example python code to read and access the **.pkl** and/or **.lbl** file contents for a given observation. It also shows some code to plot an EEG signal.

Paper explaining **.lbl** file contents (section 3.2):
https://www.isip.piconepress.com/projects/tuh_eeg/downloads/tuh_eeg_seizure/v1.5.1/_DOCS/02_annot.pdf