**Documentation**

This code seeks to apply a version of BERT called "ClinicalBERT" to data that originates from Indiana chest X-ray dataset. ClinicalBERT was initially developed by Kexin Huang, Jaan Altosaar, Rajesh Ranganath as outlined in their article "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission", and was developed by training BERT on MIMIC III hospital readmission data. The code in this project follows that described in an article by Nwamake Imasogie, where she applies a similar analysis to 2-day and 3-day clinical notes, instead of the discharge notes used in the original article.

**Code**

The code for this repo is contained in two files.
prep_caml_data.ipynb → for preprocessing the code an conducting classical NLP ML analysis
CAMLBERT_final.ipynb → for running ClinicalBERT on the data

There are flags at the top of the code for switching between Mac and Colab platforms, though Mac settings have not been tested in a while and may not work.

The original code provided by Nwamake Imasogie can be found at this repo (obtained after contacting her … the links in her original articles are broken). The NLP analysis is different rom that conducted by the author, and can be considered to be **"additional modeling" (= ablation study).**

**CAML Data Location**

The chest X-ray data provided for this project is contained in this repo (clean_caml.csv). The original train / test splits are also provided for comparison (caml_train.csv and caml_test.csv). The "binary polarity reversed" version of clean_caml.csv is rev_clean_caml.csv, and is kept in the top level directory.

**Original Papers**

This project seeks to apply the general approach taken by Nwamake Imasogie in her paper ClinicalBERT: Using a Deep Learning Transformer Model to Predict Hospital Readmission. The code repo is available here and a related article on how she preprocessed the data and applied classical NLP techniques and ML analysis can be found at Predicting Hospital Readmission Using NLP. The link to the code base for that article is broken, and the author did not have time to send it to me (she is on maternity leave), but promised to send it at a future date.

**Dependencies**

Python 3.7 or greater

transformers

pytorch

**Data Download Instructions**

**BERT Model and Code**

The pretrained ClinicalBERT Model parameters (bert_config.json) are provided under the Model folder, but the actual BERT bin file (pytorch_model.bin) is too large to store in the repo and needs to be uploaded from https://drive.google.com/drive/folders/1X_oOiKWE5WRebNDAyniafuycQZYkQYXu. The file structure needs to be property maintained, as suggested by the file paths at the top of the Jupyter Notebook CAMLBERT_final.ipynb.

| File | Location |
|---|---|
| file_utils.py | top level directory |
| modeling_readmission.py | top level directory |
| model (folder) | top level directory |

| File | Location |
| --- | --- |
| pytorch_model.bin | in the model folder |
| bert_config.json | in the model folder |