Documentations

The goal with Healthcare HRCT is to find the clinically relevant terms in documents to help identify the patients that are likely to end up with a particular diagnosis. The hope is that by identifying the patient's primary issue, treatment can be carried out earlier resulting in more positive outcomes, especially where a patient only has a small timeframe to get treatment to result in a positive outcome such as a severe sepsis diagnosis.

The package within will allow you to build a model using labelled datasets (one with the diagnosis and one without) which are required and additionally you may choose to include a third dataset consisting of notes of the same type that will be used to designate a background distribution, but this third dataset is not required to build the model.

To import the model you can use the following import statement:

from hrct_model import HRCTmodel

As you might have guessed to build the model you with use the HRCTmodel you just imported. The only other items you need are 2-3 paths to directorys consisting of PDFs of clinical notes. This model does not currently support any file types other than PDF, but will ignore items that are not of the PDF type if they are in the same directory. Code to set up and train the model can be found below:

hm = HRCTmodel(positive_diagnosis_directory, negative_diagnosis_directory, *background_directory)
*Note background_directory is optional to include

Building the model will likely take several minutes. Your HRCTmodel object will have 2-3 distributions (depending on how many were used to train with), and exploration can be done to dive into word_count, total_words, word_dictionary, and word_distribution within each distribution. Finally, each model will have a scoring function for additional PDFs that were not used to train the model.

hm.score(directory_to_score, save_file=True)
This scoring function by default will score each document and then output a file of scored documents and which group they belong to 1 = positive diagnosis and 0 = negative diagnosis.

There is one pre-trained model available to predict a diagnosis of severe sepsis provided in both windows () and linux/macos () versions. To import this model, the following code can be used with the correct version of the model for your system.

import pickle
from sepsis_model import sepsisHRCTmodel
with open ('sepsis_model_linux.pickle', 'rb') as handle:
    sepsis = pickle.load(handle)

Errors/Issues Identified:

Occassionally when running this code in Spyder, the working directory changed from the location the file was opened in to the /home directory for the system. To resolve this, you can simply import os and change your working directory before importing the HRCTmodel:

import os
os.chdir('/Users/Cristina/Downloads/sepsis_model_linux/')
from sepsis_model import sepsisHRCTmodel
Please note, while all that is required to run this project is to download all of the items into a single place on your machine, there are some packages that are required to allow this model to successfully run.

Required python packages:

- os
- pandas
- spacy==3.4.3
- pdfminer.six
- pickle ** only required to open pre-trained model