Cohere Health Machine Learning Takehome Assignment

Cohere's mission is to reduce complexity and waste in clinical care. Often the most valuable insights are hidden among clinical notes and can be extracted through machine learning techniques. We'd like you to leverage your expertise and synthesize a map of clinical diagnoses and symptoms from the data provided.

Task Description

Using the data from sampleclinicalnotes.zip and machine learning techniques:

- Determine the primary medical diagnoses of a patient.
- Identify the common underlying factors for each diagnosis.

An example of primary diagnosis and underlying factors:
Patients that have heart failure usually develop Type 2 Diabetes, high blood pressure, and coronary artery disease first. In this case, the primary diagnosis is heart failure, and the underlying factors would be the ones listed above.

BONUS TASK (completely optional): Once you have constructed this "database", convert it into a diagnosis tool. The end user of this diagnosis tool should be able to input a patient's clinical note and receive an output of the patient's primary diagnosis (i.e. heart failure) as well as the symptoms in their clinical history that support this primary diagnosis (i.e. type 2 diabetes).

This bonus task is completely optional and will not reflect negatively on you if you choose not to participate. We include it here in an attempt to provide an example of how such a task might be used as a product in our field.

Dataset Overview

You have .txt and .ann files in sampleclinicalnotes.zip.

- **.txt:** Contains sections like 'Discharge Diagnosis', 'Chief Complaint', and 'History of Present Illness'. You can focus on these sections to find helpful information about medical conditions and associated symptoms.
- **.ann:** Represents entity recognition outputs, indicating the type of medical entity in the corresponding .txt file. The included columns are: ['group', 'category', 'start_idx', 'end_idx', 'text']. The 'start_idx' and 'end_idx' indicate the character positions of the specific entity in the corresponding text file. You can choose to use only entities tagged as "Reason", since those would be the ones that map more closely to disease names.

(continued on next page)

Assume all possible medical conditions and underlying factors are within this dataset. You may also find it helpful to take a look at scispacy and their entitylinker.

Utilize the starter code in **load_data.py** and **cohere_ml_takehome_start.ipynb** to set up. You will still need to perform your own cleaning of the data.

Expectations

The goals of this exercise are:

- To evaluate your coding capabilities;
- To assess your problem solving skills;
- For you to get a taste of what it is like to work with us and our data.

Your next steps should be:

1. Review the task summary and data. Ensure that you are able to download the given dataset.
2. Implement your solution and return:
    a. functioning code in a zip file, with instructions on how to run and how to retrieve the underlying factors for a given condition;
    b. a final report as a PDF explaining the problem, solution and analysis of the results. If you chose to participate in the bonus task, please note that in your report.

Your solution will be evaluated holistically on the following criteria:

- Demonstration of an understanding of the problem and the dataset;
- Code clarity & quality
- How closely it achieves the goals laid out above (Does your solution identify the patient's primary medical diagnosis given a clinical note? Does it list out the common underlying factors for each diagnosis?)

Good luck, and please don't hesitate to reach out with questions!