Problem statement

You are working on a project that aims to extract information about patient's condition from visit notes collected in the real-world setting.

The notes contain a variety of information including demographics, diagnosis, medical history, medication/treatments, etc. For example, the sentence
"Ms. A is a 60-year-old African-American female with 12 years of education who was referred for neuropsychological evaluation by Dr. X after she demonstrated mild cognitive deficits on a neuropsychological screening evaluation during a follow up appointment with him for stroke in July. Neurological evaluation with Dr. X confirmed left hemiparesis."
contains information about patient's demographics (60 years old, female, African-American), socioeconomic factor (12 years of education), and medical conditions (stroke, mild cognitive deficits, left hemiparesis).

In order to extract all the information, the first step is to figure out how to generate vector representations of the texts in a way that can capture the clinical meanings. You have a few approaches in mind, and the plan is to assess which approach is most suitable for the project by making use of two pieces of information (attached):

1. A list of **related** medical entities pairs
   "MedicalConcepts.csv" contains a list of related medical entities pairs (Term1 and Term2 are related to each other) that has been curated by domain experts

2. Sample notes collected from patient visits
   "ClinNotes.csv" contains notes (column='notes') from three departments (column='category'), i.e. Cardiovascular / Pulmonary, Neurology, and Gastroenterology

The goal of this task is to propose different approaches to generate vector representations of texts (tokens) and assess and compare the suitability of the approach for this dataset.