Data Science Take Home

The assignment will consist of a sample business problem with a dataset (patient_transcript.csv) provided. The project will be explained below.

Business Problem Assume you are working for a small consulting firm and the client is a medical research institution. The institution records patient data in a transcription file. The company has asked you to convert these notes to tabular format for ingestion in a SQL database.

Data understanding: The dataset consists of 2 columns, user_id and transcription. The transcription contains notes about each patient's illness or condition. There may be more than one transcription per patient. Sometimes this may not make much sense, as with multiple diagnoses, but this is because I procedurally generated parts of the data. Please don't get caught up on this aspect.

For this assignment, I would like you to perform named entity recognition on the dataset and determine which column headers will be in your SQL table. You may use any ML technique you like in any language. This is the only coding / engineering work I would like you to do. The rest of this assignment will be for me to understand the way you would approach this in a business setting. Be prepared to discuss how you made this decision. What entities did you keep? Why? What are some additional considerations you would make?

In real life, this would be a cross-functional effort involving engineers, PMs, and data scientists, among others. Next steps would likely involve working with engineering to create a SQL table. Assuming this will be performed in a cloud architecture (AWS, Azure, GCP, or similar), what are some of the steps that would need to be performed next? One step will be to assign values to the rows in the dataset. Think about how you will approach this. What are some of the pitfalls?

It's important to note there are no right or wrong answers here, as with most business problems. The data are messy. You must make choices. If you include all named entities you will end up with potentially hundreds of columns and a very sparsely populated dataset. If you narrow it down, how will you make that decision? What issues might arise in assigning data to the columns? How will you deal with this?

I'm not expecting the impossible. Please spend no more than 3 hours on this. In that time you should be able to run the model and think about the business aspects outlined above. Feel free to reach out during the process if you have any questions.