IQVia assignment

NLP / MACHINE LEARNING ENGINEER

Delivery format

Solution has to be delivered as a Github repository, where you share all Python resources developed to solve the assignment.

Assignment

Much clinical information is recorded in free-text medical reports which usually contain a large number of abbreviations. Some of these abbreviations may be ambiguous therefore it is crucial for an EHR parser to be able to infer to the correct expansion from the context.

The assignment is to write a production-ready solution to disambiguate acronyms using machine learning.

- Given a text containing acronyms, the code has to determine for each acronym which of its possible expansions it refers to.
- Provide the code for training and for inference
- It has to work for the acronyms provided in the toy dataset (test_set.csv), but the solution has to be prepared to allow the extension to a much larger set of acronyms. This would require retraining of course, but no change in the code should be required.

The model is not expected to have a high accuracy, we are more interested if you can model the problem in a sensible way and deliver a production-ready approach to solve a real problem.

Code quality

We don't expect you spend a big quantity of hours on the assignment, but we do want to see a running train pipeline capable of being executed with a function call and no further human interaction. For the inference step, the preferred solution would to expose the predict function through an API endpoint (doesn't need to be deployed anywhere, but kudos if you do!)

Data

Two files have been made available which is to be used to complete the assignment:

- *test_set.csv* contains abbreviations with possible expansions with some example sentences. Since there are multiple examples per acronym-expansion combination it can be used as a small test set. It is a pipe separated file, and the explanation of the columns is as follows:
    - Acronym: The acronym (will appear multiple times because there are multiple expansions)
    - Expansion: The expansion of the acronym
    - Type: The conceptual category that the expansion belongs to (this is not needed for the NLP task, but is part of the dataset as we have it)
    - Sample: An example of the acronym in a typical sentence
- *corpus.txt* contains unannotated sentences in which the above terms appear. They can be considered an equivalent of the textual data that you would have available in a hospital to work with. In a real setting this data would be in the order of millions. Preprocess this dataset according to your needs to use it for training, such as replacing expanded forms with their acronyms.