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 a EHR parser to be able to infer to the correct expansion from the context.

The assignment is to write a piece of code to disambiguate acronyms using machine learning.

* Given an acronym and its context, the code has to determine which of its possible expansions the acronym refers to.
* Provide the code for training and for inference
* It has to work for the acronyms provided in the toy dataset, but the solution has to be generic 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 production ready code to solve a real problem.

Data

--

You received two files

* *Coding\_test\_toy\_set\_acronyms.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 colums 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.
* *coding\_test\_sentence\_data.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.

Programming language:

--

Deliver the code in python, no jupyter notebook for instance.

Time investment

--

You are not expected to spend more than 4 hours on the assignment.

----------