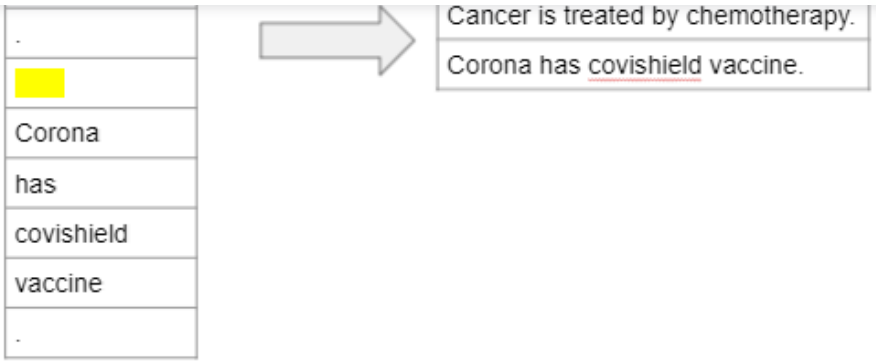
Expected Tasks

There are eight major tasks that you need to perform to complete the assignment. They are as follows:

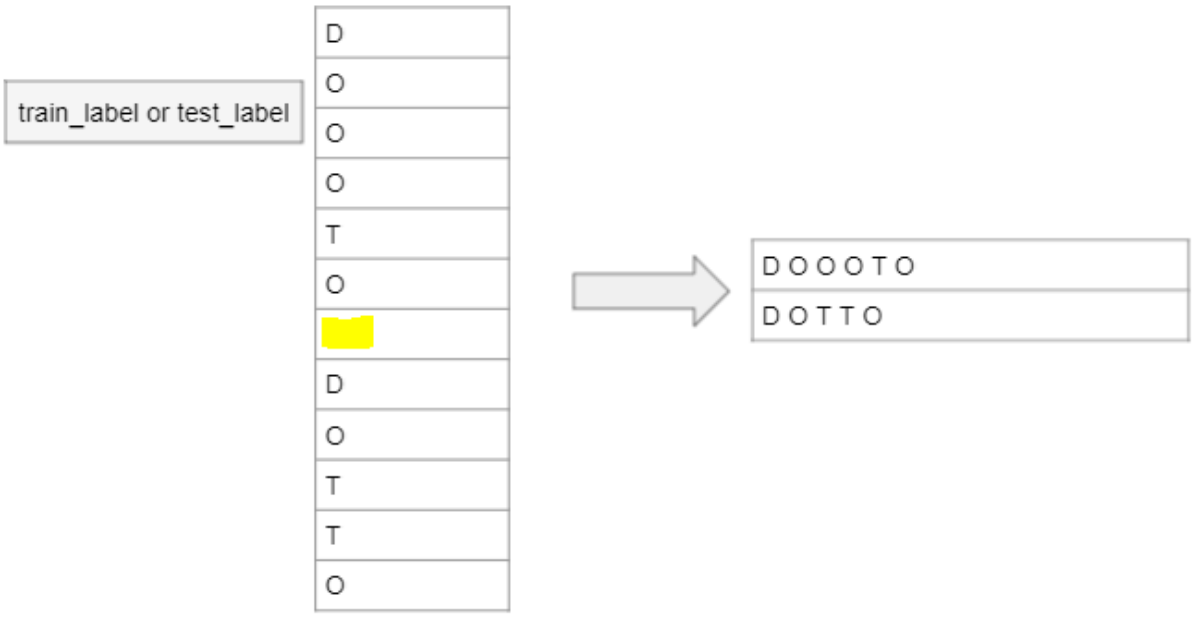
1. Data preprocessing
2. Concept identification
3. Defining the features for CRF
4. Getting the features words and sentences
5. Defining input and target variables
6. Building the model
7. Evaluating the model
8. Identifying the diseases and predicted treatment using a custom NER

Let's break down the steps into subtasks to understand this better.

Data preprocessing: As you are already aware that the dataset is in the token format instead of sentences, you need to construct the sentences from the words. There are blank lines after the completion of each sentence or a set of labels in label files ('train_label' and 'test_label') and you need to build a logic to arrange them into sentences or a sequence of labels in the case of label files. You can refer to the following two images to understand better.



A similar step is to be performed for the 'train_label' and 'test_label' datasets.



You need to do the following three tasks after processing and modifying the datasets:

- Construct proper sentences from individual words and print five sentences along with their labels.
- Print the correct count of the number of sentences in the processed train and test dataset.
- Correctly count the number of lines of labels in the processed train and test dataset.

dictionary of their counts. We will then output the top 25 most frequently discussed concepts in the entire corpus.

An important point to note here is that we are using both test and train sentences for concept identification. This is an exploratory analysis on the complete data. In this step, you need to perform the following two tasks by considering the train and the test dataset as a single unit of data:

- Use a toolkit like spaCy to extract those tokens that have NOUN or PROPN as their PoS tag and find their frequency from the **entire dataset** that comprises both the train and the test datasets.
- Print the top 25 most common tokens with NOUN or PROPN PoS tags for the **entire dataset** that comprises both the train and the test datasets.

Defining the features for CRF: Here, you need to perform the following three steps:

1. Define the features with the PoS tag as one of the features.
2. While defining the features in which you have used the PoS tags, you also need to consider the preceding word of the current word. The use of the information of the preceding word makes the CRF model more accurate and exhaustive.
3. Mark the beginning and the end words of a sentence correctly in the form of features.

Getting the features and the labels of sentences: In this step, you need to perform the following two tasks:

Defining input and target variables: In this step, you need to perform the following two tasks:

- Extract the features' values for each sentence as an input variable for the CRF model in the test and the train dataset.
- Extract the labels as the target variable for the test and the train dataset.

Building the model: You need to build the CRF model for a custom NER application using the features and the target variables.

Evaluation: Evaluate the model using the following two steps:

- Predict the labels of each of the tokens in each sentence of the test dataset that has been preprocessed earlier.
- Calculate the f1 score using the actual and the predicted labels of the test dataset.

Identifying the diseases and treatment using a custom NER:

- Create the code or logic to get all the predicted treatments (T) labels corresponding to each disease (D) label in the test dataset. You can refer to the following image to get an idea on how to create a dictionary where diseases are working as keys and treatments are working as values.

- Predict the treatment for the disease named 'hereditary retinoblastoma'.

In this way, you will be able to finish this assignment. Let's download the well-commented notebook that you can refer to solve this assignment.

You have been given the data in the form of tokens instead of sentences, and you need to process the data to get the sentences.

Please note that here, we are assuming that if there is a disease in the sentences, then the treatment mentioned in that sentence can be assumed to be the treatment for that disease. Also, there is an assumption that the same treatment can work for different diseases.

The next segment will help you understand the evaluation scheme for the assignment.

Report an error

PREVIOUS
Problem Statement

NEXT
Evaluation Rubrics