Now, let’s consider a hypothetical example of a health tech company called ‘BeHealthy’. Suppose ‘BeHealthy’ aims to connect the medical communities with millions of patients across the country.

‘BeHealthy’ has a web platform that allows doctors to list their services and manage patient interactions and provides services for patients such as booking interactions with doctors and ordering medicines online. Here, doctors can easily organise appointments, track past medical records and provide e-prescriptions.

So, companies like ‘BeHealthy’ are providing medical services, prescriptions and online consultations and generating huge data day by day.

Let’s take a look at the following snippet of medical data that may be generated when a doctor is writing notes to his/her patient or as a review of a therapy that he or she has done.

“The patient was a 62-year-old man with squamous cell lung cancer, which was first successfully treated by a combination of radiation therapy and chemotherapy.”

As you can see in this text, a person with a non-medical background cannot understand the various medical terms. We have taken a simple sentence from a medical data set to understand the problem and where you can understand the terms ‘cancer’ and ‘chemotherapy’.

Suppose you have been given such a data set in which a lot of text is written related to the medical domain. As you can see in the dataset, there are a lot of diseases that can be mentioned in the entire dataset and their related treatments are also mentioned implicitly in the text, which you saw in the aforementioned example that the disease mentioned is cancer and its treatment can be identified as chemotherapy using the sentence.

But note that it is not explicitly mentioned in the dataset about the diseases and their treatment, but somehow, you can build an algorithm to map the diseases and their respective treatment.

Suppose you have been asked to determine the disease name and its probable treatment from the dataset and list it out in the form of a table or a dictionary like this.

A screenshot of a computer

Description automatically generated

After discussing the problem given above, you need to build a custom NER to get the list of diseases and their treatment from the dataset.

You have the train and the test datasets; the train dataset is used to train the CRF model, and the test dataset is used to evaluate the built model.

First, you will understand the ‘train\_sent’ and the ‘test\_sent’ datasets. Let’s take a look at the structure of these datasets using the image provided below.

A screenshot of a computer screen

Description automatically generated

Here, you need to understand that each word in this dataset is provided in a single line. So, first, you need to club all these words together to form the sentences. Moreover, there are blank lines given in the dataset that have been highlighted in the image given above. These blank lines indicate that a new sentence is starting from the next line onwards to the next blank line.

In the image provided above, you need to make the sentences in the following way:

Sentence1: …using a Spearman-rank Correlation

Sentence2: This relationship should be taken into account when interpreting the AFI as a measure of fetal well-being.

Sentence3: The study population…

...and so on.

You can also refer to the image given below to get a better idea on how to create sentences from words.

A screenshot of a computer

Description automatically generated

In this ‘train\_sent’ dataset, there are a total of 2,599 sentences when you form the sentences from the words. Similarly, there are a total of 1,056 sentences in the ‘test\_sent’ dataset when you form the sentences from the words.

Now, let’s take a look at the next datasets that are named ‘train\_label’ and ‘test\_label’.

A line of letters and numbers

Description automatically generated

The above dataset is about the labels corresponding to the diseases and the treatment. There are three labels that have been used in this dataset: O, D and T, which are corresponding to ‘Other’, ‘Disease’ and ‘Treatment’, respectively.

These labels correspond to each word that is available in the ‘train\_sent’ and 'test\_sent' datasets. So, there is one-to-one mapping of each label available in the 'train\_label' and 'test\_label' datasets with the words that are available in the 'train\_sent' and 'test\_sent' datasets, respectively. You need to again create the lines of labels corresponding to each sentence in the ‘train\_sent’ and the ‘test\_sent’ datasets as shown below.

A screenshot of a computer

Description automatically generated

So, in this ‘train\_label’ dataset, there are a total of 2,599 lines of labels when you form the lines from the label dataset. Similarly, there are a total of 1,056 lines of labels in the ‘test\_label’ dataset when you form the lines from the label dataset.

In this assignment, you need to perform the following broad steps:

* You need to process and modify the data into sentence format. This step has to be done for the 'train\_sent' and ‘train\_label’ datasets and for test datasets as well.
* After that, you need to define the features to build the CRF model.
* Then, you need to apply these features in each sentence of the train and the test dataset to get the feature values.
* Once the features are computed, you need to define the target variable and then build the CRF model.
* Then, you need to perform the evaluation using a test data set.
* After that, you need to create a dictionary in which diseases are keys and treatments are values.