**NLP UPDATE 1 REPORT**

**Data Collection.**

Firstly, we manually scraped information from the Yale official website and separated it into individual sentences totaling 128 plus samples. The official Yale website contains English information, which is our input language. For the output language, we want Hindi, so we used Google Translate to translate the sentences we got from Yale into Hindi, which can be used for training, testing, and validation of the model.

Thus collected 128 samples were manually loaded into a spreadsheet one by one. Under column English and Hindi.

**Data Processing.**

We built a dictionary for word2index, word2count, index2word, and vocabulary, as well as a total amount of words for each language, after iterating over all phrases to remove any undesired special characters, lower the text, and eliminate any trailing white spaces. which can be used later in the training process.

**Custom Dataset creation.**

We created a custom data set class that contains a Len method that returns the length of the dateset. And a **get-item** method that returns 4 things, an input sentence tensor, an output sentence tensor, input sentence and output sentence using an index.

Example :

tensor([368, 78, 344, 369, 164, 366, 42, 370, 371, 18) tensor([420, 258, 217, 30, 195, 413, 53, 421, 251, 358, 171,1])

'we have been expanding international collaborations in many areas ', 'हम कई क्षेत्रों में अंतरराष्ट्रीय सहयोग का विस्तार कर रहे हैं।'

**Data loading**:

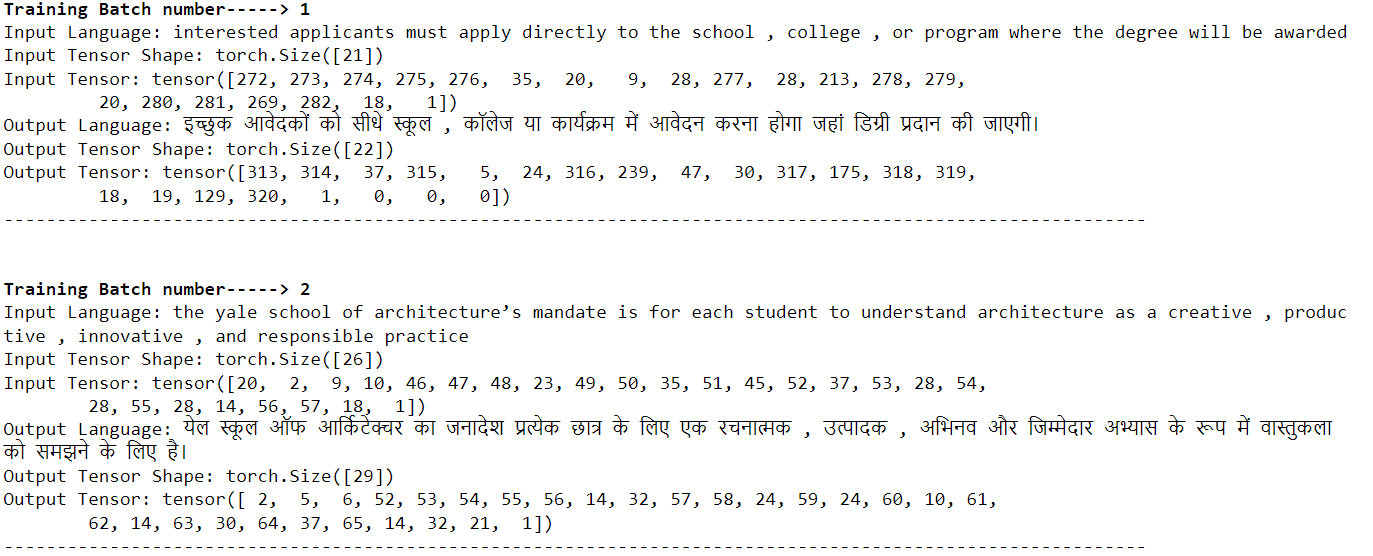
**Splitting the data:**

We have used the **train-test-split** method from sklearnto split the dataset into training, testing, and validation datasets.

**loading the dataset into a batch.**

We used **DataLoader**from PyTorch to create mini-batches of size 8, and we iterated through each mini-match and displayed the first sample. To overcome the problem of imbalanced tensor output sizes for each sample we used collate function which internally used the padding technique to solve this issue.

Example.



Dataset Link: <https://github.com/sujithkumar05/English-to-Hindi-machine-translation/blob/main/dataset.xlsx>