## Data Preparation Steps

Please follow below instructions to prepare data for NER model training:

* Once we receive the annotations data for an insurer (put respective insurer data folders inside **‘bucket’** & **‘out’** directory)
* Place **combine\_annotations\_files.py** file inside **‘out’** folder and change the directory name inside the file.Now, open a terminal and run this file inside **out** folder to combine different files(train\_page1.txt, train\_box\_page1.txt, train\_image\_page1.txt) for each annotated pages into three files namely (train.txt, train\_box.txt, train\_image.txt).
* After that run the **train\_test\_split\_script.py** file to split data into train-test split.Inside the file you can change the number of examples you want in test files.This will create total 6 files (train.txt, train\_image.txt, train\_box.txt, test.txt, text\_image.txt, test\_box.txt)
* Now, use **generate\_labels.py** file to generate labels or **labels.txt** file.Please change configuration according to use case (Motor/Health).

Note:

If you want to combine multiple files (train.txt, train\_box.txt, train\_image.txt) for different insurers, just place the respective annotated folder inside **bucket** & **out** and repeat the steps from 1-2.