Installation and Description of the Prediction Module

***The Prediction Module Package contains the following components:***

* Predictor **Vibraint\_PM\_Predictor.py**
* Initializer **Vibraint\_PM\_Initializer.py**
* Data Generator **Vibraint\_PM\_DataPredictor.py**
* Trainer **Vibraint\_PM\_Trainer.py**
* Model Encoder **encoder.h5**
* Model Decoder **decoder.h5**
* Menu Item Probabilities **Menu.csv**
* Training Sequences **Vibraint\_TrainSequence.csv**
* Test Sequences **Vibraint\_TestSequence.csv**

**Follow these instructions before running the Predictor script:**

1.Update the following variables in the Initializer script:

* **file\_path** : This variable should point to the directory in your system where the **Menu.csv** and **Vibraint\_TestSequence.csv** files are located. The output files will be stored in this path as well.
* **prediction\_time** : This variable specifies an hour slot ranging from 0 to 23 in 24-hr format. The model will predict the most probable menu items for this particular hour of the day.
* **generate\_seq** : Specify 1 if you want the program to generate synthetic data, else 0 where it will read from the corresponding input\_seq or test\_seq sequnce csv files
* **n\_samples** : The number of samples you want to generate, or thenumber of samples in the file
* **accuracy\_level** Level of accuracy required. If the value is 3, then the program will calculate Top 3 accuracy.

2. If you retrain the prediction model, the model files **decocer.h5** and **encoder.h5** must be in the same directory as the Predictor script

Description of the various components

**Predictor:**

* Run this script to obtain predictions.
* It generates a file called **Vibraint\_PM\_output.json** as its output.
* The prediction corresponds to the hour specified by the **prediction\_time** variable
* The Json file contains the top 6 menu items and their corresponding probabilities
* This program also predicts an output sequence for a given test sequence read from the Test Sequences file.
* It gives the Top-N accuracy, where N is the value of the **accuracy\_level** Variable in the Initializer. It utilizes the test sequences in the Test sequences file.
* Model output and individual probabilities for each test sequence are outputted in the **out\_target** and **out\_prob** csv files

**Initializer:**

* This is an initializer script, where variables of important variables are assigned and common functions are defined
* This program is used by all the remaining programs of the project

**Data Generator:**

* Generates random sequences of length **n\_steps\_in**
* Based on the probabilities specified in the Event Probabilities component, it generates **n\_samples** number of sequences.
* Generated sequences are written in the Training Sequences and Test Sequences files
* Output from this program are the source and expected output values in an one-hot-encoded array

**Trainer:**

* This program trains a prediction model with the samples generated or specified in the csv file
* It saves the trained model in .h5 format.
* Output of this program are model, encoder and decoder h5 files

**Model Encoder:**

* Saved encoder model from the Trainer
* This model is loaded in the Predictor module

Note: Different models have been generated and saved which are trained with different number of input sequences.

**Model Decoder:**

* Saved decoder model from the Trainer
* This model is loaded in the Predictor module

Note: Different models have been generated and saved which are trained with different number of input sequences.

**Menu Item Probabilities:**

* This is a csv file used for generating the input sequences
* This file contains the list of menu items and their corresponding probabilities distributed across Morning, Noon, Evening and Night.
* If we change the probability values, model needs to be retrained to reflect the appropriate probabilities.
* If menu items need to be added, they can be added at the end of the file. The variable **n\_features** in the Initializer module needs to be incremented by the number of menu items being added. This requires retraining the model.

**Training Sequences:**

* This is a csv file currenty generated by the Data Generator. In the future, it could contain actual patient event sequences.
* Each sequence contains 24 events, one per hour.

**Test Sequences:**

* This is a csv file currenty generated by the Data Generator. In the future, it could contain actual patient event sequences.
* Each sequence contains 24 events, one per hour.