**Chat bot assignment:**

The process I have followed for training and chatting with the chatbot are as follows:

1. Trained the chatbot with **cornell corpus** for 10 thousand iteration.
2. Now created a new preprocessing file “new\_preprocessing.py” in which am loading the **twitter dataset**.
3. Am completing the preprocessing by :
   1. Removing the emoticons.
   2. Removing the hashtags.
   3. Removing any kind of special characters
   4. Removing all the numbers
   5. Removing URLs
   6. Removing the punctuations
   7. Making all the words to lower case.
4. Then dividing the data into question and answer format for further processing the data
5. Using the prepare\_dataset() function in data.py am creating the necessary datasets for training my chatbot.

Both the cornell corpus data and twitter corpus data will are fed for training the chatbot.

**The observations and improvements are noted as follows when training the data :**

**Case 1:** Trained the chatbot for 1000 iterations, with initial bucket size. It was clearly seen that bucket 4 and bucket 5 was empty. And chat bot was returning the below output.

HUMAN ++++ hi

BOT ++++ lamella lamella lamella fiance fiance fiance subconscious subconscious subconscious subconscious subconscious subconscious wuv wuv wuv wuv lamb lamb lamb.

**Case 2**: reduced the bucket size to accommodate all the words and no bucket is left empty. For this I have analysed the data and found that longest sentence will have maximum of 45 words.

Hence I have fixed the bucket size to (5,5) (10,10) (15,15) (20,20) (35,35) (45,45) and trained the model again for 1 thousand iteration.

HUMAN ++++ hi

BOT ++++ you do you not solution.

HUMAN ++++ how are u

BOT ++++ i no have do you have.

**Case 3** : loaded the twitter dataset, preprocessed it successfully and with the preprocessed data and trained chatbot for 1 thousand iterations. This time I saw that chatbot was giving me the same reply.

HUMAN ++++ hi

BOT ++++ thanks

HUMAN ++++ hello

BOT ++++ thanks

**Case 4**. Hence again set the bot bucket size to [(5, 5), (10, 10), (15, 15), (25, 25), (35, 35), (50, 50)] and training the chatbot for 12 thousand iteration for more than 14 hours, the output from chatbot was improvised to a extent as below :

HUMAN ++++ heyya

BOT ++++ go thanks no off .

HUMAN ++++ can u reply something

BOT ++++ going to do you

HUMAN ++++ why are u behaving weird

BOT ++++ dirtbag news to do you

HUMAN ++++ who created u

BOT ++++ do you thanks going do you have

HUMAN ++++ helloo???

BOT ++++ yeah .

HUMAN ++++ hi ...

BOT ++++ thanks have nice

HUMAN ++++ wassup ..

BOT ++++ dirtbag news.

HUMAN ++++ yeah

BOT ++++ going to do you ?

Further action : From the above inferences, it is clearly seen that the chatbot needs to be trained for more number of iterations to get better result. But training the chatbot in my system takes long time due to insufficient configuration of laptop.

It took avg of 15 minutes for training 100 iterations. So it almost took 14 to 15 hours for training the chatbot for 12 thousand iterations.