2ndNov 2018

**Attendance: 10%, Continuous evaluation: 70%, Viva-20%**

**Assignment No. 6**

* Download and preprocess the sentiment analysis dataset from https://www.kaggle.com/bittlingmayer/amazonreviews/downloads/test.ft.txt.bz2/2 on Amazon reviews as was mentioned in Assignment 4.
* Download the Glove word vectors from <http://nlp.stanford.edu/data/glove.6B.zip> and extract the 100 dimensional file (glove.6B.100d.txt) from the zipped folder. Now complete the following steps.
* Create a dictionary of vocabulary obtained from preprocessing in the above step which contain index and word as key pair list. E.g:

vocab2id = {‘<PAD>’ : 0, ‘\_UNK’ : 1, ‘a’ : 2, ‘an’: 3, …...’together’ :7008}

id2vocab = {0 : ‘<PAD>’, 1 : ‘\_UNK’, 2: ‘a’, 3:‘an’,…...’together’ : 7008}

(Note: To add <PAD> and <\_UNK> token in vocab list)

* Create an embedding list(array) for vocab:

embedding\_list= np.array(len(vocab2id), 100)

Where each element represent id from vocab2id and corresponding word vector from (glove.6B.100d.txt) file.

Note: if word is not present in glove.6B.100d.txt, replace with random 50 dimensional vector ranging between (-.0.5 to +0.5).

* Create sentences in batches of length 64 by replacing each word with id from vocab2id. The dimension of batch is [64, 25, 100], where 64 is batch size, 25 is maximum word length in sentence (fill with zeros if sentence length is less than 25), 100 is the word vector size.
* Train a LSTM model with FC layer applied in final layer using the following tensorflow function:
* tf.nn.embedding\_lookup
* tf.contrib.rnn.LSTMCell
* tf.nn.dynamic\_rnn
* tf.layers.dense
* tf.train.AdamOptimizer
* Plot the graph of loss for training and validation set. Compare accuracy on test set and validation set.

Submit a report with result.