**CSCS 366 – Introduction to Natural Language Processing**

Assignment 3 Deadline: June 5th 11:59PM

**Part 1: Word2Vec**

1. Using genism, train word2vec embeddings of 300 dimensions on provided Roman Urdu dataset. Use “smsreceived” column as you text input.
2. Using your word2vec trained model, print top 10 words most similar to “police”
3. Save word2vec embeddings as text file.

**Part 2: Text Generation using LSTM**

1. Given the harry potter book, train a LSTM based language model using tensorflow keras.
2. Use as many LSTM units and as many LSTM stacked layers as you see fit.
3. Use sequence length of your choice.
4. You will need to convert your text to sequences. For input t, t+1 is your label. <start> is your starting token for inputs, and <end> is your ending token for labels.
5. For embeddings, use GloVe 300D embedding vectors to initialize your embeddings layer.
6. Once model is trained and your loss does not decrease further, generate text using initial seed words “<start> He had awoken”. Generate until you get <end> token generated by the model.

**Submission Guidelines:**

* Submit Jupyter Notebook with all four outputs above.
* Show output of all above tasks in the notebook.
* Submit on Moodle.
* Submit all the files so that I can run your code and reproduce the results.