Tutorial 5: Recurrent neural networks

1. Adapt the RNN-based language modelling codes at https://github.com/pytorch/examples/tree/master/word_language_model for Singlish SMS messages at https://github.com/jasonyip184/SGTextGenerationLSTM/blob/master/smsCorpus_en_2015.03.09_all.json as follows:
   a. Collect SMS messages from the JSON file
   b. Tokenize the messages using NLTK tokenizer (https://www.nltk.org/api/nltk.tokenize.html)
   c. Randomly split them into train (80%), validation (10%) and test (10%) subsets
   d. Train a language model with the messages
   e. Generate samples from the trained language model
   f. Try with different model type (e.g. GRU) and epochs and observe the generated texts
2. Revise Question 1 codes to use only training dataset for building vocabulary as follows:
   a. Collect vocabulary from training dataset
   b. Select a 'known' subset of training dataset vocabulary, whose tokens are most frequent and cover 99% of the train data
   c. Change unknown words in the three datasets to the special token '<unk>'