CS6910 : Deep Learning (for Computer Vision)
Programming Assignment 3

Instructor: Prof. Anurag Mittal (BSB 368)

amittal@cse.iitm.ac.in

1 Part-A : Train and Test your own Word2Vec

In this assignment, you are required to train your own Word2Vec model for the "text8" dataset. Please implement (a) Bag-of-words, (b) skip-gram and (c) LSTM-based models, as described in class. Note: You are not allowed to use any libraries like gensim for building the model.

Dataset

We are giving a python script (nlp_preprocessing.py) in Moodle to download and preprocess the dataset. You are required to install nltk and gensim to run the script. You can install these packages using pip / pip3 install.

2 Part B : Sentiment Analysis in movie reviews

You are required to classify movie reviews in one of the 5 classes based on the sentiment of the review. Make a feature representation of the review using pre-trained word embedding & your own embedding from Part-A and LSTM and then classify to one of the 5 classes mentioned below. Please try different parameters for word2vec training, such as the dimension of the representation, number of words predicted around a word etc. and create plots that describe the performance of the system as a function of these parameters. Please provide the Confusion Matrix and the classification accuracies for these.

2.1 Dataset: Rotten Tomatoes dataset. Available on Moodle as Train.tsv, Test.tsv

The sentiment labels are:
0 - negative
1 - somewhat negative
2 - neutral
3 - somewhat positive
4 - positive

Notes

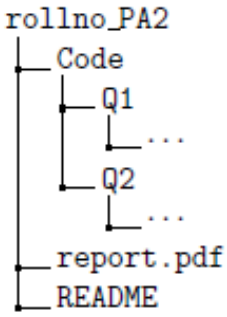
- We recommend PyTorch to implement the assignment.
- You can use pre-trained word embeddings. We recommend GloVe or FastText, which is available with the Torchtext package in PyTorch (https://torchtext.readthedocs.io/en/latest).

Plagiarism

- You should do the assignment yourself. In case you take help from others, please mention in the pdf submitted.
- No sharing of code/experiments etc. will be allowed under any circumstances and may attract disciplinary action by the institute disciplinary committee.

Submission Details

- **Deadline :** 11/30/2020 11 : 59 PM IST
- **What to submit :** You should prepare a report of the results obtained of your work. LaTeX is recommended for ease of work, but not essential. Submit a single tar/zip file containing the following files in the specified directory structure. Use the following naming convention: *rollno_PA1.tar.gz* or *rollno_PA2.zip*. A sample submission would look like this:



- **PDF & Code Upload:** On Moodle.

TAs:

- Gouthaman KV
- Arulkumar
- Asrar Ahmed
- Saikat Dutta
- Pawan Prasad

Please ask your doubts via the Moodle QA forum.