E0 334 - Deep Learning for NLP

Assignment 1

(due by 2nd Sept, 09:59 PM)

Note: Use the form available at the following link for submitting your results of Assignment 1.

`https://forms.office.com/r/9A9K1cFv99`

Problem:

The aim of this assignment is to study the use of different pre-trained word embeddings (word2vec/GloVe/fastText) for text representation and use them for text classification. You can download pre-trained word embeddings of your choice [1, 2, 3].

Design classifiers using any of the neural architectures discussed/mentioned in the class till Aug 21 for the following datasets:

1. Aug24-Assignment1-Dataset1 (Test data set, *Aug24-Assignment1-Dataset1-test*, will be made available on Sept 02 at 9:00 PM).
2. SST2 dataset (Using Train/Validation/Test splits available at `https://huggingface.co/datasets/stanfordnlp/sst2`

You could use different classifiers for each of the datasets.

Note: You can use libraries like Gensim (`https://radimrehurek.com/gensim/`) or spaCy (`https://spacy.io/`) for various text processing as well as other tasks. You can use t-SNE software[4] to visualize sentence/paragraph representations.

References

1. Word2vec (`https://code.google.com/archive/p/word2vec/`)
2. GloVe (`https://nlp.stanford.edu/projects/glove/`)
3. fastText (`https://fasttext.cc/`)
4. t-SNE Software (`https://lvdmaaten.github.io/tsne/`)
5. Iyyer *et al*, Deep Unordered Composition Rivals Syntactic Methods for Text Classification. `https://people.cs.umass.edu/~miyyer/pubs/2015_acl_dan.pdf`. Also see the code available at `https://github.com/miyyer/dan`