E0-334 Deep Learning for NLP

Assignment 2

(due by 23rd Aug, 11:59 PM)

Note: Use the following link for submitting your results.

`https://forms.office.com/r/QcyfbjUaFW`

Problem:

Consider the "20 Newsgroups" dataset available at the following site:

`http://qwone.com/~jason/20Newsgroups/20news-bydate.tar.gz`

We will consider the following three classes. The number of training and test documents are mentioned in the parenthesis.

1. rec.sport.hockey (600, 399)
2. sci.electronics (591, 393)
3. rec.auto (594, 396)

- Use TF-IDF (term frequency-inverse document frequency) to get a vector representation of each document in the corpus and design a three-class classifier.
- Report the micro-average F1 score of the designed classifier on the test set.
- Use t-SNE(t-distributed stochastic neighbor embedding) for visualizing the training set documents.