DDA4220/MDS6224/MBI6011 Deep Learning

Assignment 2: Text classification by using deep models

Due Date: 23:59, April 22, 2023

This assignment aims to train models for text classification, which will account for **15%** of the final grade.

Natural Language Processing with Disaster Tweets

This is a "Getting Started" competition from Kaggle https://www.kaggle.com/competitions/nlp-getting-started . In this assignment, you're challenged to build a machine learning model that predicts which Tweets are about real disasters and which ones aren't. 3 CSV files are provided:

- `train.csv`: it is the training set which contains:
  1) a unique identifier for each tweet;
  2) the *text* of a tweet;
  3) A *keyword* from that tweet (although this may be blank);
  4) the *location* the tweet was sent from (may also be blank);
  5) the *target* denoting whether a tweet is about a real disaster (1) or not (0).
- `test.csv` does not contain *target*, you need to format the results from your model as `sample_submission.csv`, and submit your results in Kaggle.

In this assignment:

1. You need to split the `train.csv` into a training set and validation set in a ratio of 7:3, and submit the results on the validation set.
   1) **(Baseline task, 40%)** The code base provided in Tutorial 5 is a baseline, you need to use it to implement the classification task for this assignment (The pandas library is useful to read CSV files). You need to submit reasonable results in the report, the model and code can output the results you reported.
   2) **(Improvement task, 15%)** You need to try other parameter settings, embedding methods, other networks, etc. upon the baseline to get different (even lower is OK) results. Feel free to modify the code or rewrite it. Also, you need to report your new results and submit the model and the code.
2. **(Kaggle competition, 5%)** You can predict on `test.csv`, and submit your predictions to the Kaggle competition. Please submit the result with a screen shot in the report.
3. **(Report, 40%)** You need to write a report using this template: https://www.overleaf.com/read/bxhymndwhdyp. It contains the results and methods for your provided codes. Higher scores do not mean higher points for this assignment, only give your analysis to show that the predictions of your models make sense.

Submission requirements:

1. Please submit *code+model* for the **baseline task** in one folder named `baseline`.
2. Please submit *code+model* for the **Improvement task** in one folder named `improvement`.

3. Please submit the *report* named stu_id.pdf.

All materials must be submitted to GitHub, other submissions without explanation will not be accepted this time. Learning to use Git is helpful. If you cannot upload the model due to the file-size limitation, you can also upload your model to BB or share a driver link of the model. Assignment link: https://classroom.github.com/a/sT1bEg_o

The deadline is **23:59PM, April 22**. For each day of late submission, you will lose 10% of your mark in the corresponding assignment. If you submit more than three days later than the deadline, you will receive a zero on this assignment. No late submission emails or messages will be replied to.