# 110-2
# Natural Language Processing

HW4

TA: Kuei-Chun Kao

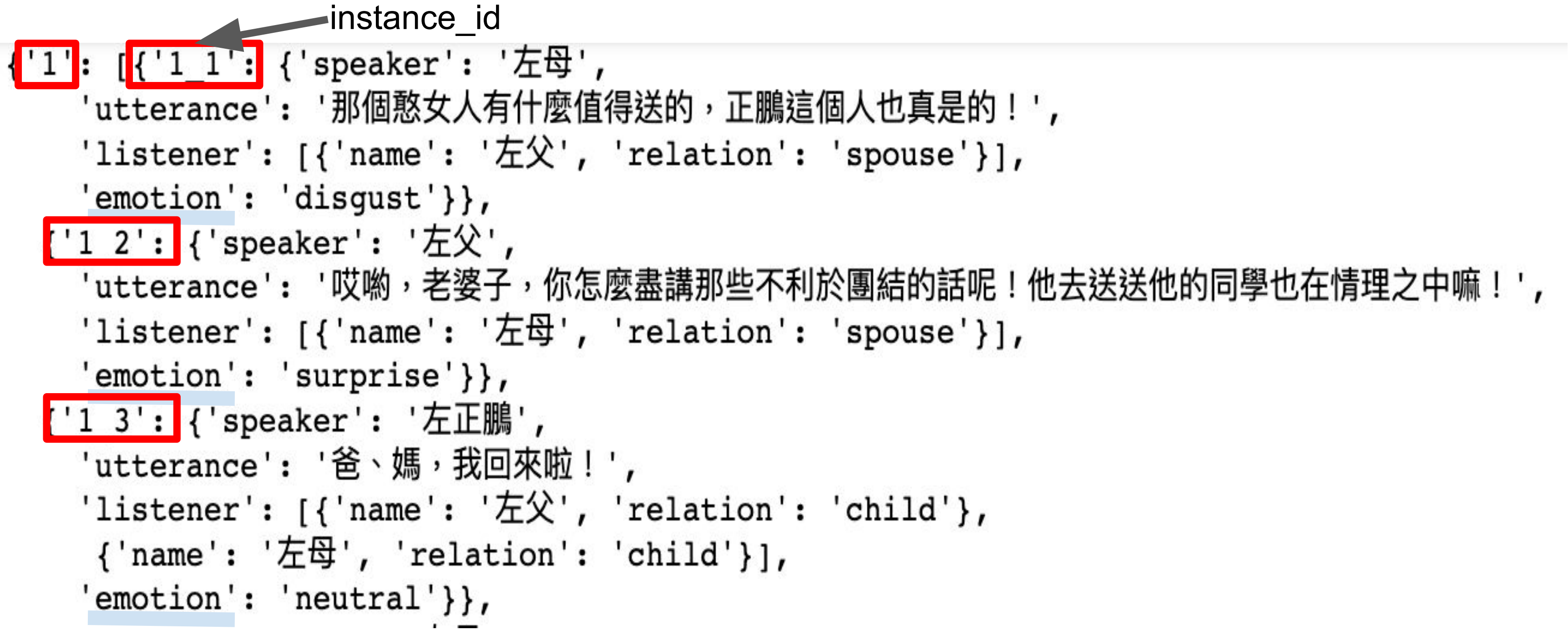
# Task introduction (Sentiment Analysis)

- What is Sentiment Analysis?
  - A chinese dialogue dataset is an indispensable resource for building a dialogue system. Additional information like emotions and interpersonal relationships labeled on conversations enables the system to capture the emotion flow of the participants in the dialogue.
- Learning objectives: train your own NN model to classify 7 emotions

# Dataset

- Dataset format: json format (need to parse by yourself)
- The file train.json & test.json contain the dialogues. Each dialogue has a unique case index value in the json file, and is a list composed of the utterances in speaking order. Every utterance in the list contains the speaker, content, and annotated labels shown in data format. The list of the listener in the utterance contains all listeners in this utterance with their relation type.
- The metadata is given in metadata.json. The file defines all the emotion, relation types, and the subclasses in the two perspectives, position, and field. The data format of metadata.json is shown as follows.

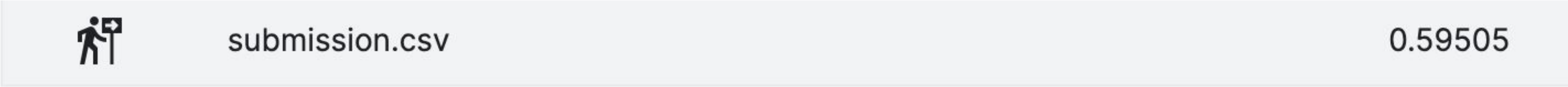
# Example (train.json)



```
{'1': [{'1_1': {'speaker': '左母',
    'utterance': '那個憨女人有什麼值得送的，正鵬這個人也真是的！',
    'listener': [{'name': '左父', 'relation': 'spouse'}],
    'emotion': 'disgust'}},
  {'1_2': {'speaker': '左父',
    'utterance': '哎喲，老婆子，你怎麼盡講那些不利於團結的話呢！他去送送他的同學也在情理之中嘛！',
    'listener': [{'name': '左母', 'relation': 'spouse'}],
    'emotion': 'surprise'}},
  {'1_3': {'speaker': '左正鵬',
    'utterance': '爸、媽，我回來啦！',
    'listener': [{'name': '左父', 'relation': 'child'},
     {'name': '左母', 'relation': 'child'}],
    'emotion': 'neutral'}},
```

# Outputs

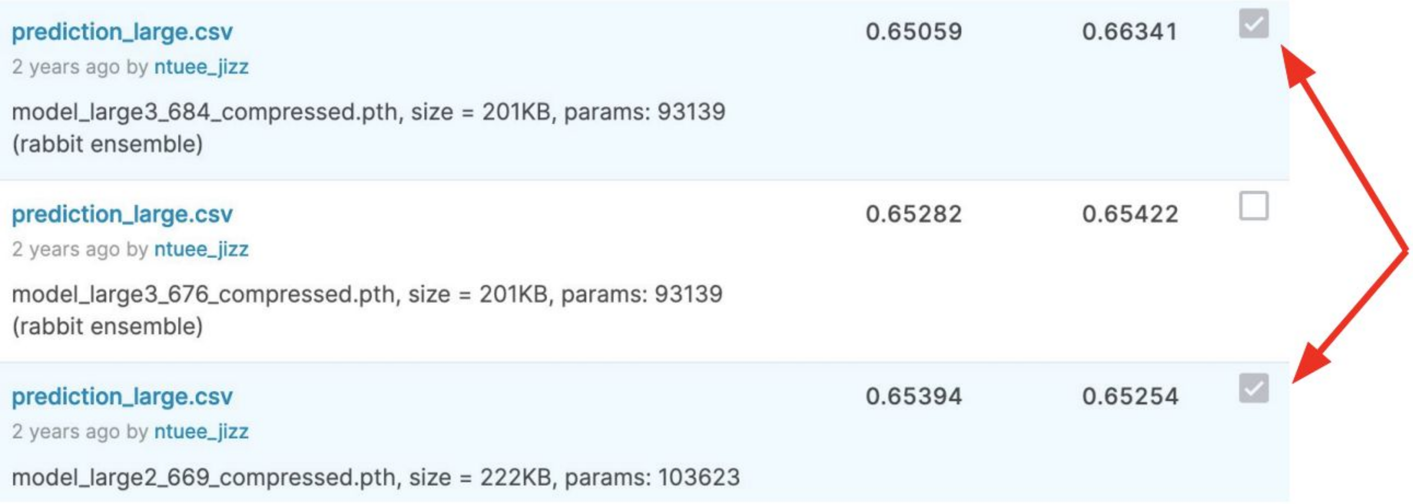
- Each line consists of two fields separated by horizontal whitespace (a single tab or space character).  The first field is the instance ID of a dialogue from the Json file.  The second field is the emotion type.
- Kaggle Link:
  https://www.kaggle.com/t/6dd77085715c4802996103cd26676c00
- Displayed name: <student_ID>
- Submission format: .csv file (You can also see from sample_submission.csv)
- Evaluation metric: Accuracy



submission.csv 0.59505

# Kaggle submission

- You may submit up to 5 results each day (UTC).
- Up to 2 submissions will be considered for the private leaderboard



| | | | |
|---|---|---|---|
| prediction_large.csv<br>2 years ago by ntuee_jizz<br>model_large3_684_compressed.pth, size = 201KB, params: 93139 (rabbit ensemble) | 0.65059 | 0.66341 | ☑ |
| prediction_large.csv<br>2 years ago by ntuee_jizz<br>model_large3_676_compressed.pth, size = 201KB, params: 93139 (rabbit ensemble) | 0.65282 | 0.65422 | □ |
| prediction_large.csv<br>2 years ago by ntuee_jizz<br>model_large2_669_compressed.pth, size = 222KB, params: 103623 | 0.65394 | 0.65254 | ☑ |

remember to select **2** results for your final scores before the competition ends!

# Reference (You can follow some ideas here)

- You can try different variant of models and use ensemble method to give final prediction
- You can split your train folder into training set and validation set
- You can use the additional information from dataset. (i.e. relationship between speakers, social field…etc in metadata.json)
- Consider local and global features from dialogue
- Use the packages/tools I allow:
    - Python 3.8 / 3.9 and Python Standard Library
    - Pytorch,tensorflow
    - SpaCy, sentencepiece, nltk, stanfordcorenlp, huggingface
    - datasets, json, tqdm, itertools,numpy,pandas…etc
    - Dependencies of above packages/tools.

# Requirements

- Python only
- Use publicly available pre-trained BERTs and their variants.
- Can’t use model trained with other Dialogue data.
- No plagiarism!
- At the top of your Source code

```
#Author: Kuei-Chun Kao

#Student ID: 1234567

#HW ID: Hw4

#Due Date: 01/30/2020
```

# Submission

- Deadline
  - Submit Zip to E3 before 6/5 11:59 PM
  - No Late Submission, thanks!
- Format
  - Source code: Hw4_<StudentID>.py (py only, your main file), you can split anther utils files, if you needed. Be careful for your import path.
  - Report file: Hw4_<StudentID>.pdf (pdf only)
  - Make sure the .py file contains the correct execution results and formats.
  - If can’t compile correctly, no score for you
  - Zip file: Hw4_<StudentID>.zip (zip only)
- Any question can ask me on E3, answer your question ASAP

# Grading policy

- Ranking score in Kaggle Leaderboard (50%)
- Report (40%)
- Code comment, file format, display name  (10%)
- I can only see your last submission.
- Do not submit your model or dataset.
- If your code is not reasonable, your final grade will be multiplied by 0.8!
- You should NOT modify your prediction files manually.
- Do NOT share codes or prediction files with any living creatures.
- Do NOT use any approaches to submit your results more than 5 times a day.

# Ranking score in Kaggle Leaderboard (50%)

- Public leaderboard (20%): Your public leaderboard score > baseline, you can get 20% of this part; Otherwise, you can only get 10% of this part.
- Private leaderboard (30%): Your private leaderboard score * 30%
- This part score = public leaderboard + private leaderboard

# Report (40%)

1. Your model design and concept (8%)
2. Error Analysis and Experiment Discussion (8%)
3. 請描述嘗試過的方法，並且討論曾經遇到的問題以及解決的方法 (8%)
4. 請根據實驗結果分析哪一類的情緒較容易預測？哪一類的情緒較難預測？較難預測的情緒容易被誤認為何種情緒？(可使用 confusion matrix 進行分析討論)並探討可能的原因（給一些實際例子)且提出解決辦法 (8%)
5. 請分析給定不同上下文的資訊量(句數)，對於預測該句話情緒的影響 (8%)

# Q1: Data processing (2%)

1. Chinese Preprocessing and Tokenizer(1%):
    a. Describe in detail about the tokenization algorithm you use. You need to explain what it does in your own ways.
    b. Describe your preprocessing for Chinese term.

2. Answer Prediction (1%):
    a. After your model predicts the probability of answer, what rules did you apply to determine the final prediction?

# Q2: Modeling with LMs and their variants (6%)

1. Describe (4%)
    a. your model (configuration of the model) and details
    b. performance of your model.
    c. the loss function you used.
    d. The optimization algorithm (e.g. Adam), learning rate and batch size.

2. Plot
    a. Learning curve of loss (1%)
    b. Learning curve of Accuracy (1%)

# Bonus

- If your ranking is top 3 in class, you can get 3 points bonus in this hw final score!
- Last HW, many thanks to everyone~