Introduction to Natural Language Processing

HW2

TA : WEI-LING HSU

weiling.hsu.cs11@nycu.edu.tw

Task introduction

- **Sentiment analysis** with neural network

For example : “This movie is so funny !” 

- You can use any external package you want, e.g. Pytorch, Keras…
- Tips: Use RNN with word embedding and attention to build your model
- You can only use RNN or CNN, other advanced model (e.g. BERT) is not allowed in hw2.

- Requirement:

1.Submit a report and your source code to E3

2.Upload your submission to Kaggle

Dataset

https://drive.google.com/drive/folders/1PD38uQGZwQfIeHIw-t9d2YKcW21KYAPI?usp=share_link

You can get the dataset from the link or E3,

it contains : < train_HW2dataset.csv >

< dev_ HW2dataset.csv >

< test_ HW2dataset.csv >

You can use < train_ HW2dataset > and < dev_ HW2dataset > to tune your model,

then predict the emotion of Utterances in < test_ HW2dataset >

Column Specification

Utterance	Speaker	Emotion	Dialogue_	Utterance_	Old_Dialc	Old_Utter:	Season	Episode	StartTime	EndTime
also I was	Chandler	neutral	0	0	0	0	8	21	00:16:16,(	00:16:21,731

Column Name	Description
Utterance	Individual utterances from EmotionLines as a string.
Speaker	Name of the speaker associated with the utterance.
Emotion	The emotion (neutral, joy, sadness, anger, surprise, fear, disgust) expressed by the speaker in the utterance.
Dialogue_ID	The index of the dialogue starting from 0.
Utterance_ID	The index of the particular utterance in the dialogue starting from 0.
Season	The season no. of Friends TV Show to which a particular utterance belongs.
Episode	The episode no. of Friends TV Show in a particular season to which the utterance belongs.
StartTime	The starting time of the utterance in the given episode in the format 'hh:mm:ss,ms'.
EndTime	The ending time of the utterance in the given episode in the format 'hh:mm:ss,ms'.

Output

- Your model need to predict which type of emotion is for Utterance in < test _HW2dataset.csv >
- We have 7 types of emotion at all, please map them into 0~6.

  neutral -> 0, anger -> 1, joy -> 2

  surprise -> 3, sadness -> 4, disgust -> 5, fear -> 6

For example:

Input Utterance “Good morning, John.”, model output 0.

Input Utterance “This movie is so funny!”, model output 2.

Kaggle submission(50%)

	A	B
1	index	emotion
2	0	0
3	1	1
4	2	6
5	3	2
6	4	3
7	5	3

- Kaggle link: https://www.kaggle.com/t/964d464205694362a12061ddc60b1e47
- Display team name:<student ID>
- Submission format:
  - A 3401 X 2 .csv file, first row is for the column name and the last 3400 rows for your result.
  - Column name should be index and emotion.
  - Result should be the index and emotion(0~6),please make sure the order of your result is wright!
- There will be two bassline(strong and simple).

I will provide hints about simple baseline on 11/29 14:00.

Get bonus if you achieve the simple baseline before that time.

Kaggle submission(50%)

- You can submit at most 5 times each day.
- You can choose 2 of the submissions to be considered for the private leaderboard, or will otherwise default to the best public scoring submissions.
- The scoring metric will be macro F1, not accuracy!

Report Submission(50%)

- Submit a report contains 4 questions:

  1. Describe how you build your model ? How did you do to preprocess your data from dataset ? The distribution of the emotion is unbalance, what did you do to improve the accuracy on those emotion which are in small scale?(30%)

  2. Have you tried pretrain word embedding?(e.g. Glove or Word2vec).What is the influence of the result after you using them?(30%)

  3. Have you tried attention on your model? What is the influence of the result after you using them? Which text your model attention on when it predict the emotion?(30%)

  4.Have you used other information form dataset to improve your model performance?(e.g. Speaker) What is the influence of the result after you using them?(10%)

Please answer the questions in detailed!

E3 Submission

- Deadline:
    - Submit Zip to E3 before 12/8 11:59PM
    - No Late Submission!
- Format:
    - Source code : Hw2_<student ID>.py
    - Report file : Hw2_<student ID>.pdf
    - Zip file : Hw2_<Student ID>.zip

Grading policy

- Kaggle(50%):

  - Basic score (according to the public leaderboard):

    under simple baseline: 40

    over simple baseline: 60 *1.2 If you achieve before TAs provide hints(10/25 14:00)

    over strong baseline: 80

Ex:

  - If you achieve simple baseline after 11/29 but you don't achieve strong baseline in the end, your basic score will be 60.
  - If you achieve simple baseline before 11/29 but you don't achieve strong baseline in the end, your basic score will be 72.
  - If you achieve simple baseline before 11/29 and you achieve strong baseline in the end, your basic score will be 80.

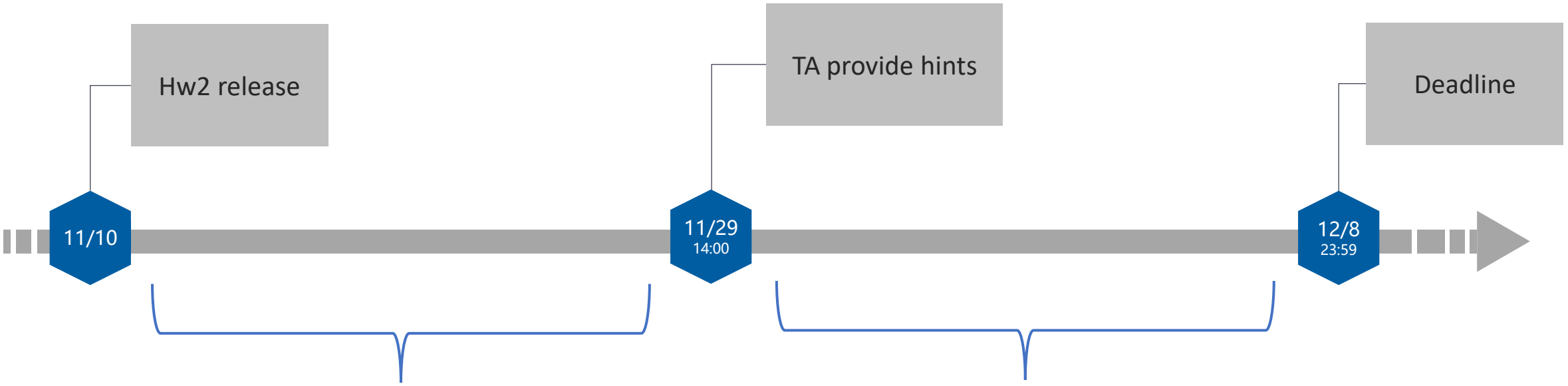
  - Ranking score (according to the private leaderboard) :

    score= 20-(20/N)*(ranking-1), N=number of people

- Source code and Report(50%):

  - the more detail you make, the higher score you get

Timeline



Simple baseline score: 60*1.2=72
Strong baseline score: 80

Simple baseline score: 60
Strong baseline score: 80

If you have any question about HW2, please feel free to contact with TA: WEI-LING HSU, through email weiling.hsu.cs11@nycu.edu.tw