# Natural Language Processing

## HW1

TA: Hsiu-Hung Lee

# Task introduction

- Identify whether the extracted subject, verb, and object are valid

- You can use the packages you need, e.g., Spacy, NLTK, …etc.

- Tips: You can start from POS-tagging.

- Requirements:
  1. Submit a report and your source code to E3
  2. Upload your submission to Kaggle

# Inputs

https://drive.google.com/file/d/10apUEA8DVjM8DI8J1Dconc91SlOG1vH1/view?usp=sharing

This file contains 1262 rows,each contain:

- ID
- Sentence
- Subject (S)
- Verb (V)
- Object (O)

| ID | Sentence | S | V | O |
|---|---|---|---|---|
| 1 | Rhodes discovered he had cancer last October after he felt a sharp pain in his right leg . | he | felt sharp pain in | his right leg |

# Outputs

- You need to determine whether S contains the subject of the sentences. (verb, object as well)

- Output 1 if all the S, V, O contain the corresponding text,else output 0

- EX: ”I eat a delicious burger.”
  If S contains ”I” and V contains ”eat” and O contains “burger”, output 1, else output 0.

# Kaggle submission(40%)

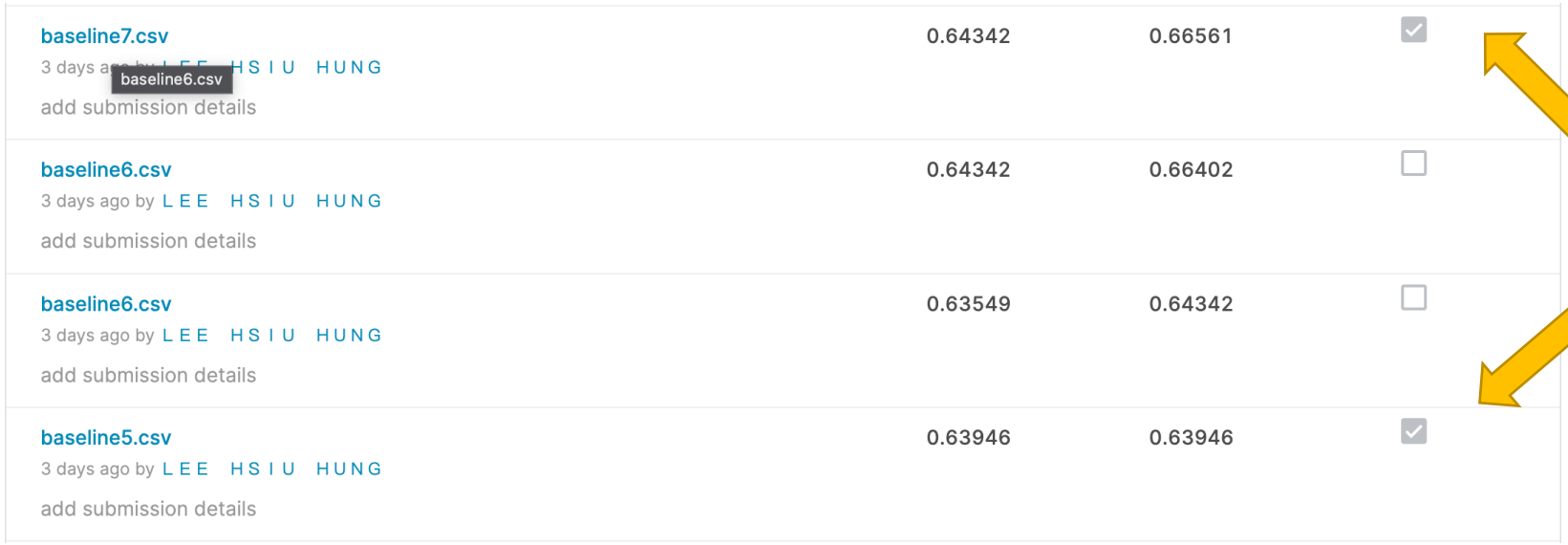
| index | T/F |
|---|---|
| 0 | 0 |
| 1 | 0 |
| 2 | 0 |
| 3 | 0 |
| 4 | 0 |
| 5 | 0 |
| 6 | 0 |
| 7 | 0 |
| 8 | 0 |
| 9 | 0 |
| 10 | 0 |

- Kaggle link:
  https://www.kaggle.com/t/36fbc49fc9d644f0ae9cfc76dae3331b
- Display name:<student ID>
- Submission format:
  A 1263 X 2 .csv file containing :
  - 1 row of column name and 1262 rows of result.
  - The column name must be “index” and “T/F”
  - Each row contains an index(from 0 to 1261) and a prediction(0 or 1)

- There are two baseline (simple and strong)，
  I will provide hints about simple baseline on 10/25 14:00.
  Get bonus if you achieve the simple baseline before that time.

# Kaggle submission(40%)

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




Remember to select 2 results for your final score before the competition ends!

# Report Submission(60%)

Please submit a report containing 3 questions:

1. Describing  your methods in detail.(50%)

2. Is there any difference between your expectations and the results? Why?(20%)

3. What difficulties did you encounter in this assignment?
   How did you solve it?(30%)

# Requirements

- Python only
- No plagiarism !
- At the top of your Source code:

```
#Author: Hsiu-Hung Lee
#Student ID: 1234567
#HW ID: HW1
```

# E3 Submission

- Deadline:
  - Submit Zip to E3 before 11/3 11:59PM
  - No Late Submission!
- Format:
  - Source code : Hw1_<student ID>.py
  - Report file : Hw1_<student ID>.pdf
  - Zip file : Hw1_<Student ID>.zip

If you have any question, ask all the TAs by Email.

# Grading policy

- Kaggle(40%):

  - 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 10/25 but you don't achieve strong baseline in the end, your basic score will be 60.
  - If you achieve simple baseline before 10/25 but you don't achieve strong baseline in the end, your basic score will be 72.
  - If you achieve simple baseline before 10/25 and you achieve strong baseline in the end, your basic score will be 80.

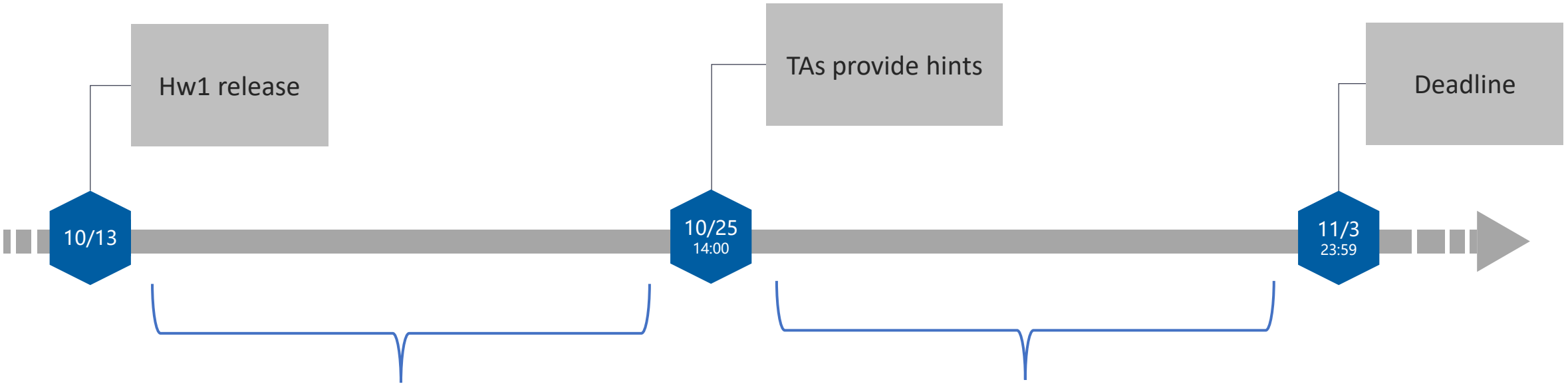
  - Ranking score (according to the private leaderboard) :

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

- Source code and Report(60%):

  - 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