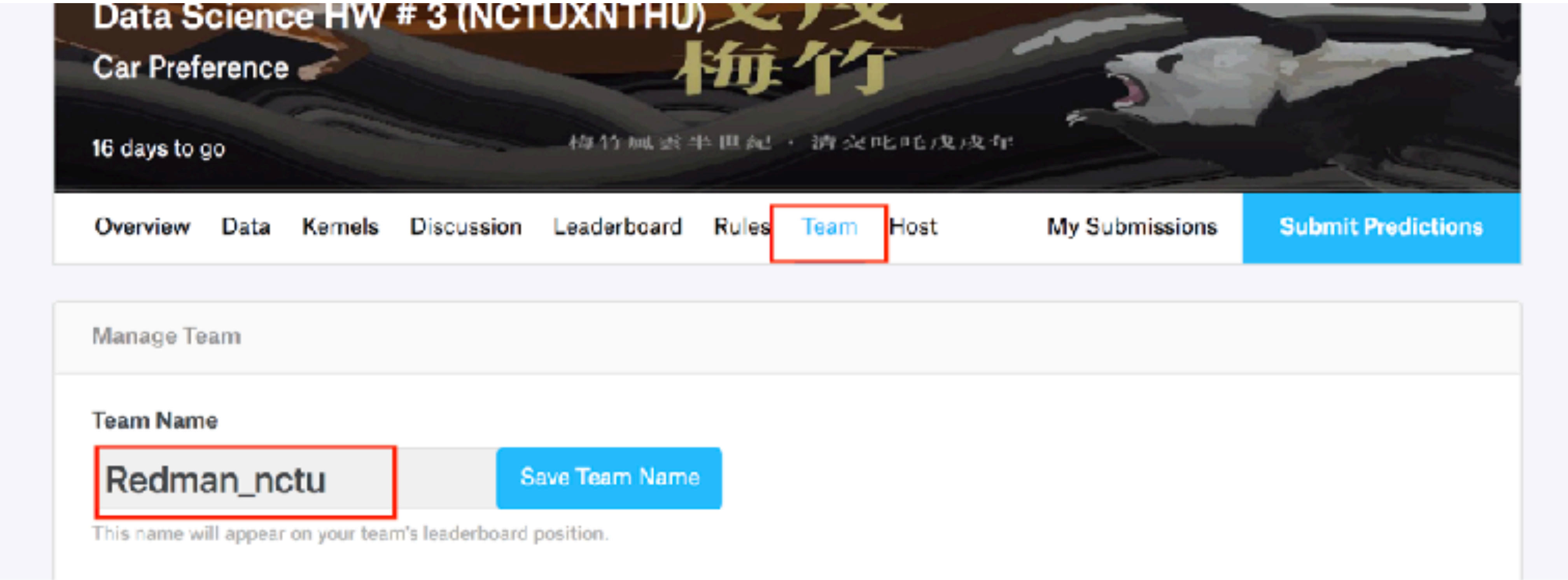
Data Science HW5

Jokes Rating

陳泓仁

HW #5

- Kaggle in-class
  - https://www.kaggle.com/t/fc5eca936a5e4740878789ba1c684d0d
  - Deadline: 06/12/2018 11:59 PM
  - Please add your university name.



Data Science HW # 3 (NCTUXNTHU)

Car Preference

16 days to go

Manage Team

Team Name

Redman_nctu

This name will appear on your team's leaderboard position.

Dataset

Anonymous Ratings from the Jester Online Joke Recommender System.

We take 100 jokes and 10000 users who have rated 36 or more jokes and the user ratings ranging from -10.00 to +10.00 for 100 jokes(the value "99" corresponds to "null" = "not rated").

- #of training data: 777978
- #of testing data: 222022
- #of users: 10000
- #of jokes: 100

Dataset

The text of the jokes:

1. 100 files
2. Each file has title init_.html, where _ is 1 to 100
3. The titles correspond to the ID's of the jokes in the Excel files above

You need to crawl the text of the jokes by yourself.

Dataset

Training Data:

- User ID - an id unique to a given user
- Item ID - the id of a joke
- rating - the rating

Testing Data:

- User ID - an id unique to a given user
- Item ID - the id of a joke

Dataset

Sample training data:

train

user_id	item_id	rating
15	32	5.83
2	33	3.06
9	2	6.07
17	3	3.11
17	78	99

Sample testing data:

test

user_id	item_id
10	1
9	34
19	74
2	84
8	51

Submission

1. The maximum number of daily submissions is 10.

2. The submission file should be CSV file and contain two columns:

- uer_id-item_id : an id unique to a given user and jokes.
- rating : the rating needs to real value(ranging from -10 to +10).

3. The competition will take 50% of the test data to calculate the RMSE.

   Final rank will show on E3 after the competition.

4.The evaluation for this competition is Root Mean Square Error (RMSE)

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

Submission

sample

user_id-item_id	rating
10-1	0
9-34	0
19-74	0
2-84	0
8-51	0
3-48	0
5-66	0
15-48	0
16-31	0

Grading policy

Kaggle rank:

Beyond baseline (4): 0

top 10%: 100
top 25%: 90
top 50%: 80
top 75%: 75

Others: 70

Requirements

Please archive your code, testing result and submit on E3.
Deadline: 06/12/2018 11:59 PM

Submission folder (your team name on Kaggle) should contain 2 files:

- [Student ID].py
- answer.csv

EX.
Redman_nctu :
- 0310707.py
- answer.csv

Contact Information

- If you have any questions, please email 陳泓仁.
  - Gmail: 0226.hjc@gmail.com

- FB Group:
  - www.facebook.com/groups/156477025052136/

Questions?