Data Science HW4

Start 2019/11/26 Now

Deadline 2019/12/17 23:59

Task Description

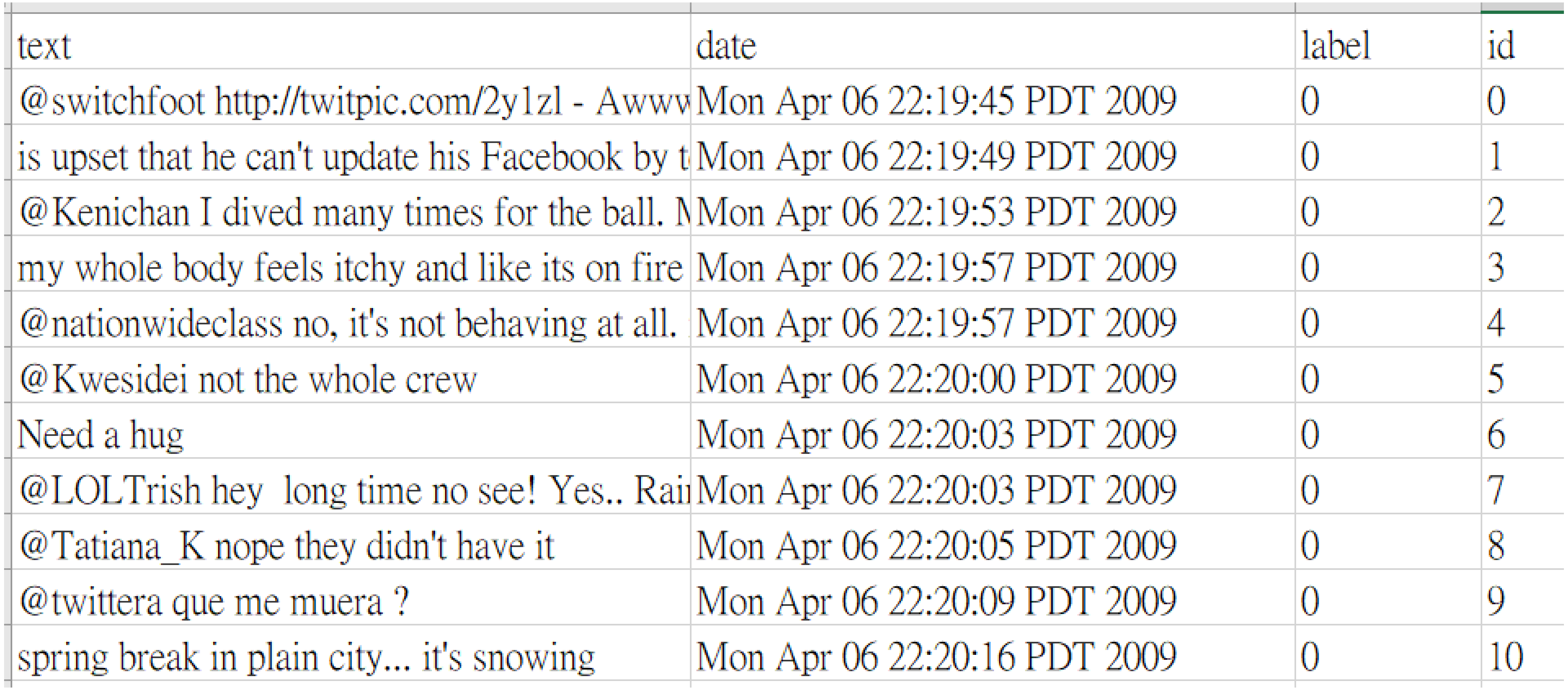
- In this competition, you need to predict the polarity of a tweet based on the **content** and the **date** (you can choose one or two). There are 1,600,000 data, while 80% for training and 20% for testing (50% **public** and 50% **private**). The label is the polarity of a tweet (0 for negative and 1 for positive). The score will be evaluated **by mean square error**.

Dataset Description

There are 4 attributes for one instance:

- **id**: the index for the dataset
- **label**: the polarity of the tweet (0 = negative, 1 = positive)
- **date**: the published date (e.g. Sat May 16 23:58:44 UTC 2009)
- **text**: the content of a tweet (Need a Hug)

Training Data



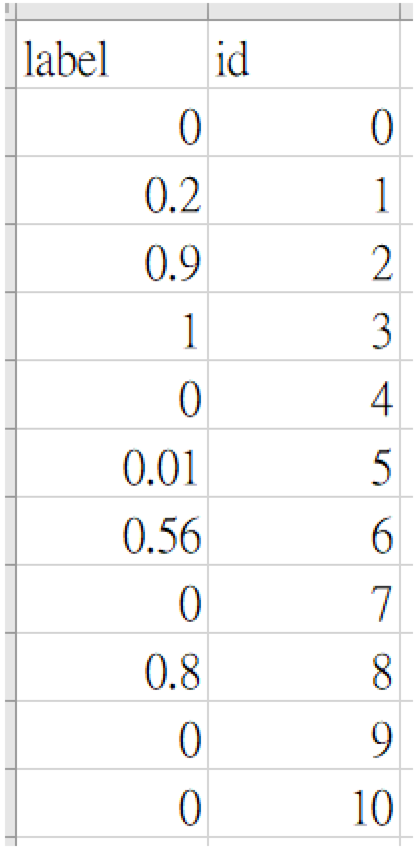
text	date	label	id
@switchfoot http://twitpic.com/2y1zl - Awww	Mon Apr 06 22:19:45 PDT 2009	0	0
is upset that he can't update his Facebook by t	Mon Apr 06 22:19:49 PDT 2009	0	1
@Kenichan I dived many times for the ball. M	Mon Apr 06 22:19:53 PDT 2009	0	2
my whole body feels itchy and like its on fire	Mon Apr 06 22:19:57 PDT 2009	0	3
@nationwideclass no, it's not behaving at all.	Mon Apr 06 22:19:57 PDT 2009	0	4
@Kwesidei not the whole crew	Mon Apr 06 22:20:00 PDT 2009	0	5
Need a hug	Mon Apr 06 22:20:03 PDT 2009	0	6
@LOLTrish hey long time no see! Yes.. Rai	Mon Apr 06 22:20:03 PDT 2009	0	7
@Tatiana_K nope they didn't have it	Mon Apr 06 22:20:05 PDT 2009	0	8
@twittera que me muera ?	Mon Apr 06 22:20:09 PDT 2009	0	9
spring break in plain city... it's snowing	Mon Apr 06 22:20:16 PDT 2009	0	10

Submission format

- You can only submit the testing results 10 times/day.
- The submitted file should be CSV file, with 2 columns:

  id: the index of the original order

  label: the real number between 0 to 1



label	id
0	0
0.2	1
0.9	2
1	3
0	4
0.01	5
0.56	6
0	7
0.8	8
0	9
0	10

Evaluation

- There are two phases: public and private.
- You can only see the result of public testing data (50%) when the competition going. After the competition, the private testing result will be revealed.
- The evaluation metric is MSE (Mean Square Error)

$$L_{mse} = \frac{1}{N}\sum_{n=1}^{N}(\widehat{y_i} - y_i)^2$$

Grading Policy

- Top 10%: 100
- Top 25%: 90
- Top 50%: 80
- Top 75%: 75
- Others: 70
- Below baseline(loss>0.15): 0

Important Information

- Kaggle competition link: https://www.kaggle.com/t/73e25da0b0b64029b4a9ec5c5178ea62
- Your team name should be your student id (i.e. 0750123).
- Never borrow others’ code or submission file, it should be done individually. (Or you might get 0 point)
- Hand in the source code and a readme file indicating the used libraries and how to generate you submission file on e3 (in one zip file). TA should be able to regenerate the submitted result. After unzipping, it should appear a single directory, containing the source code and readme. Name the zip file as your student id (i.e. 0750123.zip).

Important Information

- Deadline: 2019/12/17 PM 23:59
- Good Luck
- TA: Yun-Zhu yunzhusong.eed07g@nctu.edu.tw