

**HW # 5: If you talk to a man in a language he
understands, that goes to his head. If you talk to him in
his language, that goes to his heart:
Text Sentiment Prediction**

- Kaggle In-class
- Competition Link: <https://www.kaggle.com/t/367d86aef4d74639afa4fda3b08639e1>
- Deadline: 12/31/2018 23:59 PM

HW # 5: If you talk to a man in a language he understands, that goes to his head. If you talk to him in his language, that goes to his heart:

Text Sentiment Prediction

- predict sentiment for given tweets in the sentiment140 dataset. It contains 1,600,000 tweets, 80% for training, 20% for testing(50% for private and 50 % for public), extracted using the twitter api . The tweets have been annotated (0 = negative, 4 = positive).

Data

- There are 6 attributes for one instance. Below is a brief overview of type and values for various features in the dataset.
sentiment: the polarity of the tweet (0 = negative, 4 = positive)
id: The id of the tweet (2087)
date: the date of the tweet (Sat May 16 23:58:44 UTC 2009)
query: The query (lyx). If there is no query, then this value is NO_QUERY.
user: the user that tweeted (robotickilldozr)
text: the text of the tweet (Lyx is cool)

Training data

sentiment	id	date	query	user	text
0	1760244167	Sun May 10 21:01:23 PDT 2009	NO_QUERY	cgomezZz	drizzy's gon be big and overplayed now i want drake to be the best kept secret, ahh.
0	2049203163	Fri Jun 05 16:28:32 PDT 2009	NO_QUERY	lingerlonger247	Change In plans, have to go to drivein sat nite. Oldest has softball @ 9am. At least my ex tol
4	1692219096	Sun May 03 19:31:40 PDT 2009	NO_QUERY	BlackaRicanMa	I'm soooo excited rite now!
0	2016410015	Wed Jun 03 06:43:43 PDT 2009	NO_QUERY	dee2283	no sleep again ... dam it sucks when you want someone so much that you can even sleep in
4	1969701899	Sat May 30 00:55:16 PDT 2009	NO_QUERY	cruznicol	time to out..... see yah!!!!!!!!!!!!!! bye 4 now
4	1970922598	Sat May 30 05:40:14 PDT 2009	NO_QUERY	beyondbasscamp	B*B*C has a write-up in the latest issue of Bass Guitar Magazine - hence the reason for the e
4	2188342623	Mon Jun 15 21:43:27 PDT 2009	NO_QUERY	Tastelikecrazy	@jdgaby Free Iran, dear. Cause everyone deserves equal and fair representation.
0	1956737727	Thu May 28 22:33:53 PDT 2009	NO_QUERY	foodeater	@LAsnark I am not a spammer and was sad you unfollowed but then again i dont want peop
0	2221651817	Thu Jun 18 05:46:26 PDT 2009	NO_QUERY	RealSamCarter	@roystern09 Im still awaiting a phone call for rights to some music

Testing data

id	date	query	user	text
1976713459	Sat May 30 18:50:51 PDT 2009	NO_QUERY	saamanthajaane	@HelloLizzi sick and tired. and sore. and pissed off. i have exams tomorrow : har
2000683349	Mon Jun 01 22:41:41 PDT 2009	NO_QUERY	jessicaconnell	June 5th - aunt June 12th - 8am (zzz...) - counsin June 12th - 5pm () - BFF
2062038727	Sat Jun 06 22:11:12 PDT 2009	NO_QUERY	silverishblue04	is going to Cedar Point on July 11th with Dom!!!! so excited!
2052931597	Sat Jun 06 02:35:46 PDT 2009	NO_QUERY	BlurryJoy	@chellanglo - hi there. i am quite pleasant
2032189115	Thu Jun 04 10:54:05 PDT 2009	NO_QUERY	Listersmate	@TwittyWoman Hey! I'm doing alright.
1753254815	Sun May 10 00:46:51 PDT 2009	NO_QUERY	wrecktify	@krystynchong Love to see that!
2204375493	Wed Jun 17 02:03:05 PDT 2009	NO_QUERY	Chopstyx	@MamaGaea Hey beautiful, hate to say but this week isn't good I forgot what we
2051217775	Fri Jun 05 21:10:25 PDT 2009	NO_QUERY	taliasunset	Watching lightning storms & blowing bubbles off the balcony

Evaluation

- The competition will take 50% of the test data to calculate the accuracy.
Final rank will show on E3 after the competition.
- The evaluation for this competition is MSE(Mean Square Error).

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

Submission files

- The maximum number of daily submissions is **10**.
- The file should be **CSV file** contains two columns:
ID: the index
Sentiment: 4 = positive
0 = negative

ID	Sentiment
0	1.93
1	2.26
2	1.91
3	1.99
4	1.98
5	1.92
6	1.89
7	1.91
8	1.89
9	1.93
10	2.01

Grading policy

- Below baseline (2.4): 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: 12/31/2018 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