In this project we prepare to do a user reliability score evaluate system. We have some problems need to discuss with you in data preparation stage.

***(Will it proper to use the formula below to do our data labelling work?)***

Step1 Data preparation

• Using yelp open source dataset(https://www.yelp.com/dataset) and do **data labelling** work.

**Explanation:** We have 16 attributes from our dataset as dictionary showing in appendix. Since we need to do score prediction but there's no score label in this dataset. We plan to calculate or labelling reliability score by using formula below:

$$Score = 1 - \frac{\sum_1^{RC} |RS_i - BAS|}{RC \cdot 4}$$

*RC* sum of one user's reviews

*RS score the user marked to each merchant*

*BAS* current score of each corresponding merchant

**This formula is aim to and more likely to calculate a average accuracy ratio.**

We can not just use the average score the user gave(average of RS) as the finally score. Because reliability can not just be baed on user-side comments, we should take the bias into consideration. You can see this formula in a more abstract way:

**Score = 1-average percentage error= average accuracy= reliability score**

Example: here is list of a user who gave 5 comments on yelp:

| ***RS*** | ***BAS*** |
|---|---|
| 4.5 | 5 |
| 4 | 4.5 |
| 5 | 4 |
| 3 | 4.5 |
| 5 | 4.5 |

1.We will get a score like: score=1-(0.5+0.5+1+1.5+0.5)/5*4=0.8

2.Take this score, multiply by 5( full score we can give or merchant can get on yelp), the average score this user gave to merchants is : 0.8 *5=4.

3. the reason why we divide a constant number 4 in this formula is that the biggest deviation value we could get is (5-1) or(1-5) ,which absolute value of this deviation is 4.

## Appendix:

attribute dictionary

| Attribute | Attribute Description |
|---|---|
| *user_id* | unique user id marked by 22 digits |
| *name* | first name of user |
| *review_count* | the total number of reviews they've written |
| *yelping_since* | years since user joined yelp |
| *friends* | list of user's friends marked by user_id |
| *useful* | number of useful votes sent by the user |
| *funny* | number of funny votes sent by the user |
| *cool* | number of cool votes sent by the user |
| *fans* | number of fans the user has |
| *elite* | the years the user was elite |
| *average_stars* | average rating of all reviews |
| *compliment_hot* | number of hot compliments received by the user |
| *compliment_more* | number of more compliments received by the user |
| *compliment_propfile* | number of profile compliments received by the user |
| *compliment_cute* | number of cute compliments received by the user |
| *compliment_list* | number of list compliments received by the user |