**Data Extraction steps to create user-friend-review benchmark Dataset from Yelp Data set**

1. The Dataset was created using review and user JSON files from Yelp Dataset(<https://www.yelp.com/dataset>)
2. In review file extracted the columns user\_id and text(review text) and removed all the users who has less that 50 reviews.
3. In user file extracted columns user\_id and friends and removed all the user\_ids which are not in the extracted user\_id list from review file.
4. Removed all the user\_ids who has less than 100 friends from step 3 extracted data
5. Removed the friends for each user who has less than 50 reviews or no reviews
6. Again, removed all the user\_ids who has less than 100 friends after step 5
7. The above steps resulted a list of 1054 target users with each user having more than 100 friends and both target users and friends having more than 50 reviews
8. Consolidated all the reviews for each target users
9. Created a separate list for friends
10. Consolidated 10 reviews for each friend
11. Combined the friends reviews for each target user( 10 review per friend according to step 9)
12. Target users had an average of 260 reviews and 220 friends with each friend having more than 50 reviews

**Dataset 0** -**(Ground truth):** Five user personality trait score for each selected target user.

Data used for this dataset is 1054 target users and consolidated reviews of each target user. IBM Watson Personality Insights API was used for ground truth deduction and consolidated review for each user was used as text input to the API . Initial output was JSON file containing all the personality data for each user. Extracted the raw and normal value of each of the five-personality trait from the JSON file and saved as dataset 0. JSON file for each user was saved separately.

**Dataset 1 (Users only + limited social footprints):** This dataset contains each target user and her selected 3 reviews.

Selected 3 reviews for each target user and combined them as review data for each target users.

**Dataset 2 (Users only + sufficient social footprints):** This dataset contains each target user and ALL her own reviews.

**Dataset 3 (Users + limited social footprints + friends’ footprints):** This dataset contains each target user and her selected 3 reviews, together with ALL the reviews posted by all her friends.

Selected 10 reviews for each friend and combined them. Consolidated reviews from all the friends for each target user. Combined 3 selected reviews from target user and consolidated friends reviews of each target user and saved as Dataset 3.

Eg: A target user with 100 friends would have 3 + (100\*10) =1003 reviews combined