**TraitSpot: Improving Personality Detection using Social Relationships in Latent Neural Embedding Space**

Haiquan Chen and Xiao Qin

**1. Goal.**

The goal of project is to investigate whether the footprints of social friends can improve the detection of the personality traits.

**2. Roadmap**

**Stage 1:** Extract the social networks from yelp.

* Datasets

<https://www.yelp.com/dataset/challenge>

**Any other large dataset with both social relationship and social posts?**

* Methodologies:
  + Distribution of the number of friends per user
  + Distribution of the number of reviews per user
  + First Cut: Remove all the users with less than 50 reviews
  + Second Cut: For all the remaining users, keep users with **at least 100 friends**
  + How many user are still there?

**Stage 2:** Fetch thepersonality trait ground truth for each remaining user using API <https://www.ibm.com/watson/services/personality-insights/>, based on the Big Five personality traits:

* Openness.
* Conscientiousness.
* Extraversion.
* Agreeableness.
* Neuroticism.
* Deliverables: The five separate ground truth datasets for the above trains

**Stage 3: Training dataset creation for personality trait detection**

* **Dataset 0 (Ground truth):** Five user personality trait score for each selected target user. Query the public API by using all the reviews in the original dataset from each target user
* **Dataset 1 (Users only + limited social footprints):** This dataset should contain each target user and her selected 3 reviews.
* **Dataset 2 (Users only + sufficient social footprints):** This dataset should contain each target user and ALL her own reviews.
* **Dataset 3 (Users + limited social footprints + friends’ footprints):** This dataset should contain each target user and her selected 3 reviews, together with ALL the reviews posted by all her friends.

**Here we train model using both each user’s and her friends’ posts.**

**Stage 4: Model implementation**

* Methodologies:
  + NN + TFIDF
  + SVM + TFIDF
* Metrics:
  + Compare the RMSE and R2 score for each model on each training dataset