**A Local Business Popularity Prediction System**

**Investigators:** Peng Zhao, Danlu Liu (PZ and DL are both responsible for all steps)

**Problem & Significance:** Yelp is a popular local business review publisher. For many people it is the first site to visit before making a dining or shopping decision. Although Yelp has a large amount of local business review data, users are only allowed to rank businesses by the relevance to search keywords, total number of reviews, or star-rating. Ranking by total number of reviews can be risky because a bad business may ranked top by receiving a lot of complaint reviews. Ranking by star-rating is also not a good idea because it can be biased if a business receives very few number of reviews. A smarter way to find a good business is in need. What’s more, popular businesses must have their reasons. It is important to identify the key elements for popularity so that unpopular businesses can improve themselves accordingly.

**Dataset & Source:** Yelp Dataset Challenge<https://www.yelp.com/dataset_challenge>

**Approach:** (1) Build a supervised business popularity prediction model based on a new popularity metric and display the predicted popular local businesses to users. (2) Identify frequent patterns of attributes in more popular businesses with clustering and data mining techniques, such as frequent pattern mining.

**Novelty:** (1) We will create a new popularity metric by combining the number of reviews and star-rating. It can overcome the shortcomings of these 2 existing evaluation methods. (2) We will use big data platform to handle large scale computation. (3) We will find the key elements for popularity and recommend them to unpopular businesses.

**Procedure:**

1. Integrate and preprocess data (missing value handling, data normalization, feature categorization, feature selection and dimension reduction, etc.)
2. Create a new popularity metric by combining the number of reviews and star-rating. This will be calculated for every business and will be used as the class value in prediction.
3. Train multiple supervised prediction models (Bayesian, KNN, decision tree, etc) and choose the best one.
4. Recommend good businesses to users based on their location (by ranking the predicted popularity scores for local businesses) and display the results using a user-friendly data visualization on a map.
5. Use clustering and data mining techniques such as frequent pattern mining to find the frequent patterns of attributes in more popular businesses.

**Timeline:** (1) Preprocess data till 4/15. (2) Train different prediction models till 4/30. (3) Perform clustering and frequent pattern mining on early May. (4) Create data visualization and write reports in the end.