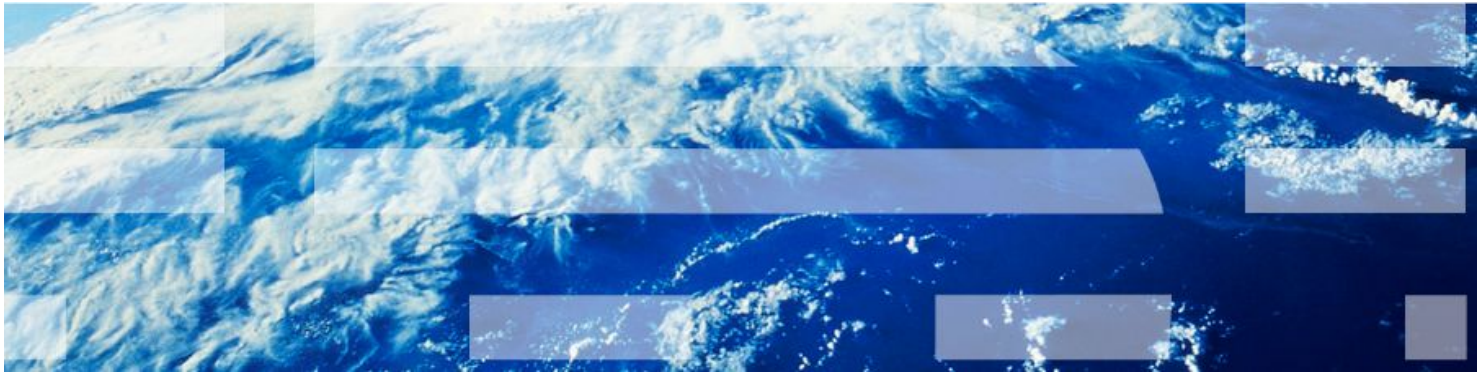
E6893 Big Data Analytics [201812-46]

Yelp Rating Interpretation with Text-based and Graph-based features

Zhuoran Liu (zl2621)
Mingye Chen (mc4414)



Motivation

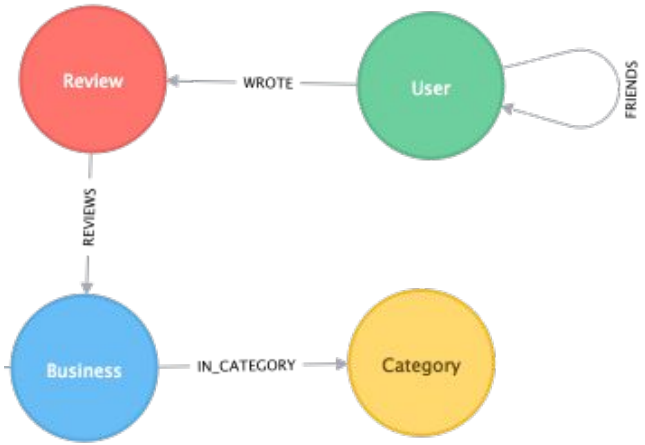
- Problem we want to solve:
  Given specific user community and/or specific type of restaurant, find factors significantly contribute to the rating?

e.g. why Chinese community in CU thinks Shake Shack is a good place for fast food?
possible answer: friend recommendation (Graph Feature), wholesome (Review Text)

- Rating polarity predictions with interpretability (positive & negative review)
- Features to be explored:
  - User Communities Detection
  - User & Restaurant Graph
  - Review Texts Mining
  - ……

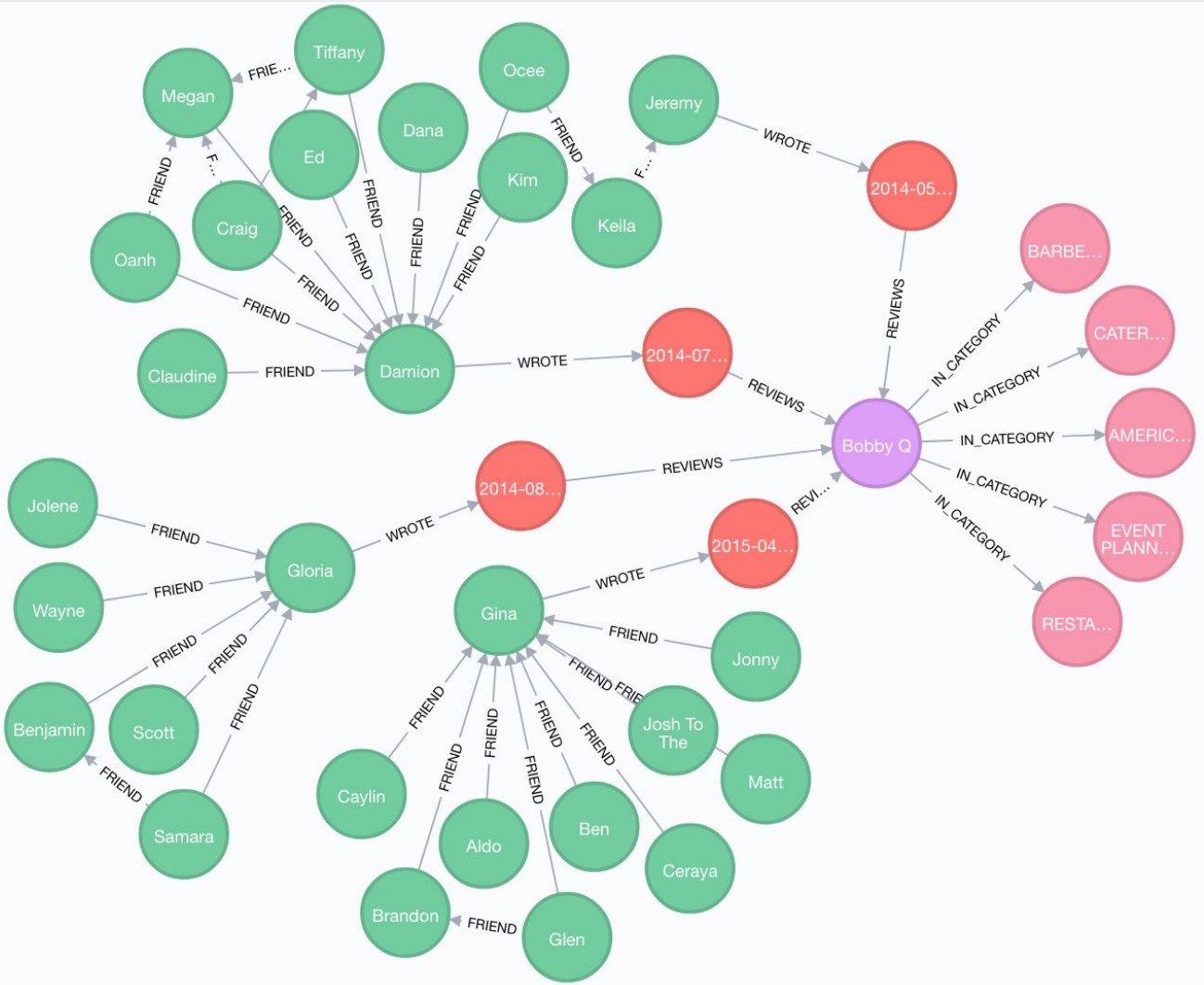
Dataset, Algorithm, and Tools

- Yelp dataset challenge (Round 12), we select restaurants of Phoenix, AZ
- 376,172 reviews, 142,286 users, 3,833 restaurants
- Graph algorithms: user community detection, user-restaurant path finding, ...
- Text features engineering: n-gram model of review texts (bag of words) , …
- Random forest for rating polarity prediction, with text & graph features
- We can get feature importances by mean decrease impurity (gini importance)
- Neo4j, sklearn, ...



Features come with datasets

- User: elite endorsement, avg star
- Restaurant: zip code, avg star, restaurant category, city
- Review: raw text, stars -> polarity (1, 2, 3: negative; 4, 5: positive)

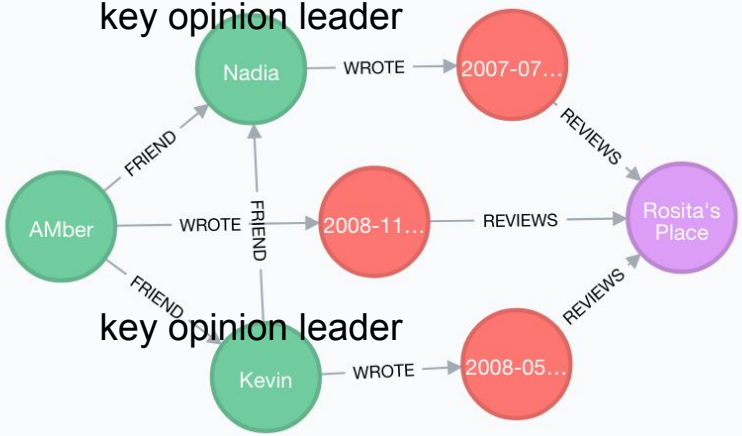


Graph Intro & Meta Model

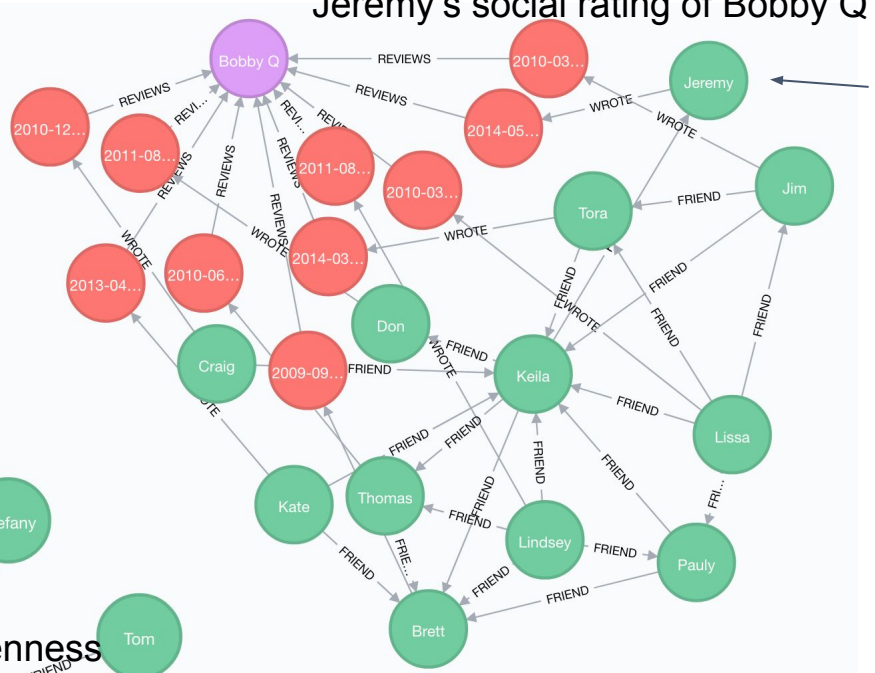
Graph features

- User Community: Louvain algorithm (maximizes a modularity for communities)
- User Centrality: PageRank, Betweenness, Closeness
- Key opinion leader: top 500 PageRank users
- Restaurant favored by different community (top 17 communities):
  users in community X's average rating of the restaurant more than user's total average rating in community X.
- Review:
  (1) any key opinion leader followed by this user wrote reviews to this restaurant
  (2) average social rating (1st & 2nd degree social circle reviewed this restaurant)

key opinion leader

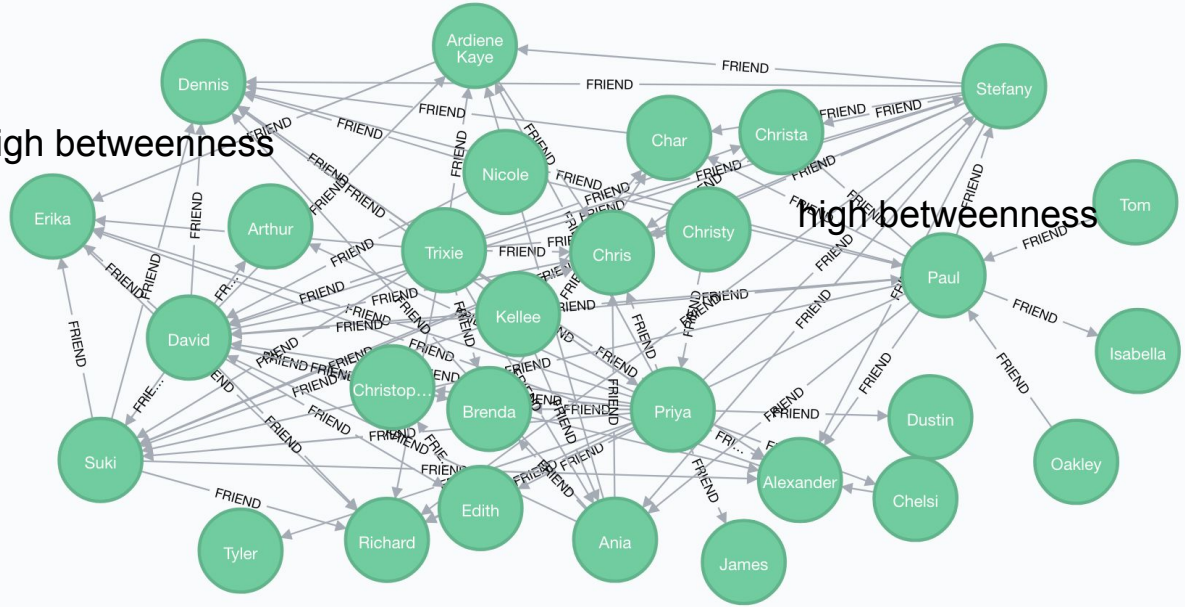


Jeremy's social rating of Bobby Q



high betweenness

high betweenness



Text Features

- Interpretability-performance trade-off: catch-phrase based features
- Catch Phrases: frequently recurring phrases / ngrams signifies polarity w.r.t. certain aspects of the restaurant
- E.g. ‘delicious’
  - 65969 occurrence;
  - positive on food;
- E.g. ‘attentive’
  - 12817 occurrence;
  - positive on service
- E.g. ‘reasonable prices’
  - 1427 occurrence;
  - positive on price;
- Strong signal for polarity classification

Text Features

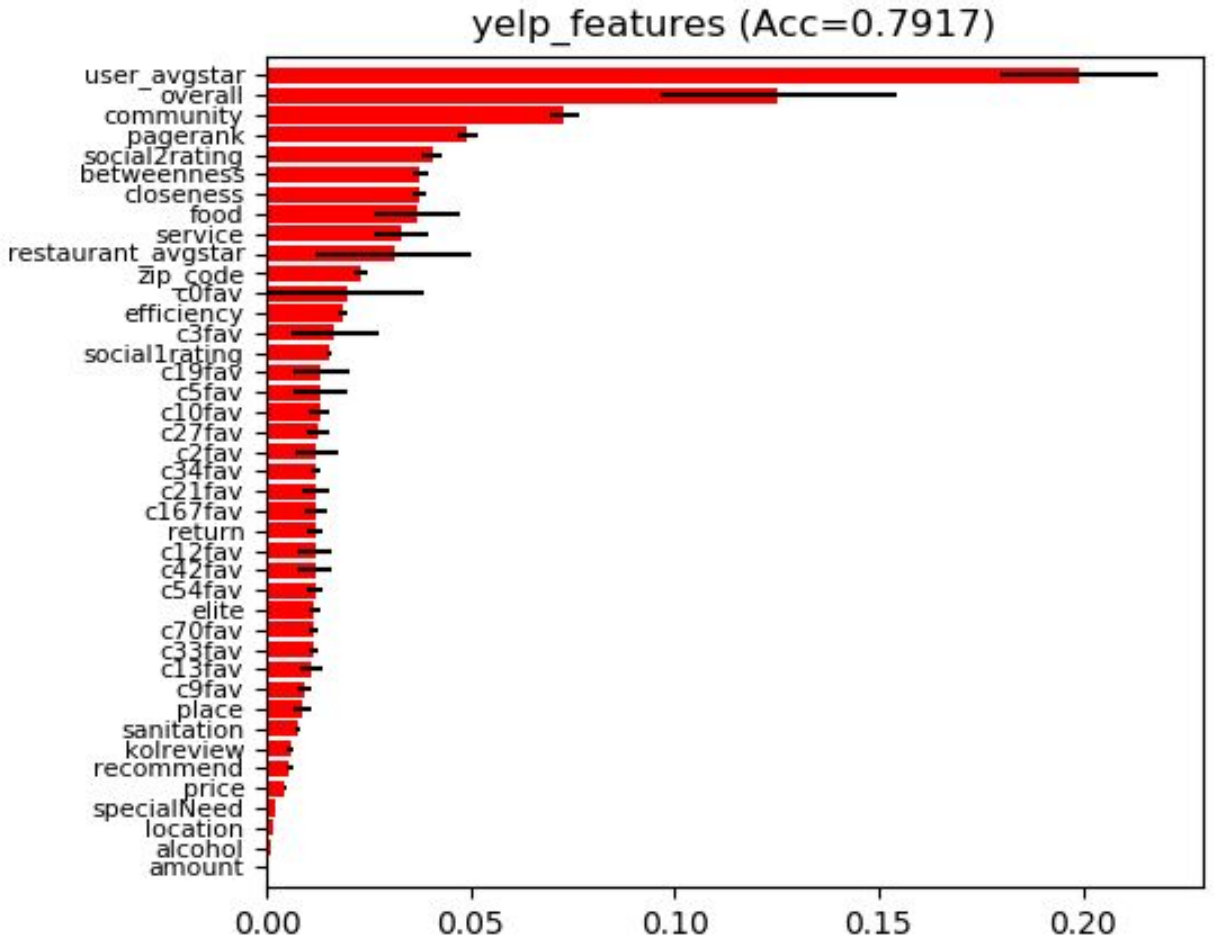
- Preprocessing: lowercasing / tokenizing / stopwords removal
- Manually selected catch-phrases top-ranking uni-/bi-/tri- grams
- Covered most reviews
- Catch phrases classified into 13 categories
  resulting in 13 features
  - food overall
  - special need
  - alcohol
  - food amount
  - service overall
  - service efficiency
  - would return
  - would recommend
  - restaurant overall
  - location
  - place
  - price
  - sanitation

Experiments

- Multiple groups of experiments
- Do different factors weigh differently in different communities / restaurants types?
- Extracted subsets of reviews by
  - identified communities
  - restaurant categories
- 44 subsets + full data → **45 groups** of experiments
- Results are interesting
- We selected several groups of experiments below

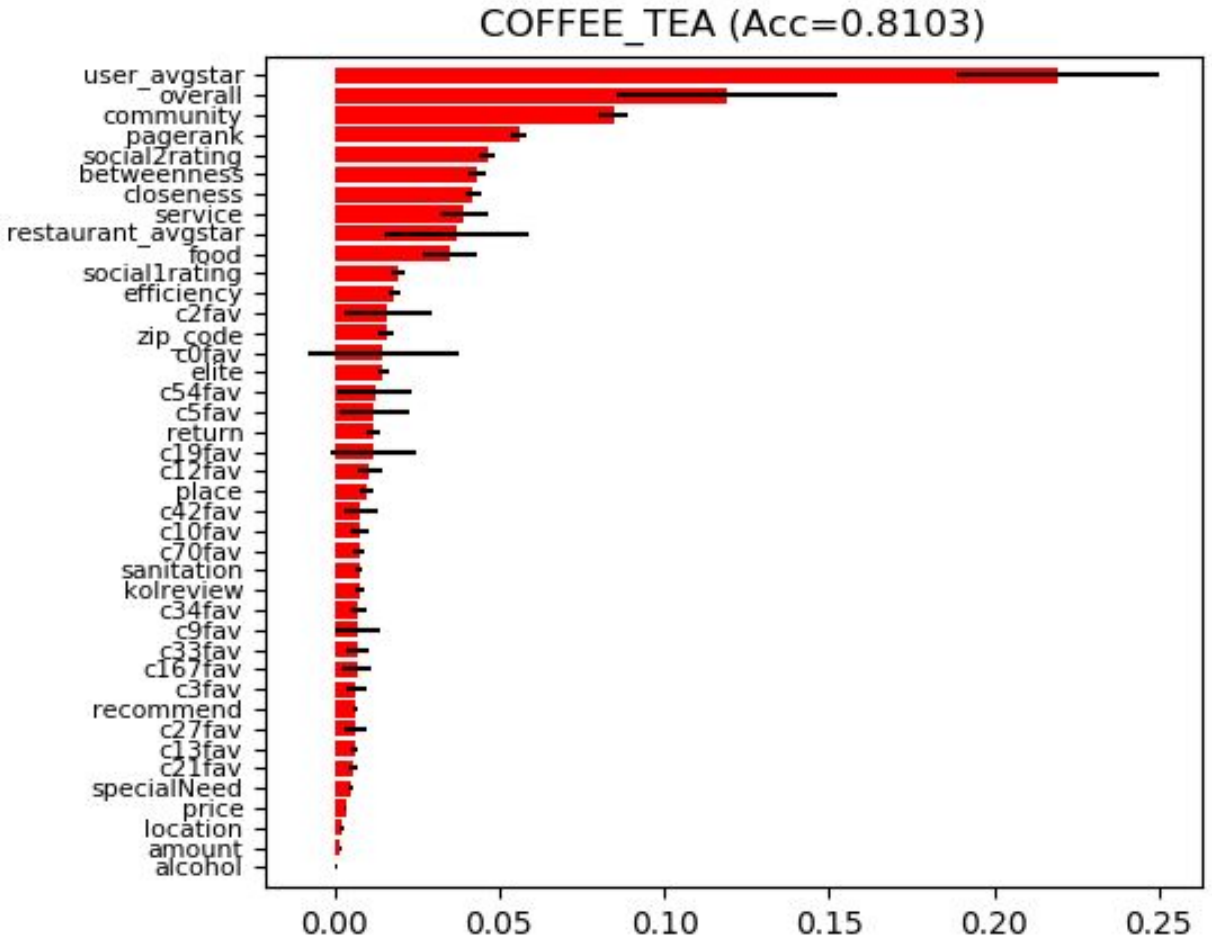
Ex. 1

- Full yelp dataset
- user_avgstar is used as baseline
- Conclusion: Users' criteria on number of stars vary greatly



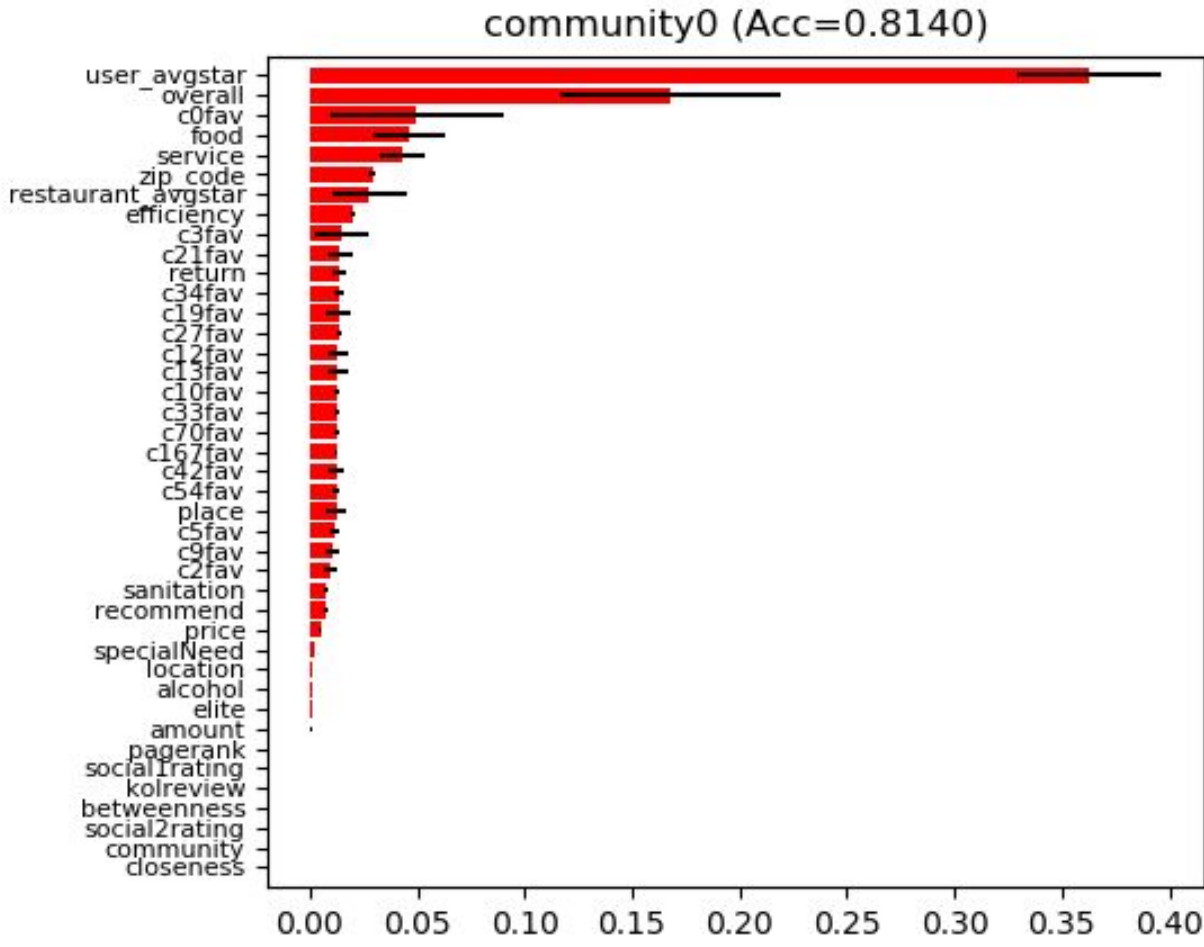
Ex. 2

- Coffee-tea restaurants
- Community/social features are important
- Conclusion: People's preference on coffee / tea are sensitive to user group

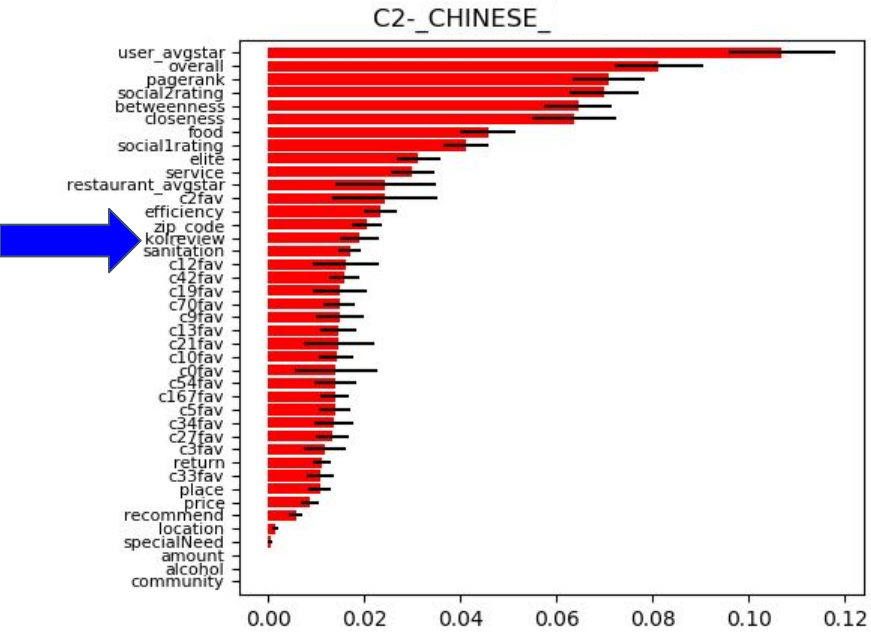


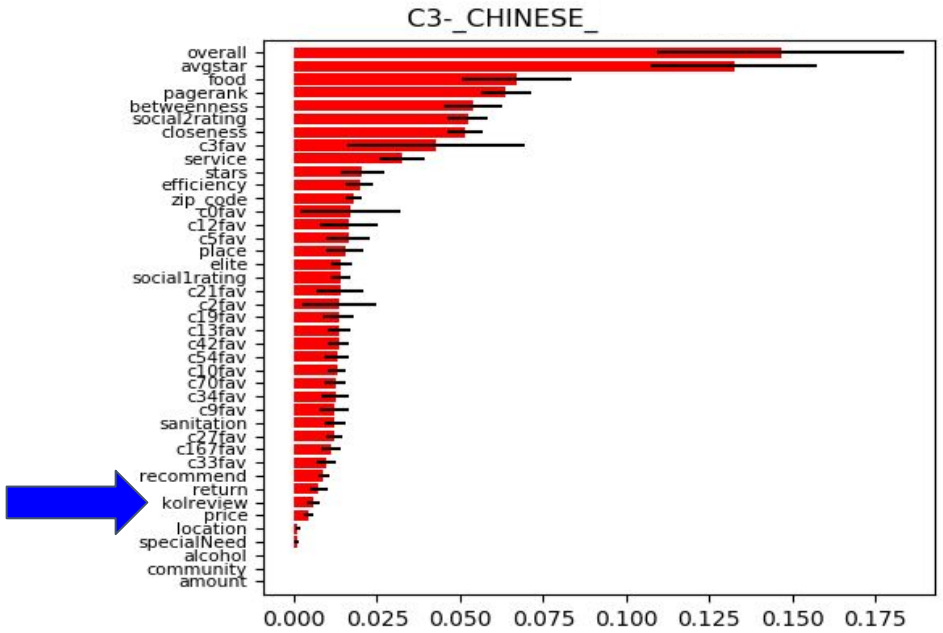
Ex. 3

- Community 0
  - mixed user group with large number of members
- The following features are dominating:
  - restaurant overall
  - food overall
  - service overall
  - service efficiency
- Conclusion:
  In mixed user group, ratings are sensitive to judgement on intrinsic qualities of the restaurant



Ex. 4 KOL review





- Community 2 / 3 review to Chinese Restaurants
- Importance of key opinion leaders' introduction varies a lot
- Conclusion: KOL opinion may have different degree of influence in different communities

Conclusion

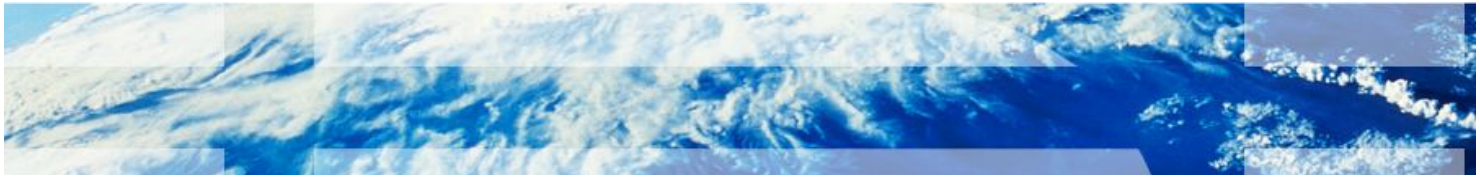
- Our classifier achieves ~80% accuracy on the full dataset;
- For different communities / types of restaurant, factors contribute different weight to final rating;
- Different users have different baseline rating scores;
- Restaurant owners can use our model to develop the customer strategy accommodating to different communities.

E6893 Big Data Analytics:

Yelp Rating Interpretation with Text-based and Graph-based features

Zhuoran Liu (zl2621)
Mingye Chen (mc4414)

Thanks!



Youtube link:

https://www.youtube.com/watch?v=NwBog99CCNg