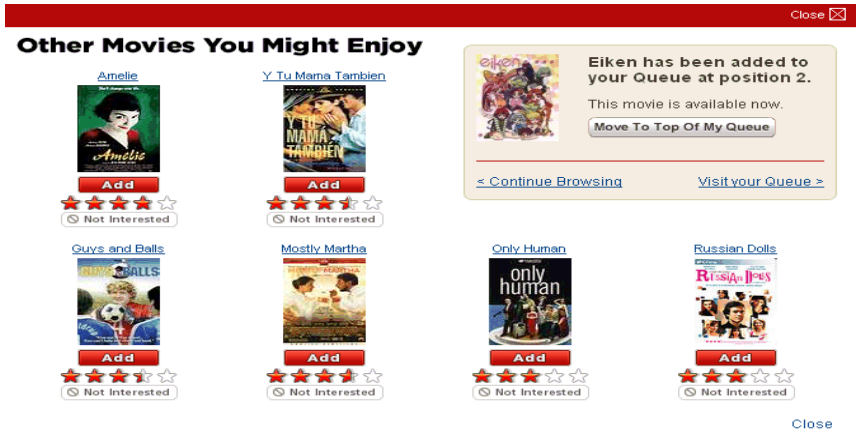
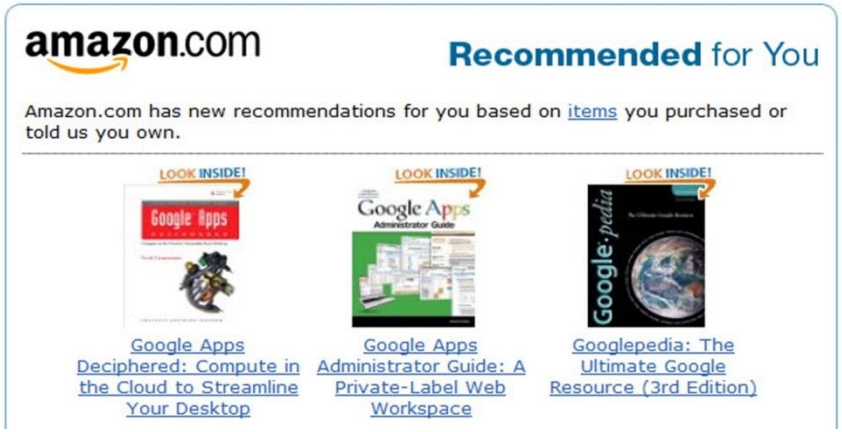
Yelp Recommender System:

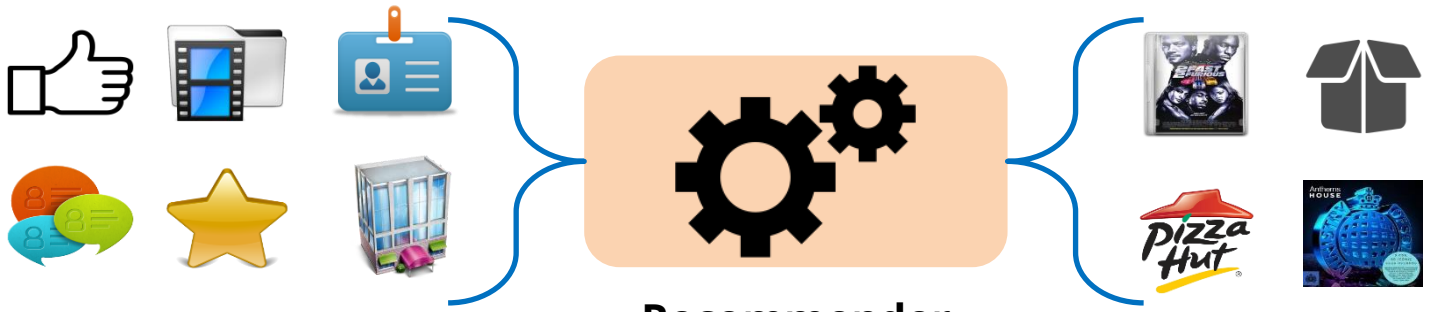
Connecting people with restaurants of their choice and taste

Reporter: Ildar Abdrashitov
Mentor: Nemanja Radojković

General overview



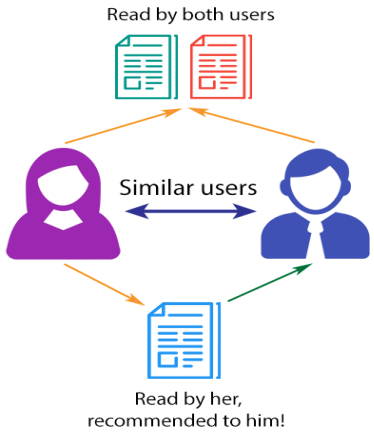




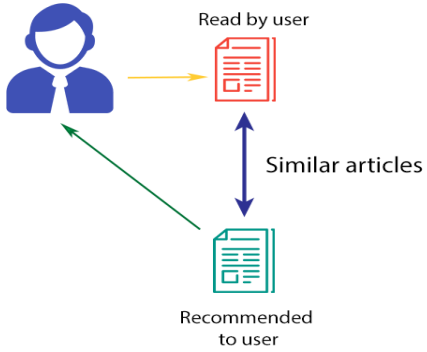
Spotify NETFLIX ebay

Recommendation techniques

- **Collaborative filtering**:

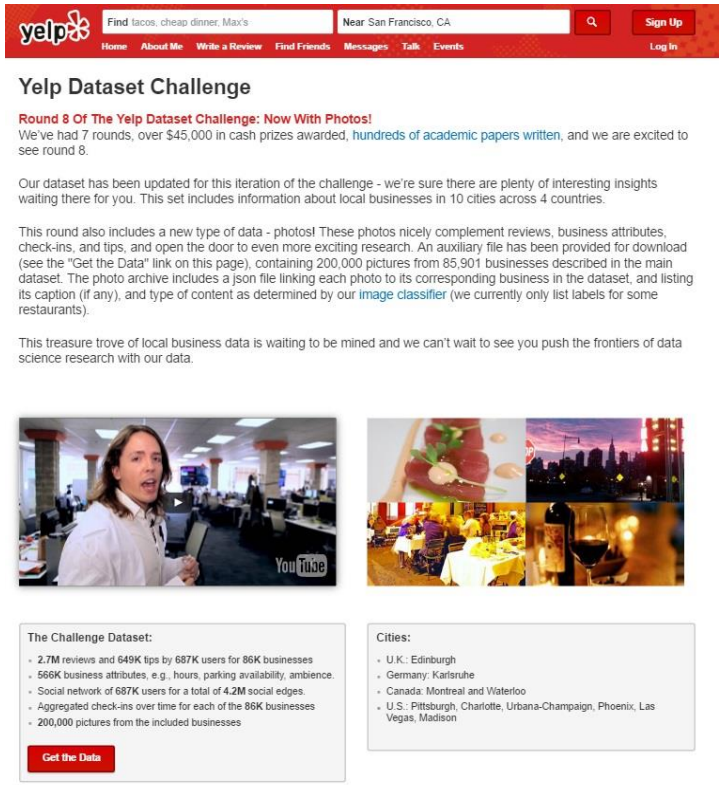


- **Content-based filtering**:



- **Hybrid: Collaborative + Content-based**

Data overview



5 Data Sets:

- **Businesses**
- **Check-ins**
- **Reviews**
- **Tips**
- **Users**

Data Format: JSON

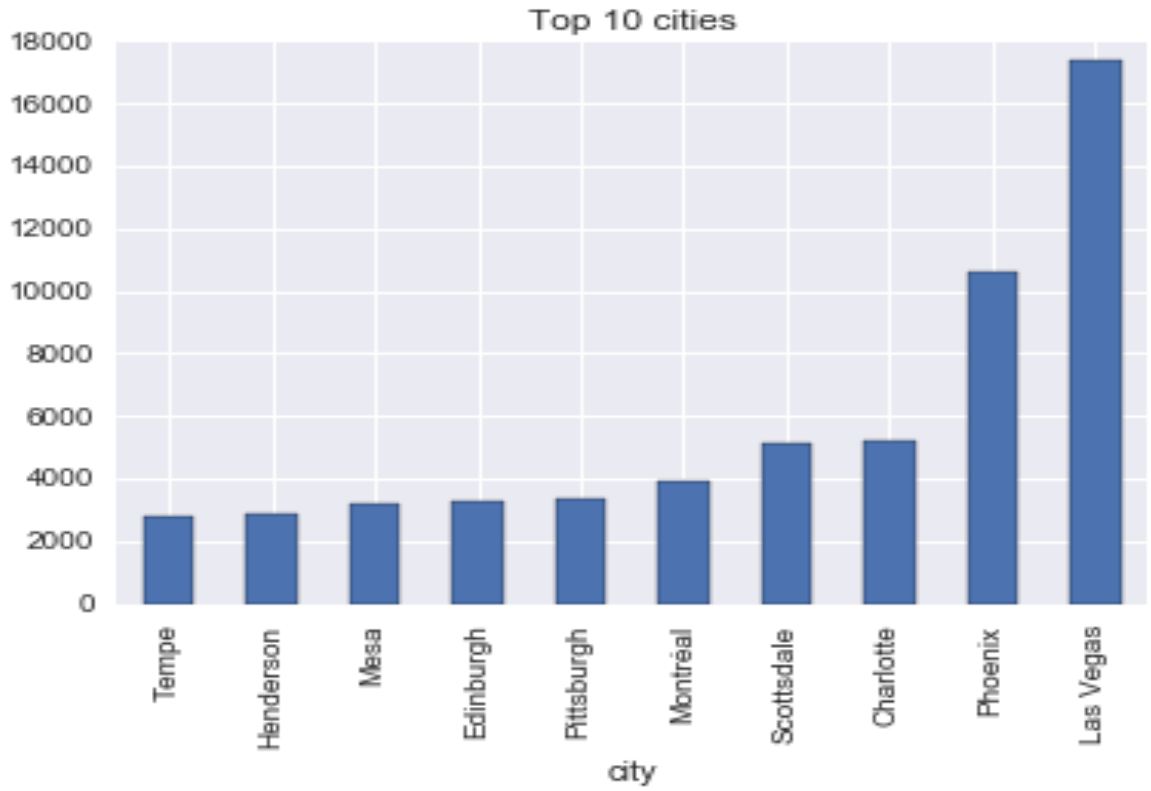
Overall Size: ~ 2.2 Gb

Data Source: https://www.yelp.com/dataset_challenge

Data: First look

Businesses:

- **77445 unique businesses**
- **412 cities**
- **31.6 avg. review count**
- **3.6 avg. stars**

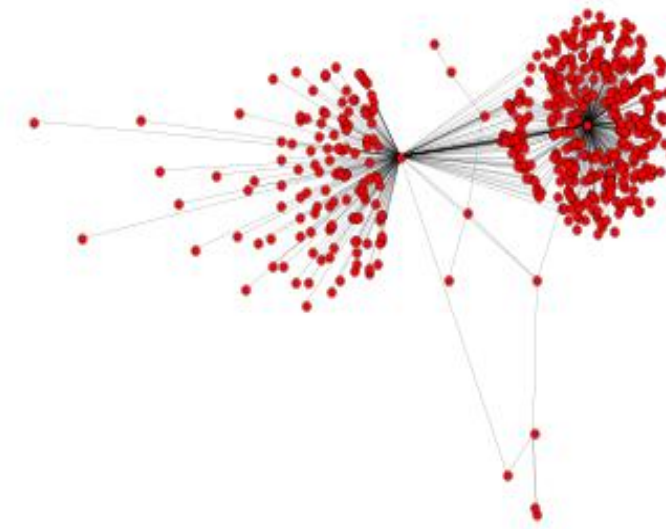


Data: First look

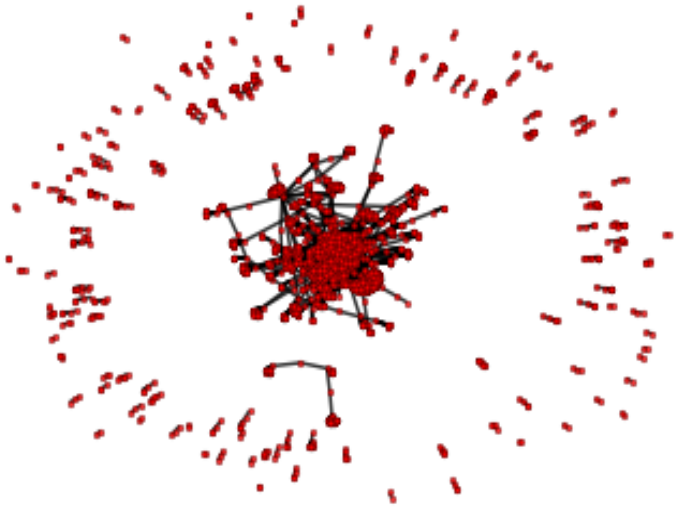
Users:

- 552 339 unique users
- 3.7 average stars
- 3 812 max n of friends
- 0 min n of friends (most common => 302 899 users)

5 users SN graph



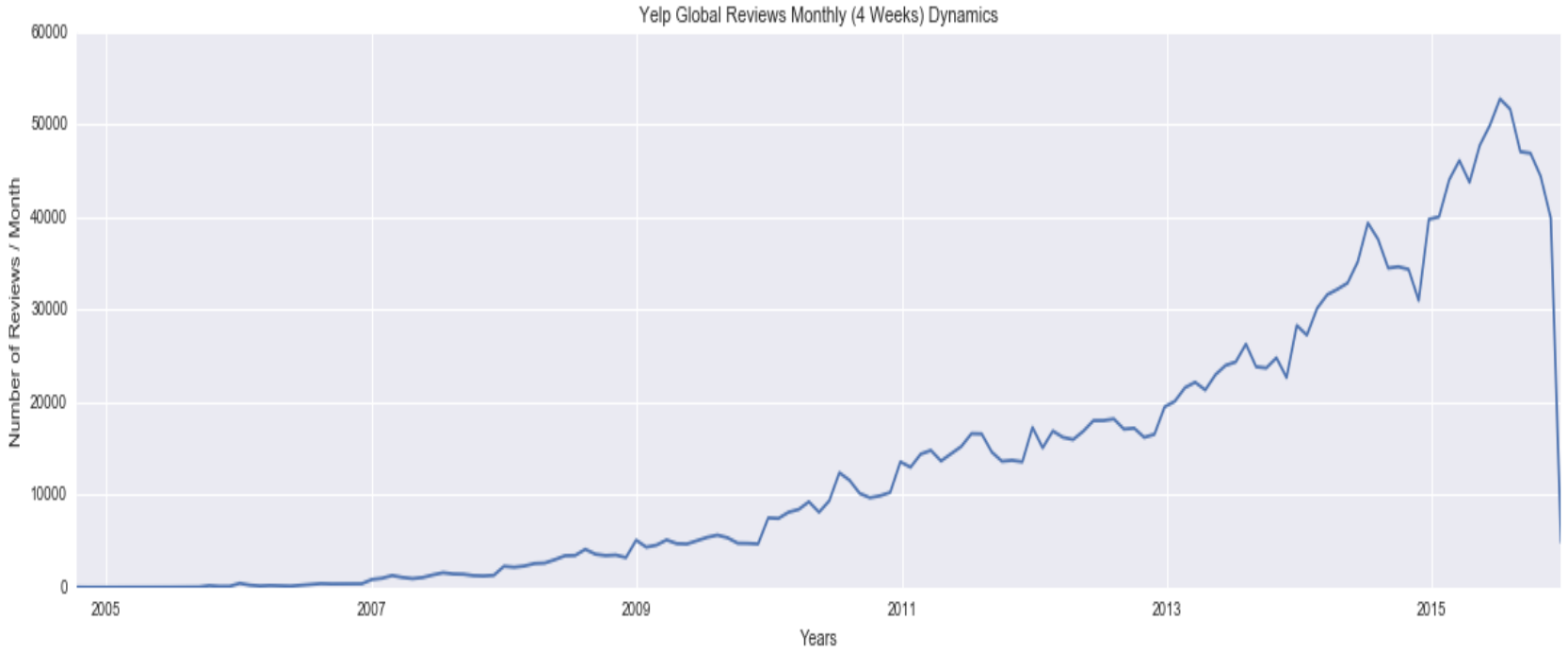
500 users SN graph



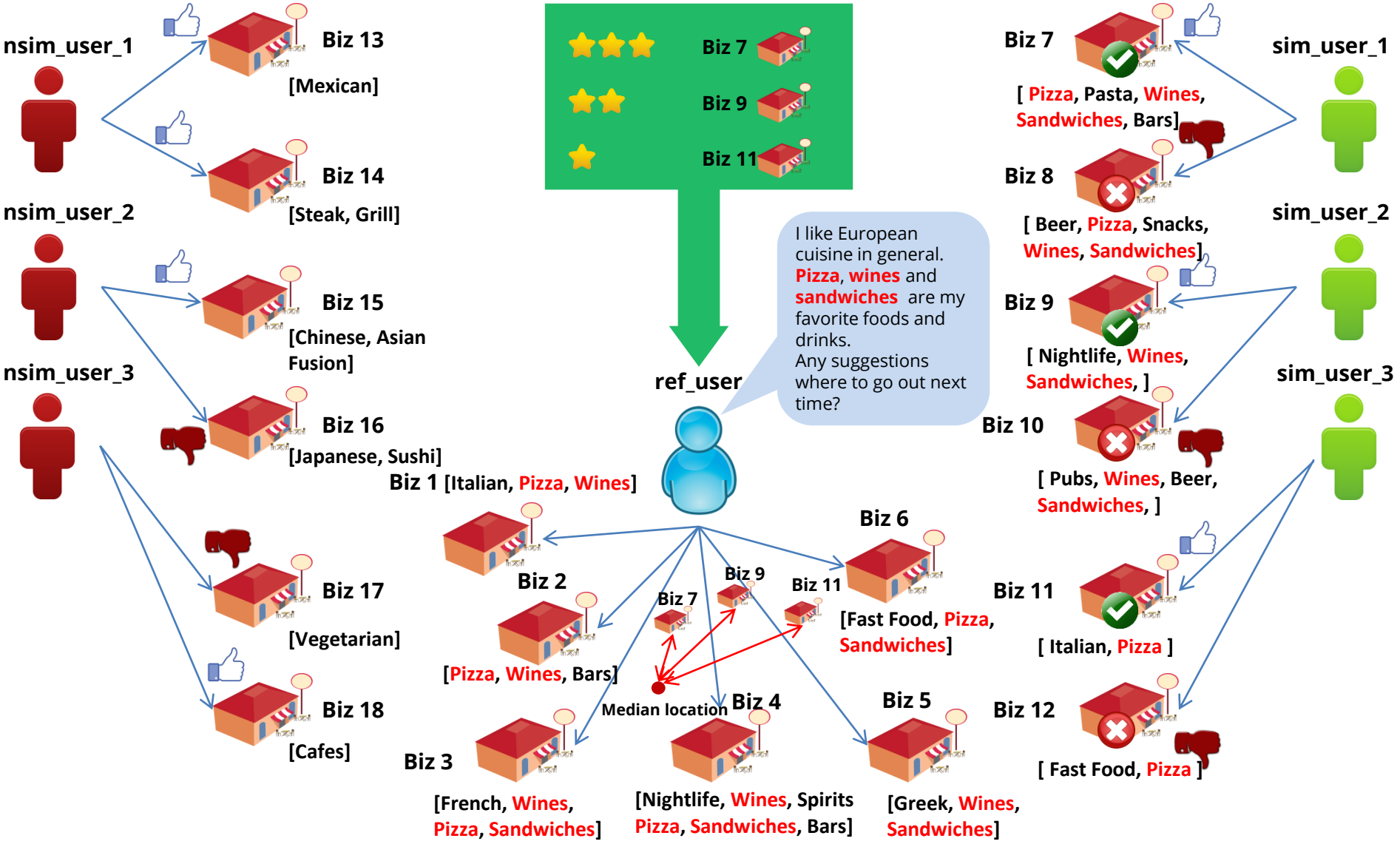
Data: First look

Reviews:

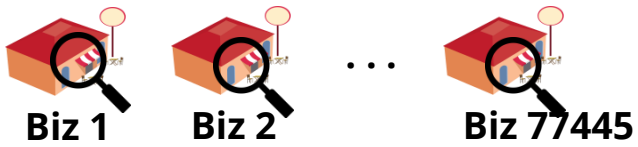
- 2 225 213 reviews
- 77 079 unique businesses
- 552 339 unique users
- 3.75 avg. stars



Recommender model overview



Learning users tastes and preferences



[category_1,
category_2,
category_3,
. . . ,
category_n]

User preference vector

user_1

2 reviews

10 stars max cumulative rating

user_1 user_2 . . . user_n

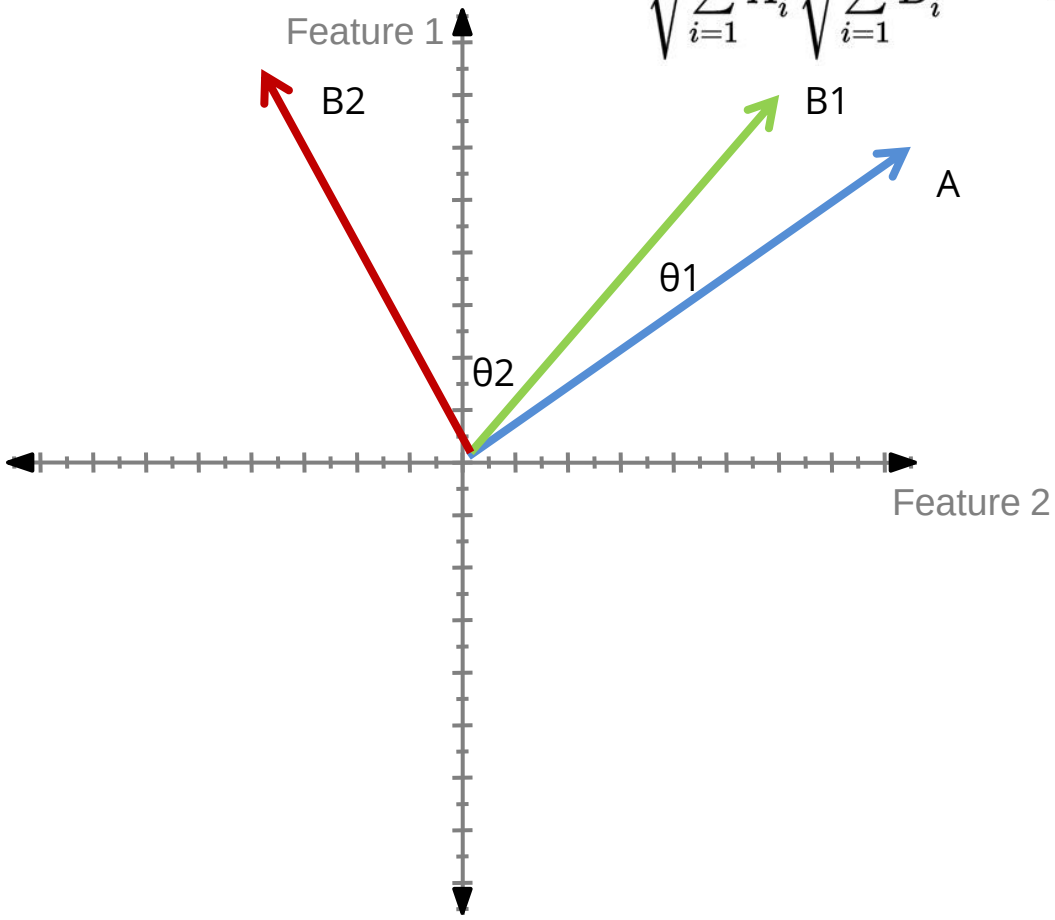
	Irish	Juice Bars & Smoothies	Korean	Cheese Shops	Arts & Entertainment	American	Peruvian	Shanghain
-0itF0VWVBe3k2AdfURe GA	0	0	0.012346	0	0	0.012336	0	0
-0p7UrjssdF5FvDKl5y6RA	0	0	0	0	0	0.072727	0	0
-0wUMy3vgInUD4S6KJInnw	0	0	0	0	0	0.044444	0	0
-19j4ExVKyIlJU8JcEV55Q	0	0	0	0	0	0.030769	0	0
-24OurJb1F03IWgo0EbMEv	0	0	0	0	0	0.063281	0	0
-26J4wzYrjU7IIQAkR1EWg	0	0.092261458	0	0	0	0.09157	0	0
-2EuoueswhqEERWezJY8gv	0	0	0	0	0.02	0.013556	0	0
-2PFmYGWB6rQhZtRh6PlPA	0.007712	0	0.012847	0	0.01025641	0.007712	0	0
-2RVxn8RRN3j9bOiqGkjug	0	0	0	0	0	0.1	0	0
-2gxrMHX8_F1Ep_ppWYTk	0	0	0	0	0.090909091	0.090909	0	0
-2iNr0vd-jCCDGcXW6lDdg	0.08	0	0	0	0	0.08	0	0
-2r-OO0dtcCQQ3BLYdHXlg	0	0	0	0	0	0.019501	0	0
-3AYE63sussmBAylWvDn0v	0	0	0	0	0	0.038023	0	0
-3HcO3CnxzhxdEW8Soplig	0	0	0.055926	0	0	0.044444	0	0
-3HkgwyxVPf49MXM4dT8T	0.028776	0	0	0	0.02877551	0.028777	0	0

Distinguishing similar users

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

$\mathbf{A}$, $\mathbf{B}$ - 2 users pref. vectors

A_i, B_i - components of the corresponding pref. vectors



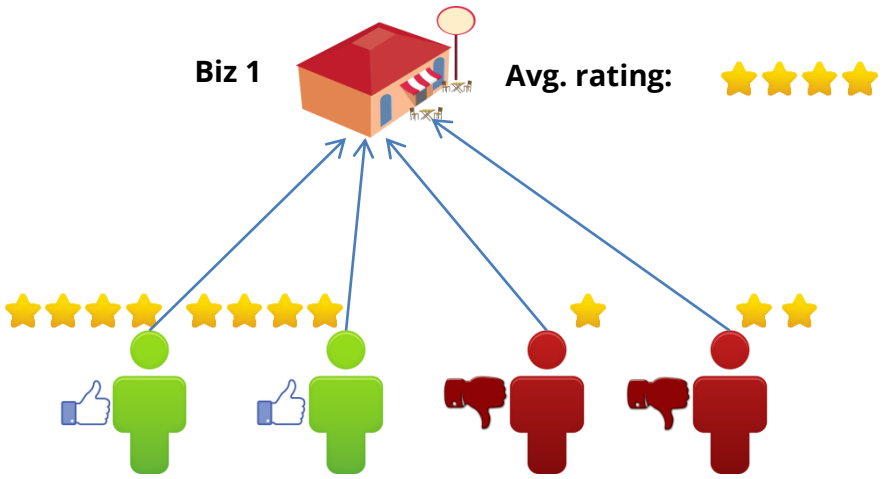
ref_user — ref_user`s pref. vector

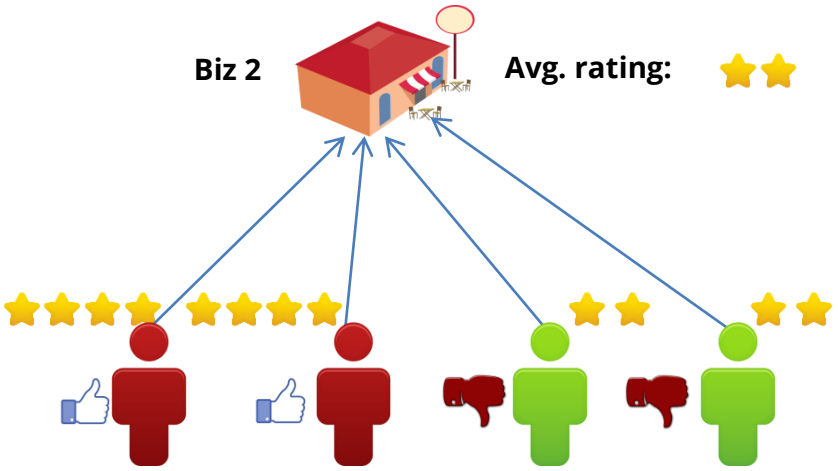
sim_user — sim_user`s pref. vector

nsim_user — nsim_user`s pref. vector

0 > = similarity < = 1

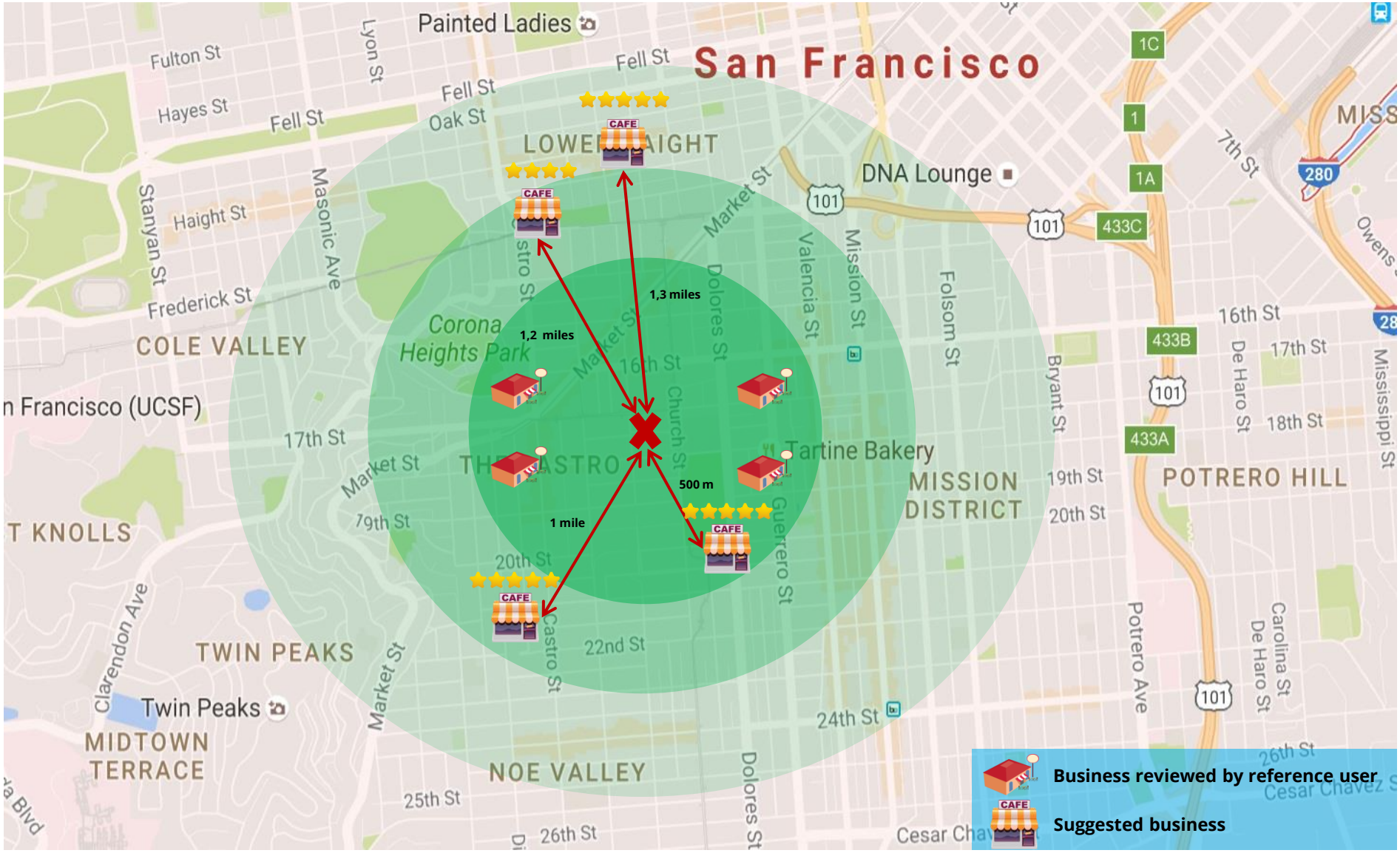
Business similarity weighted rating



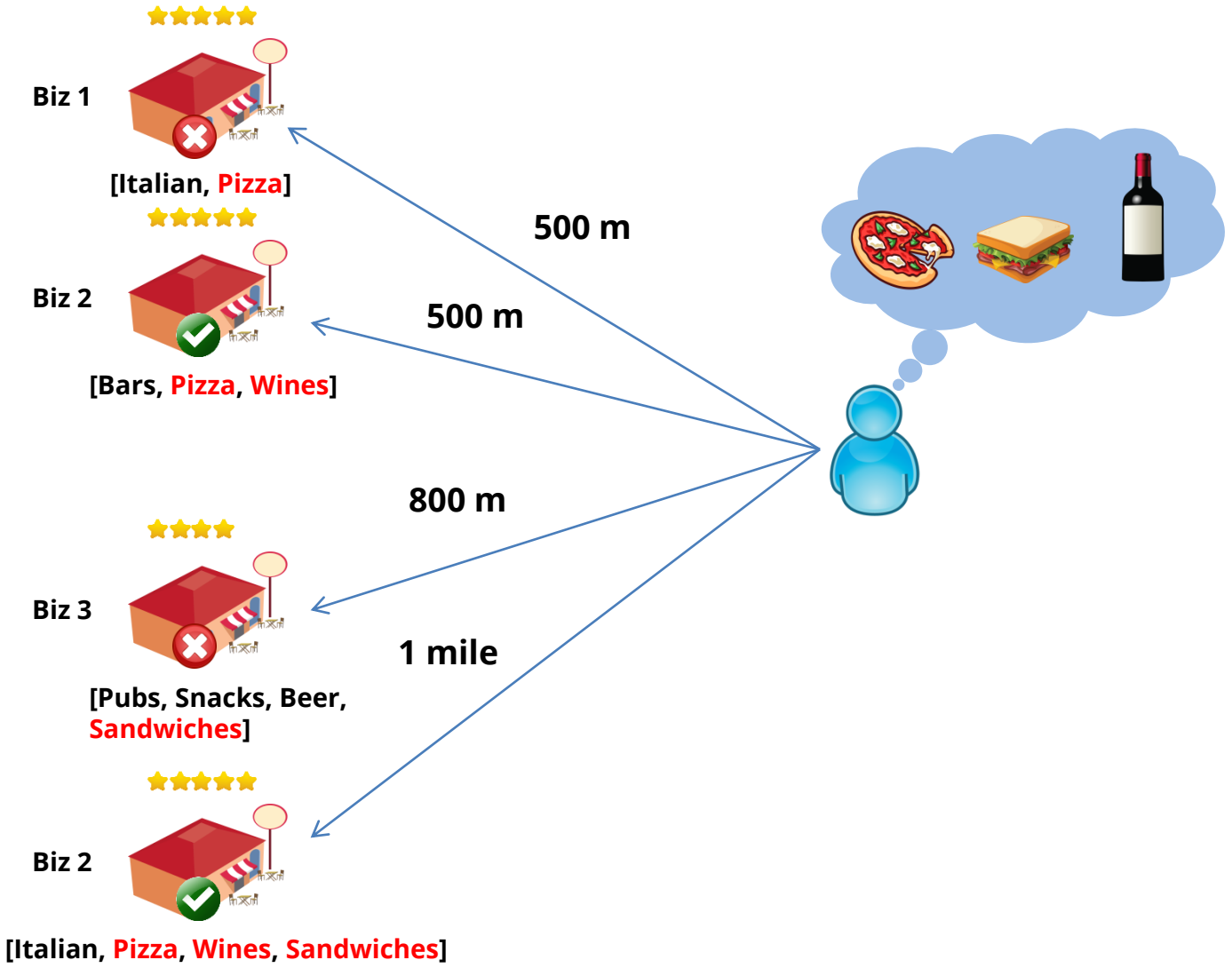


$$\frac{\sum_{i=1}^{n} rating_i \times similarity_i{}^2}{\sum_{i=1}^{n} similarity_i{}^2}$$

Business combination metric



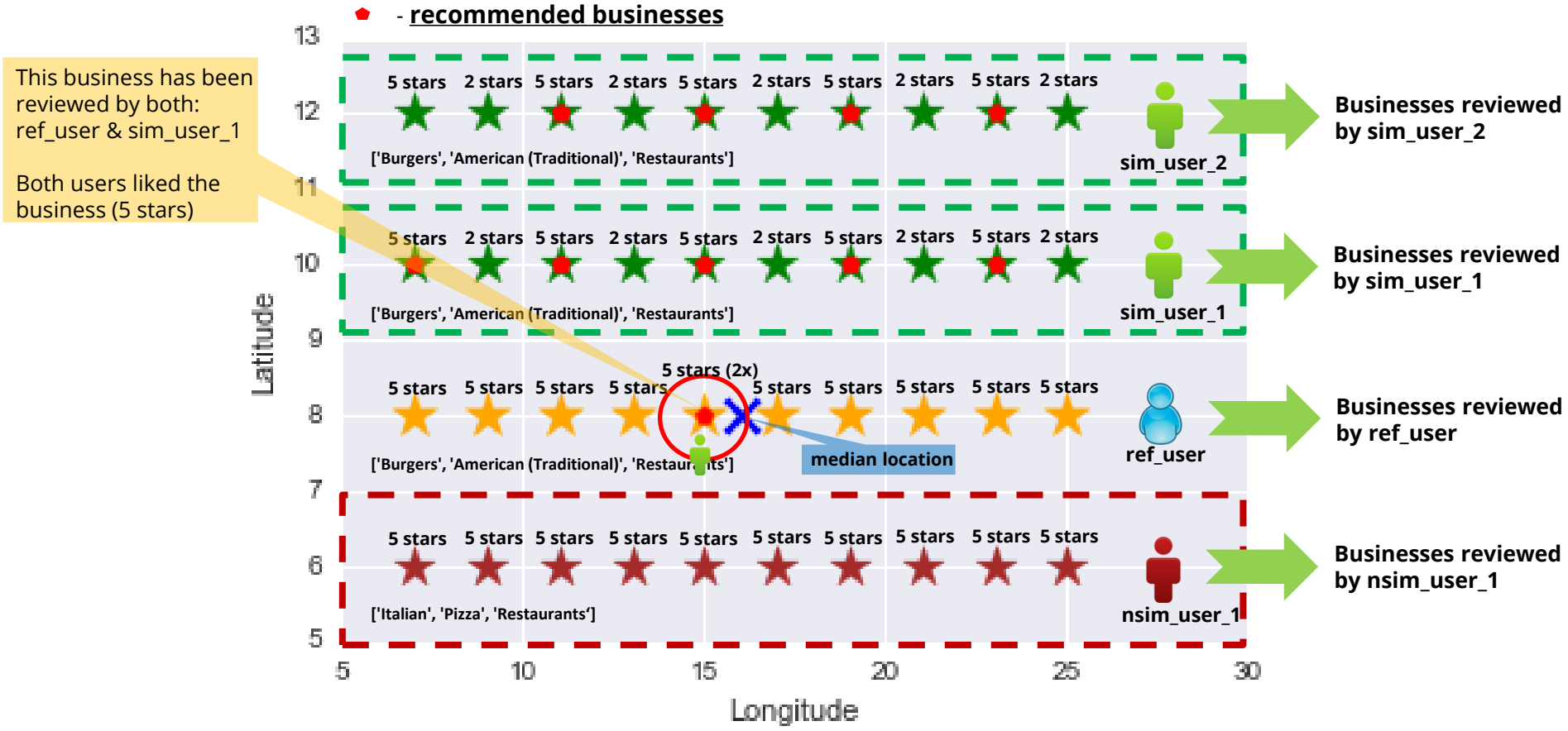
Business adjusted combination metric



Model testing

Test case 1: - 1 reference user
- 2 similar users
- 1 not similar user

- 40 businesses (30 same category)
- sim_users: 5 stars reviews count >> 11
- sim_users: 2 stars reviews count >> 10
- n_sim_user: 5 stars reviews count >> 10



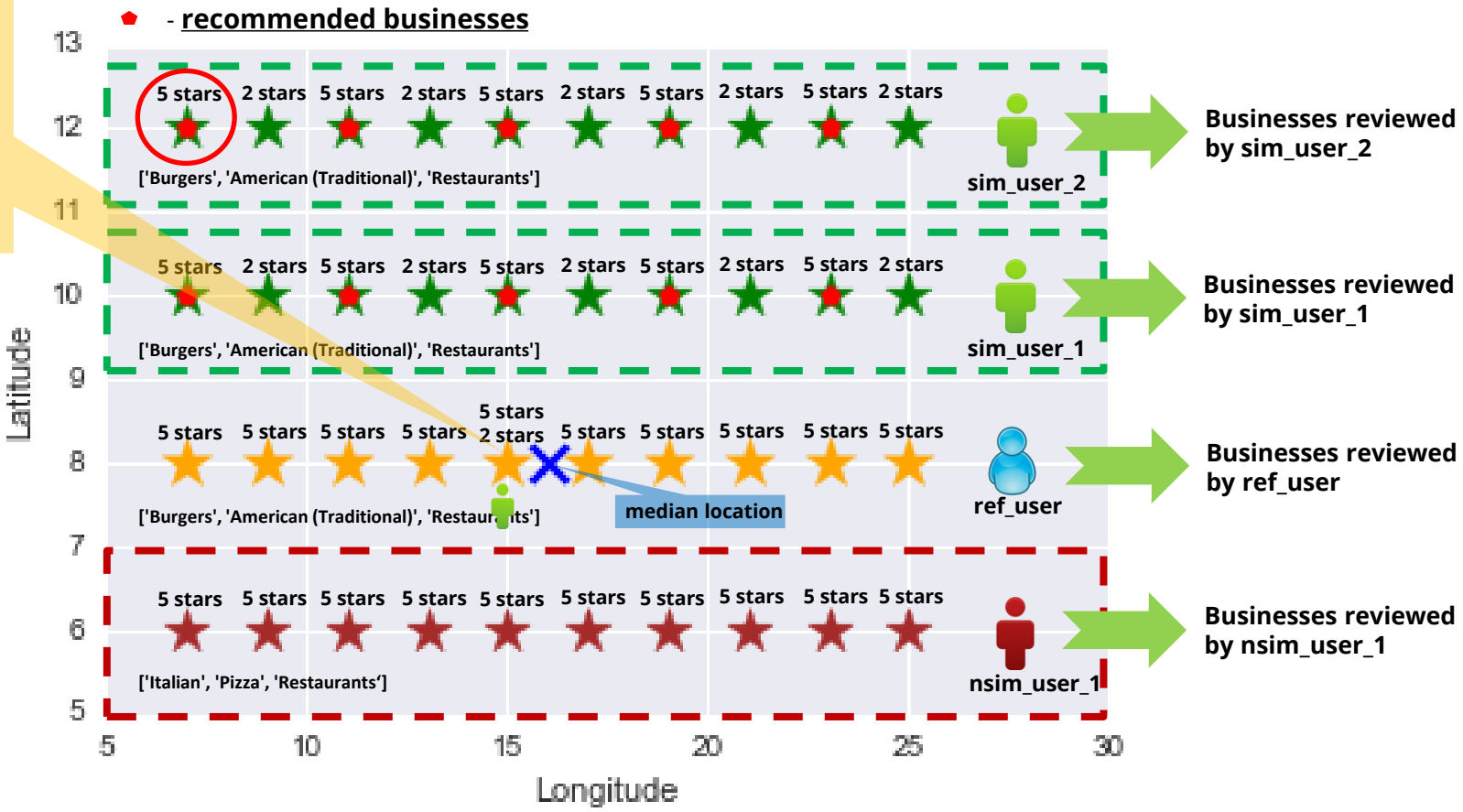
Model testing

Test case 2: - 1 reference user
- 2 similar users
- 1 not similar user

- 40 businesses (30 same category)
- sim_users: 5 stars reviews count >> 10
- sim_users: 2 stars reviews count >> 11
- n_sim_user: 5 stars reviews count >> 10

This business has been reviewed by both: ref_user & sim_user_1

However, this time sim_user_1 did not like the business (2 stars)



Model testing

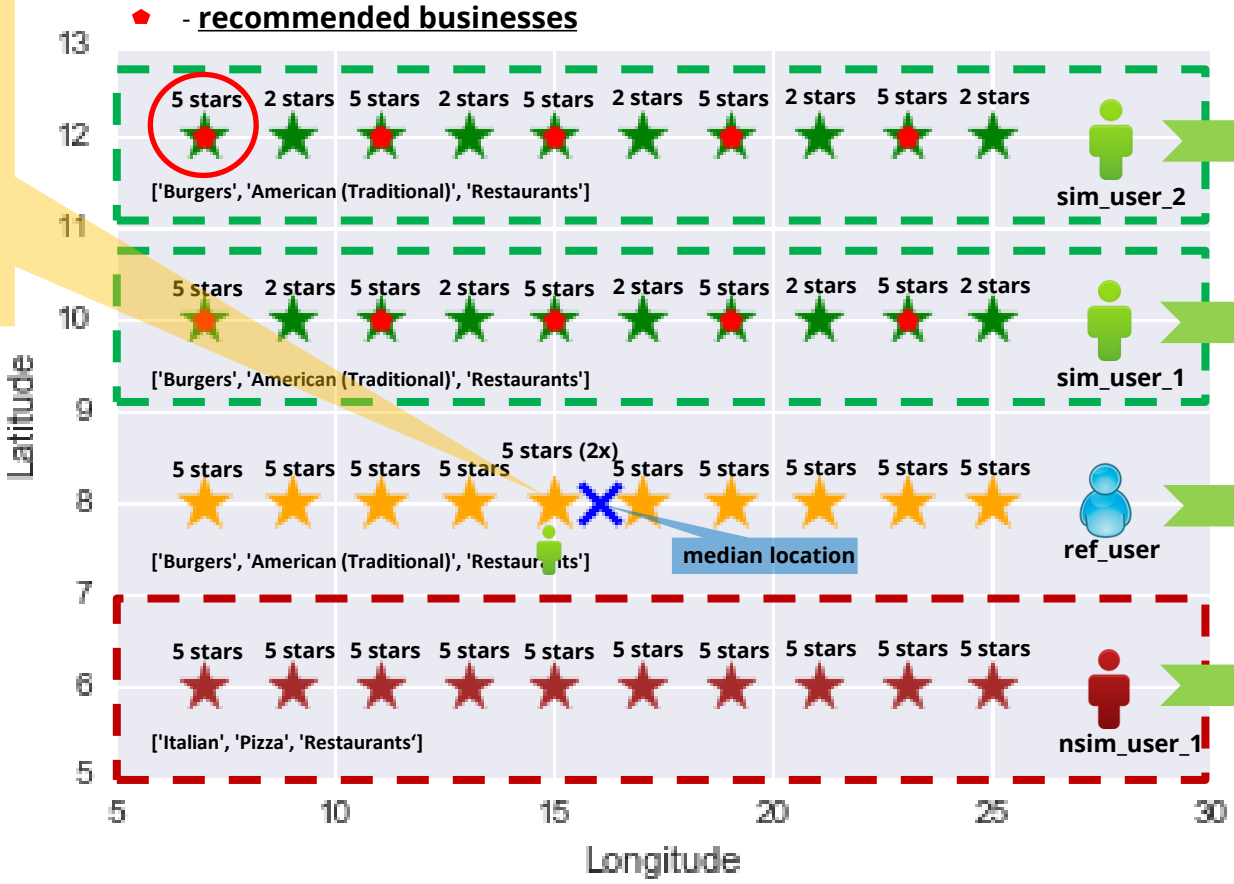
Test case 3: - 1 reference user
- 2 similar users
- 1 not similar user

- 40 businesses (30 same category)
- sim_users: 5 stars reviews count >> 11
- sim_users: 2 stars reviews count >> 10
- n_sim_user: 5 stars reviews count >> 10

This business has been reviewed by both: ref_user & sim_user_1

Both users liked the business (5 stars)

However, this business belongs only to ['Burgers'] category

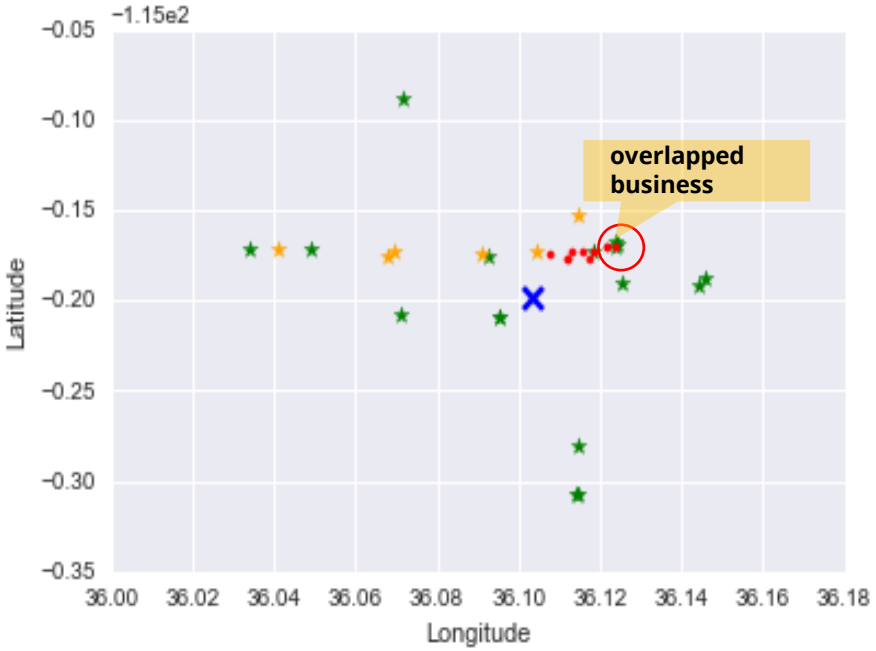


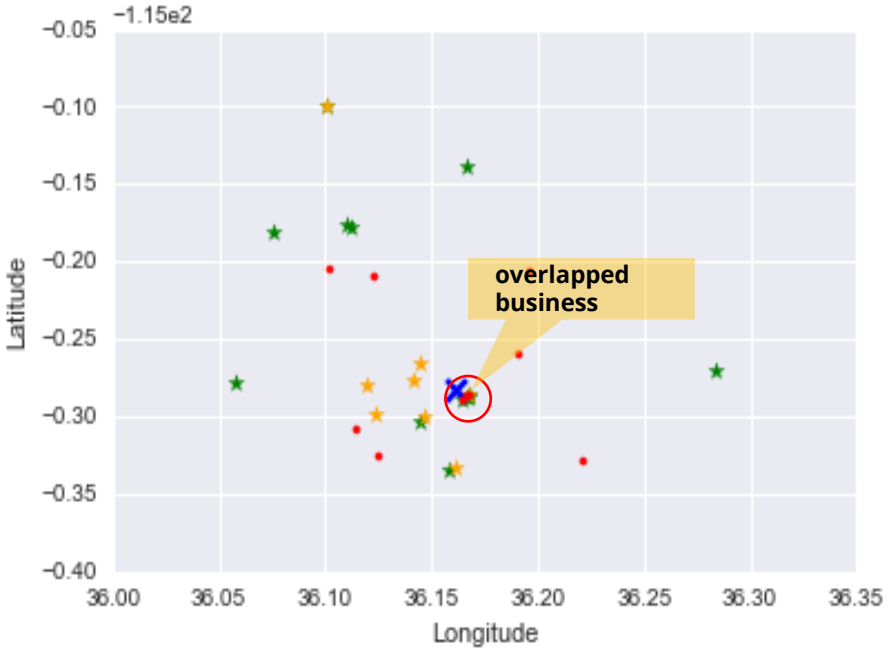
Running model on real data

Tipped business

Reviewed business

Recommended business





$$RMSE_{100\ users} = 1,48$$

Thank you for attention!

Contacts:

i.abdrashitov.ds@gmail.com

radojkovic.n@gmail.com

Github: https://github.com/iabdrashitov/Recommendation_engine_yelp