Dataset Analysis
Yelp Dataset

Team 14
Razvan Soos (40035034)
Wu Wen Tang (40028075)
Zhen Yee (40028478)

Intro

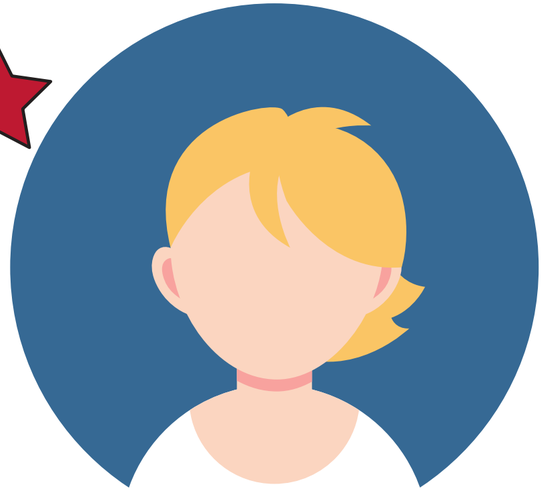
Yelp Dataset

- Company that publishes crowd-sourced reviews about businesses
- A subset of businesses, reviews and user data, written on the yelp review website

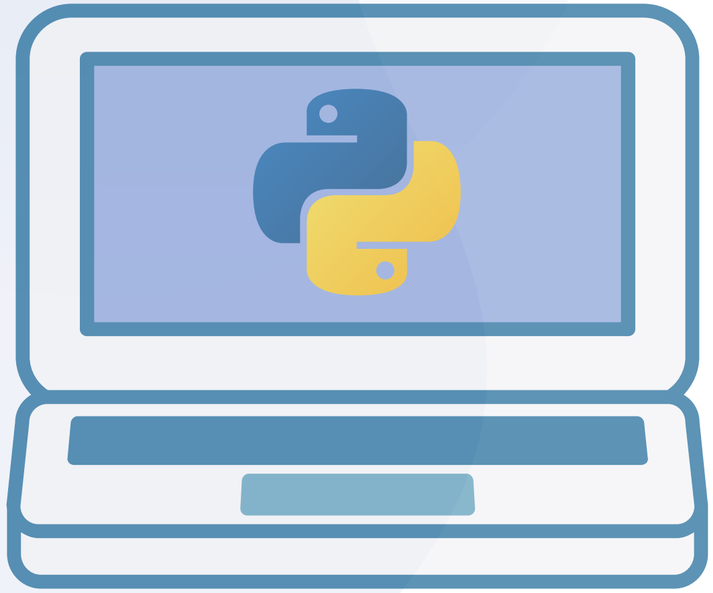
Objective

- Build a collaborative filtering recommender system based on user ratings of restaurants

Using user A's rating of restaurants, the system recommends a new restaurant to User B, predicting that it will be highly rated.



Materials & Methods



Technology

Algorithms

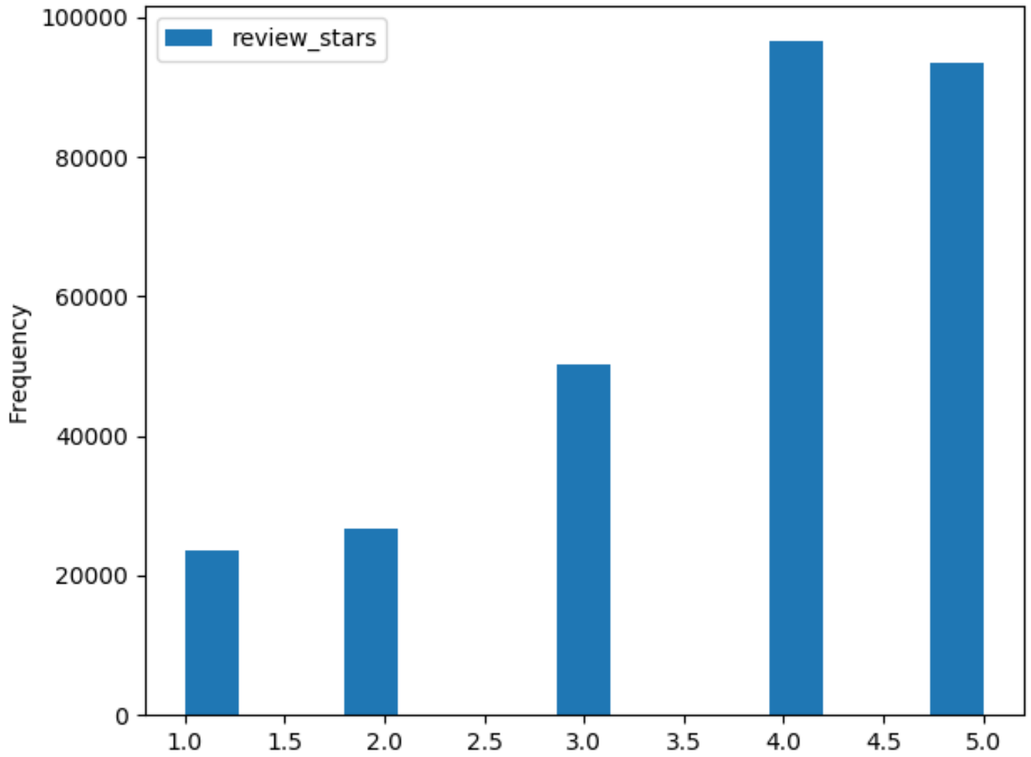
- Frequent Itemsets
- ALS

The dataset

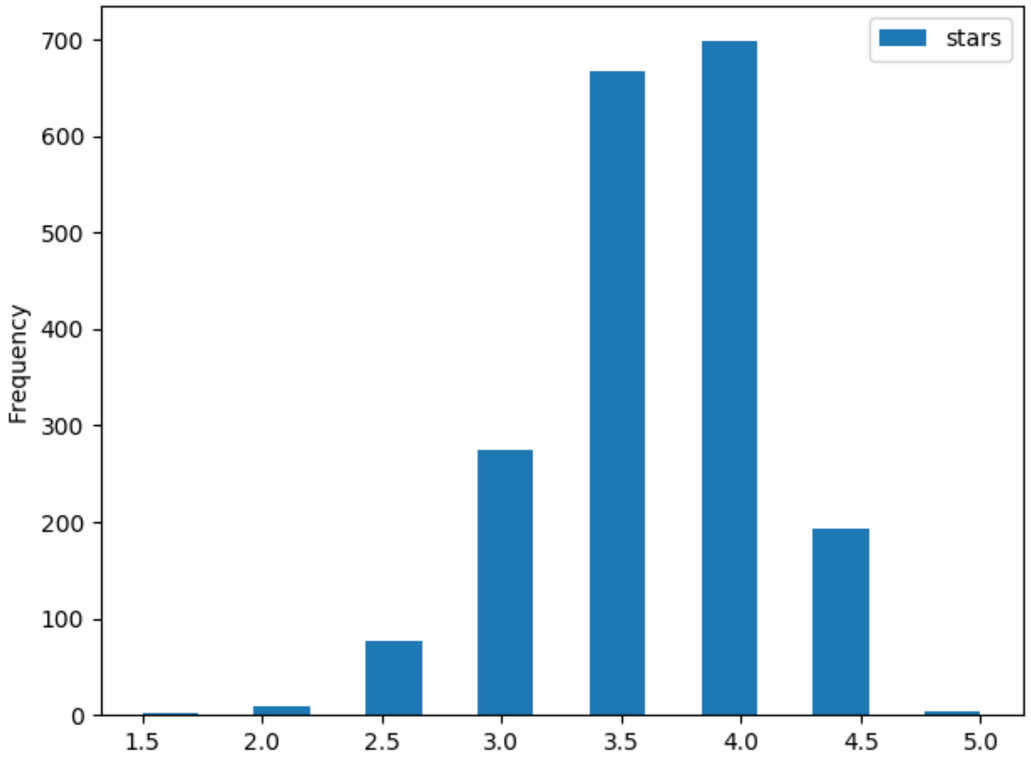
- **10 Gb**
- **5 files**
- **Only business.json and review.json were necessary**

- **Filtered to only keep open restaurants in Toronto that have over 50 reviews**
- **Millions of reviews, used chunks to pass through them and inner merge with edited businesses to only keep related reviews**
- **290900 reviews left, 1925 restaurants, 80343 users**

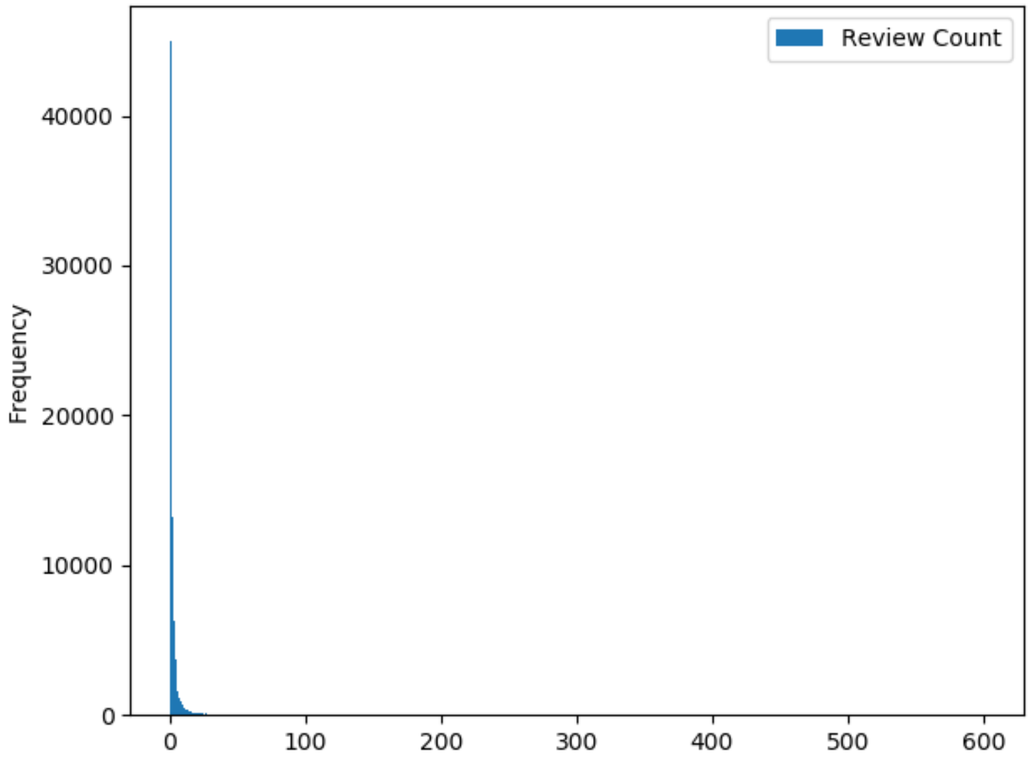
Average rating



Average stars for restaurants



Reviews per person



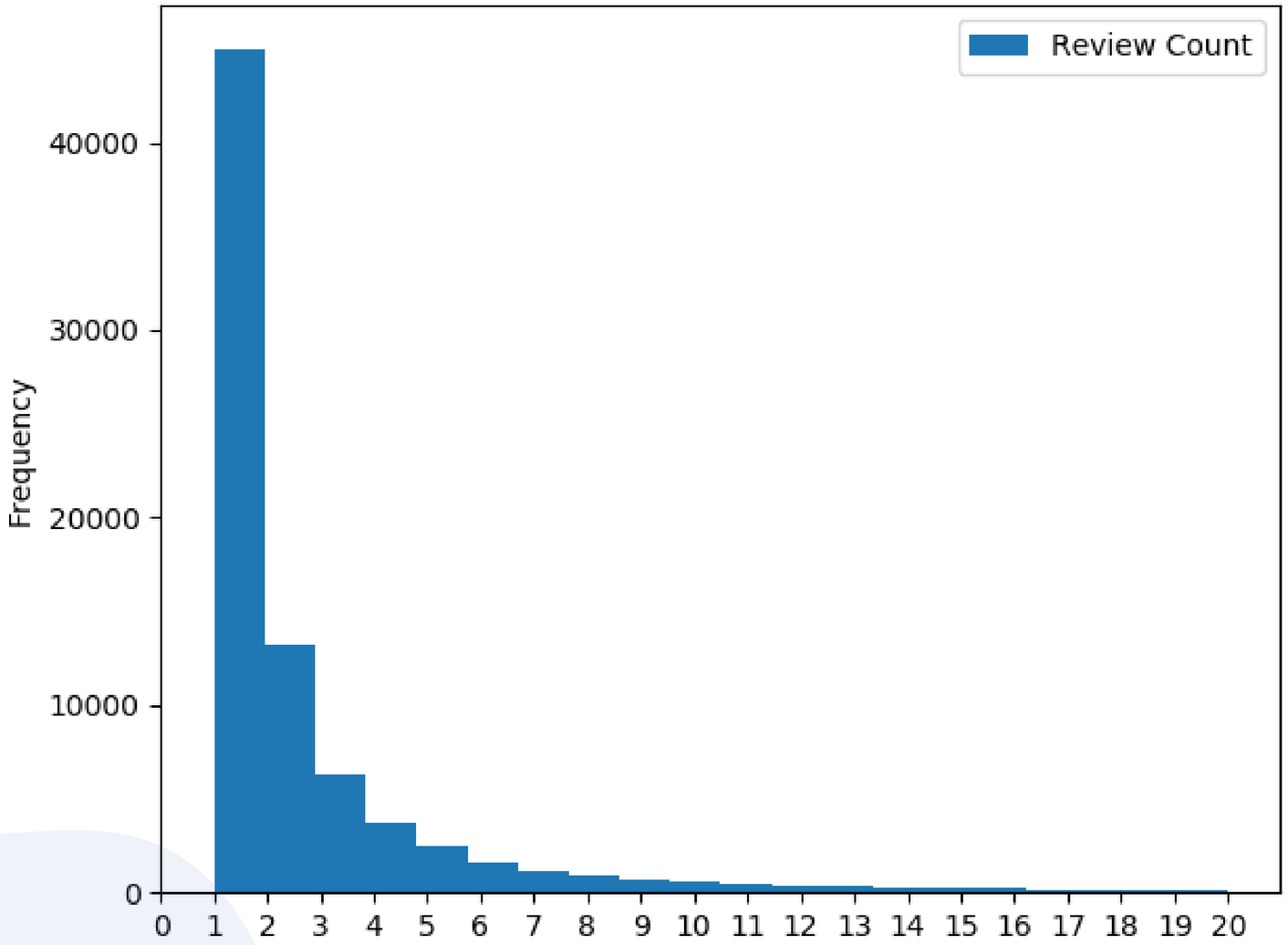
Outliers

user_id	Review Count
CxDOIDnH8gp9KXzpBHJYXw	599

Statistics

```
Review Count    80343
dtype: int64
Review Count    80199
dtype: int64
Review Count    78153
dtype: int64
```

Number of reviews for the average person



Frequent Itemsets

Association rules derived from frequent itemsets:
people who like {X} => will like {Y}

Minimum Support:

- 0.001
- Restaurant appearing 1/100 reviews

Minimum Confidence

- 0.2

Frequent itemsets

items	freq
[0a20150ytxrDjDzXNfRWkA, B70iTJjcPkuYn8ouUewWgw]	76
[N93EYZy9R0sdlEvubu94ig, RtUvSWO_UZ8V3Wpj0n077w]	74
[aLcFhMe6DDJ430zelCpd2A, N93EYZy9R0sdlEvubu94ig]	72
[8I5U8OQ06nSxX2y4PPOWzQ, 0a20150ytxrDjDzXNfRWkA]	69
[aLcFhMe6DDJ430zelCpd2A, RtUvSWO_UZ8V3Wpj0n077w]	66
[4m_hApwQ054v3ue_OxFmGw, B70iTJjcPkuYn8ouUewWgw]	66
[DE89UdHFMCN6DtYWZuer5A, RtUvSWO_UZ8V3Wpj0n077w]	63
[DE89UdHFMCN6DtYWZuer5A, N93EYZy9R0sdlEvubu94ig]	61
[N93EYZy9R0sdlEvubu94ig, B70iTJjcPkuYn8ouUewWgw]	61
[B70iTJjcPkuYn8ouUewWgw, RtUvSWO_UZ8V3Wpj0n077w]	61
[A7waf6G3cvnLfAqKeLL8DA, B70iTJjcPkuYn8ouUewWgw]	60
[8I5U8OQ06nSxX2y4PPOWzQ, B70iTJjcPkuYn8ouUewWgw]	60
[DE89UdHFMCN6DtYWZuer5A, 0a20150ytxrDjDzXNfRWkA]	59
[RwRNR4z3kY-4OsFqigY5sw, RtUvSWO_UZ8V3Wpj0n077w]	58
[0a20150ytxrDjDzXNfRWkA, RtUvSWO_UZ8V3Wpj0n077w]	58
[Yl2TN9c23ZGLUBSD9ks5Uw, B70iTJjcPkuYn8ouUewWgw]	58
[MS-hfug4QDXqb_Mws3qlzA, N93EYZy9R0sdlEvubu94ig]	57
[MS-hfug4QDXqb_Mws3qlzA, B70iTJjcPkuYn8ouUewWgw]	57
[N_2yEZ41g9zDW_gWArFiHw, B70iTJjcPkuYn8ouUewWgw]	56
[aLcFhMe6DDJ430zelCpd2A, B70iTJjcPkuYn8ouUewWgw]	55

only showing top 20 rows

Association rules

antecedent	consequent	confidence	lift
[2i0dDzxLuQQAwzxDZVR-1g]	[B70iTJjcPkuYn8ouUewWgw]	0.21120689655172414	13.153017241379311
[9ot8oInkYZTt6wkkGe__vQ]	[JMiaNitMzMbJm6Kh0RbT5A]	0.1794871794871795	18.92717505676469
[5aeOewSy4RiI8sLLWpeNGA]	[0a2O150ytxrDjDzXNfRWkA]	0.16878980891719744	11.304518011760285
[A7waf6G3cvnLfAqKeLL8DA]	[B70iTJjcPkuYn8ouUewWgw]	0.16759776536312848	10.437236347052789
[Yl2TN9c23ZGLUBSD9ks5Uw]	[B70iTJjcPkuYn8ouUewWgw]	0.16111111111111112	10.03327664399093
[HUYEadSbGSQNHXFmT2Ujjw]	[DE89UdHFMCN6DtYWZuer5A]	0.15857605177993528	11.929610403857566
[4m_hApwQ054v3ue_OxFmGw]	[B70iTJjcPkuYn8ouUewWgw]	0.15529411764705883	9.671020408163267
[C8_zdU7zGLUK3uC4e5AepQ]	[0a2O150ytxrDjDzXNfRWkA]	0.14647887323943662	9.810266813501034
[MS-hfug4QDXqb_Mws3qlzA]	[B70iTJjcPkuYn8ouUewWgw]	0.1360381861575179	8.471847450197263
[MS-hfug4QDXqb_Mws3qlzA]	[N93EYZy9R0sdlEvubu94ig]	0.1360381861575179	9.738897948613861
[HkHTdTvzbn-bmeQv_-2u0Q]	[RtUvSWO_UZ8V3Wpj0n077w]	0.12760416666666666	7.112038622526636
[N_2yEZ41g9zDW_gWArFiHw]	[B70iTJjcPkuYn8ouUewWgw]	0.12727272727272726	7.925974025974026
[8I5U8OQ06nSxX2y4PPOWzQ]	[0a2O150ytxrDjDzXNfRWkA]	0.12321428571428572	8.252145796590241
[aLcFhMe6DDJ430zelCpd2A]	[N93EYZy9R0sdlEvubu94ig]	0.11538461538461539	8.260320324836455
[RwRNR4z3kY-4OsFqigY5sw]	[RtUvSWO_UZ8V3Wpj0n077w]	0.11218568665377177	6.252687174867297
[N93EYZy9R0sdlEvubu94ig]	[RtUvSWO_UZ8V3Wpj0n077w]	0.10850439882697947	6.047510009507358
[JMiaNitMzMbJm6Kh0RbT5A]	[0a2O150ytxrDjDzXNfRWkA]	0.1079913606911447	7.232606576659053
[8I5U8OQ06nSxX2y4PPOWzQ]	[B70iTJjcPkuYn8ouUewWgw]	0.10714285714285714	6.672376093294461
[JMiaNitMzMbJm6Kh0RbT5A]	[9ot8oInkYZTt6wkkGe__vQ]	0.10583153347732181	18.92717505676469
[aLcFhMe6DDJ430zelCpd2A]	[RtUvSWO_UZ8V3Wpj0n077w]	0.10576923076923077	5.895064980681419

only showing top 20 rows

Predictions

user_id	business_id_list	prediction
wuRimFgfFMdi10VT8...	[0a2O150ytxrDjDzX...	[N93EYZy9R0sdlEvu...
Wx5xa--Qm_2vk1jYy...	[5hOhYW3728u9kT4l...	[RtUvSWO_UZ8V3Wpj...
mrru6hbeQLvagU4vk...	[ICFYlS9nwsoAyaiR...	[N93EYZy9R0sdlEvu...
wrHibYfHq_bHs4FnA...	[M_fzy4_KslKoXeyZ...	[N93EYZy9R0sdlEvu...
ZalEExPCJHmSR6Mgr...	[B70iTJjcPkuYn8ou...	[RtUvSWO_UZ8V3Wpj...
GvbQhIi9rrkyJpOrR...	[G24p1oGGfY3t-m8Z...	[RtUvSWO_UZ8V3Wpj...
cpWhL9gDc_Tkr9sbA...	[QgdO29fGB-eBNe1B...	[RtUvSWO_UZ8V3Wpj...
atO_69ftKA8MXsnr5...	[D1lAVtlav4atQTJn...	[DE89UdHFMCN6DtYW...
Kj9cFO70zZOQorN0m...	[7ZBh-3wWVQ5zkd6K...	[RtUvSWO_UZ8V3Wpj...
ERQXrV4abqYrxvsfb...	[AsPW2a72MNVuV8LT...	[RtUvSWO_UZ8V3Wpj...
SBsQvmEEYJsD6xeRz...	[ERnG-1q3igX3VSgm...	[RtUvSWO_UZ8V3Wpj...
XTGGEzW5V-vQWutcn...	[B70iTJjcPkuYn8ou...	[RtUvSWO_UZ8V3Wpj...
EK8TCpE_e0wOhOtui...	[8cr7Kdx1bT51CnKs...	[RtUvSWO_UZ8V3Wpj...
Ko4y7jIissMR8ZGQl...	[1K4qrnfyzKzGgJPB...	[RtUvSWO_UZ8V3Wpj...
Grn2WeXGx0jqxoG6K...	[8J0NuWmoFfSGe5Lu...	[RtUvSWO_UZ8V3Wpj...
VqO3p3fs8_G2VIgkC...	[Mksyt1TZQ-7vzCZj...	[N93EYZy9R0sdlEvu...
sX8HqvlKZJBD2WRHN...	[0a2O150ytxrDjDzX...	[RtUvSWO_UZ8V3Wpj...
S_wlsi3NYVYK-SGk-...	[AnemBUVAb9NRasel...	[RtUvSWO_UZ8V3Wpj...
SHipPHbt7gI9hAtI_...	[B70iTJjcPkuYn8ou...	[RtUvSWO_UZ8V3Wpj...
EgotOxec2MuH4xvIv...	[2Y-28XrkMeeA4FFv...	[RtUvSWO_UZ8V3Wpj...

only showing top 20 rows

prediction
[N93EYZy9R0sdlEvubu94ig, aLcFhMe6DDJ430zelCpd2A, 8I5U8OQ06nSxX2y4PPOWzQ,
[RtUvSWO_UZ8V3Wpj0n077w, N93EYZy9R0sdlEvubu94ig, aLcFhMe6DDJ430zelCpd2A,
[DE89UdHFMCN6DtYWZuer5A, 0a2O150ytxrDjDzXNfRWkA, N93EYZy9R0sdlEvubu94ig,
[N93EYZy9R0sdlEvubu94ig, 8I5U8OQ06nSxX2y4PPOWzQ, N_2yEZ41g9zDW_gWArFiHw,
[RtUvSWO_UZ8V3Wpj0n077w, aLcFhMe6DDJ430zelCpd2A, 8I5U8OQ06nSxX2y4PPOWzQ,
[RtUvSWO_UZ8V3Wpj0n077w, N93EYZy9R0sdlEvubu94ig, aLcFhMe6DDJ430zelCpd2A,
[N93EYZy9R0sdlEvubu94ig, aLcFhMe6DDJ430zelCpd2A, 8I5U8OQ06nSxX2y4PPOWzQ,
[RtUvSWO_UZ8V3Wpj0n077w, N93EYZy9R0sdlEvubu94ig, aLcFhMe6DDJ430zelCpd2A,
[RtUvSWO_UZ8V3Wpj0n077w, N93EYZy9R0sdlEvubu94ig, aLcFhMe6DDJ430zelCpd2A,
[RtUvSWO_UZ8V3Wpj0n077w, N93EYZy9R0sdlEvubu94ig, aLcFhMe6DDJ430zelCpd2A,
[RtUvSWO_UZ8V3Wpj0n077w, N93EYZy9R0sdlEvubu94ig, aLcFhMe6DDJ430zelCpd2A,
[RtUvSWO_UZ8V3Wpj0n077w, N93EYZy9R0sdlEvubu94ig, aLcFhMe6DDJ430zelCpd2A,
[RtUvSWO_UZ8V3Wpj0n077w, N93EYZy9R0sdlEvubu94ig, aLcFhMe6DDJ430zelCpd2A,
[N93EYZy9R0sdlEvubu94ig, aLcFhMe6DDJ430zelCpd2A, 8I5U8OQ06nSxX2y4PPOWzQ,
[RtUvSWO_UZ8V3Wpj0n077w, N93EYZy9R0sdlEvubu94ig, aLcFhMe6DDJ430zelCpd2A,
[RtUvSWO_UZ8V3Wpj0n077w, N93EYZy9R0sdlEvubu94ig, aLcFhMe6DDJ430zelCpd2A,
[RtUvSWO_UZ8V3Wpj0n077w, N93EYZy9R0sdlEvubu94ig, aLcFhMe6DDJ430zelCpd2A,
[RtUvSWO_UZ8V3Wpj0n077w, N93EYZy9R0sdlEvubu94ig, aLcFhMe6DDJ430zelCpd2A,
[RtUvSWO_UZ8V3Wpj0n077w, N93EYZy9R0sdlEvubu94ig, aLcFhMe6DDJ430zelCpd2A,
[RtUvSWO_UZ8V3Wpj0n077w, N93EYZy9R0sdlEvubu94ig, aLcFhMe6DDJ430zelCpd2A,

ALS Algorithm

rmse: 1.264359277139277

	user_id	recommendations
0	--7gjElmOrthETJ8XqzMBw	[(2X5I7wY_bqYWOub-q2Ba3Q, 8.786605834960938), ...
1	--C93xIlmjtgQfSOIpcQSA	[(MOB5oCEKCw3S76SsiuQjqA, 3.898888111114502), ...
2	--cd_gA-9Q8gM9P2cTxEsQ	[(TBzgzTFSa7pJXiLD7emYaQ, 11.425394058227539),...
3	-00AazqEBd6ZK1lgsbatyA	[(qQsrcouREdFUk4adim1uEA, 9.2857027053833), (T...
4	-018WmPPk8qlp3TEiqqMVw	[(Ud5FqhVC8AdZg9RLzNW_Wg, 1.4020780324935913),...
5	-04VHV-dxkadvMWL81XDxw	[(iPuY6dR5w5X_gOM_2mx-hA, 4.99386739730835), (...
6	-0AyZxS5C--WySnbW_Q8yQ	[(IvSzF1r0hhTwGS1LGecGmA, 6.807080268859863), ...
7	-0Mt0sE9r_3AckmHX4Klrw	[(ptvN80KgWDINzWu80mHK5Q, 7.185107707977295), ...
8	-0OE9Pn8vSK-WjJeRtHDtw	[(PyjnsWzNUTo1n2ohq0QznA, 2.539547920227051), ...
9	-0SKtG3AAAIs97gj4H-83g	[(j7Rgozl24WrmDeqFCCKwQg, 3.655378580093384), ...
10	-0ZildkQkxhgqGWpkzv9Jg	[(HD10w7sMM9HkF4pM8BlJZQ, 7.071044921875), (iP...

Conclusion

- Many users only have one restaurant rated
- ALS isn't precise in these situations since it's hard to recommend a restaurant to someone who only rated one restaurant
- Frequent item sets is a bit better for this specific situation but still not great