

# My Yelper

## A Hybrid Recommendation Engine

Sami Mustafa

«Lecture 06»

October 2017

# Key Types of Recommendation Systems

## **Content-based Filtering**


Produces recommendations based on the similarity between items attribute and user's profile

## **Collaborative Filtering**

Produces recommendations based on the knowledge of users' attitude to items, that is it uses the "wisdom of the crowd" to recommend items

## **Hybrid Recommender**

Overcomes disadvantages of a single recommender type by merging more than one recommender type together

A decorative graphic at the bottom of the slide consisting of several overlapping, wavy lines in shades of yellow, orange, and green, creating a sense of motion and flow.

# Yelp Dataset

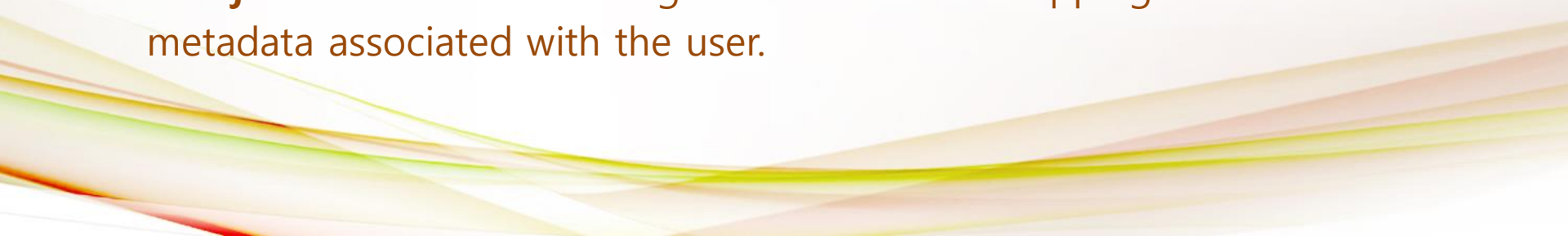
- 4,700,000 reviews
- 156,000 businesses
- 1,100,000 users
- 12 metropolitan areas

available at <https://www.yelp.com/dataset>

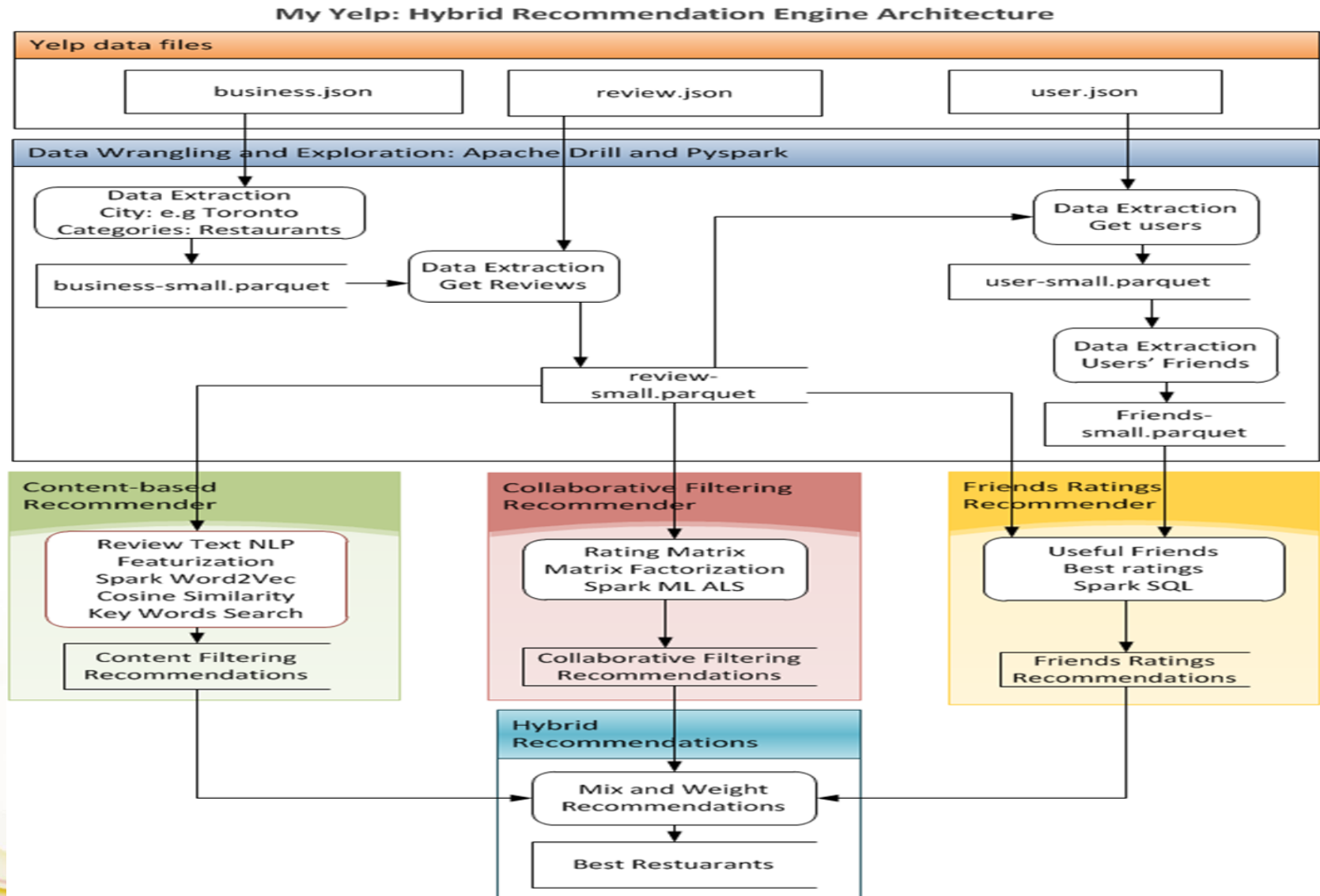
**business.json:** Contains business data including location data, attributes, and categories.

**review.json:** Contains full review text data including the user\_id that wrote the review and the business\_id the review is written for.

**user.json:** User data including the user's friend mapping and all the metadata associated with the user.



# Approach / Architecture

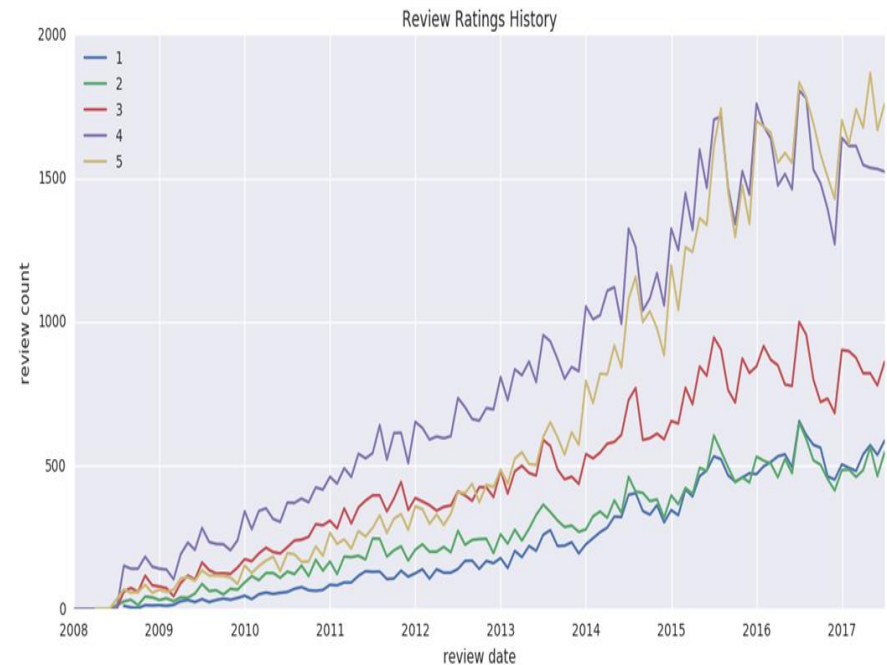
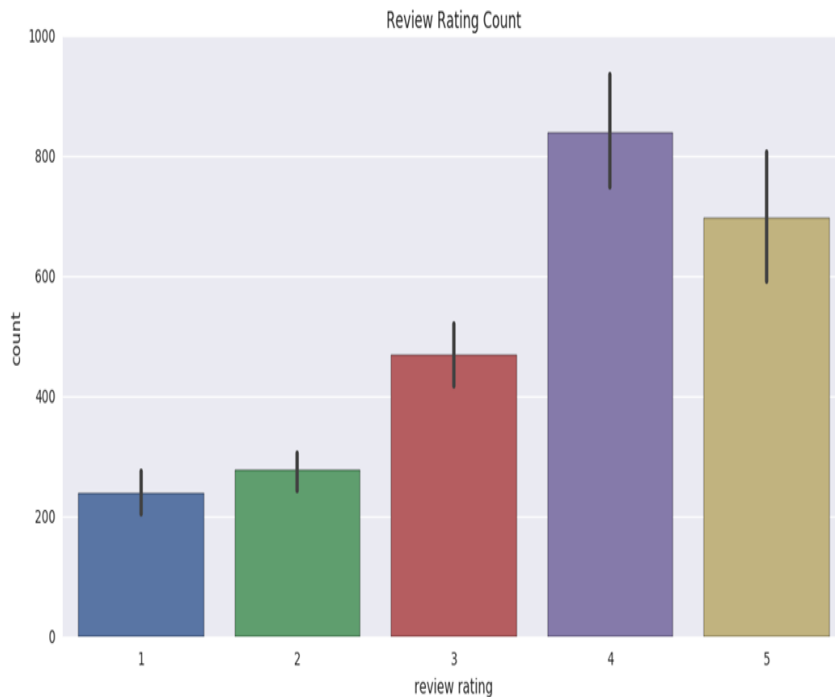


# Data Loading and Exploration

business_id	business_name	address	city	state	postal_code	latitude	longitude	stars	review_count	categories
g6AFW-zY0wDvBI9U82g4zg	Baretto Caffe	1262 Don Mills Road	Toronto	ON	M3B 2W7	43.744768	-79.346527	5.0	146	[Italian, Restaurants, Cafes]
J9vAdD2dCpFuGsxPIn184w	New Orleans Seafood & Steakhouse	267 Scarlett Road	Toronto	ON	M6N 4L1	43.677592	-79.506363	5.0	62	[Cajun/Creole, Seafood, Steakhouses, Restaurants]

user_id	user_name	review_count	yelping_since	useful	funny	cool	fans	average_stars
om5ZiponkpRqUNa3pVPiRg	Andrea	2559	2006-01-18	83681	10882	40110	835	3.94
Wc5L6iuvSNF5WGB1qIO8nw	Risa	1122	2011-07-30	26395	4880	19108	435	4.10
uxKSnOV AoEj4I6X9YhLB1g	Vivian	73	2013-03-02	34	5	2	8	3.54
s8bVHRqx6cl8F8HGf3A_og	Colleen	32	2014-12-18	19	3	7	2	4.15

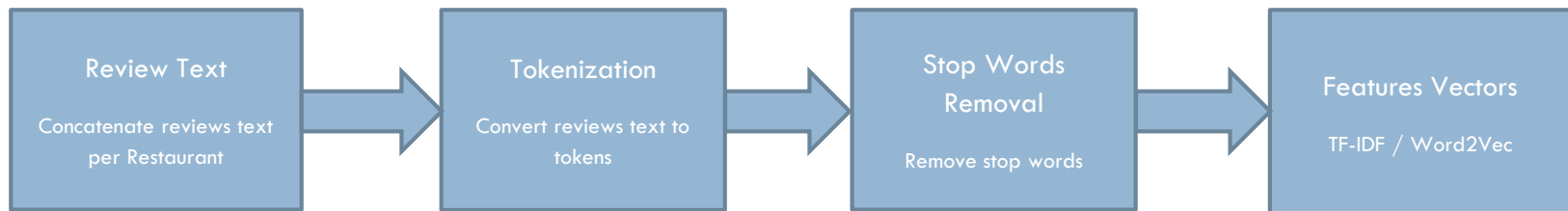
review_id	user_id	business_id	stars	review_date	review_text	useful	funny	cool
Z5I99h18E3_g1GLDSsWqA	djpMXOA1ic5wv3FPtubHNw	mr4FiPaXTWlJ3qGzp4-7Yg	3	2009-07-21	I left Table 17 feeling very ambivalent. Meh a...	3	0	0
Z3Fw292i0Eg8liWODT1jsw	-pXs08gJq9ExIk275YLvPg	mr4FiPaXTWlJ3qGzp4-7Yg	3	2008-12-13	for the time being, for all its worth, i am go...	1	0	0
hsKINx1dIKeFTDe-ZiCvgA	PTj29rhujYETuFlAZaDi3w	mr4FiPaXTWlJ3qGzp4-7Yg	5	2013-10-12	Love this place. I went there with me boyfrien...	1	0	1
oviMS8F4ACfGysxsXKmw	3hLMY2dBEP1kYbd_ywTsCQ	mr4FiPaXTWlJ3qGzp4-7Yg	5	2013-02-17	Had a lovely evening last night at Table 17. ...	0	0	0



# Content-based Filtering

1. Create items attributes from review text (Feature Extraction)
2. Create user profile from user's rating history
3. Compute cosine similarity between user's profile and items ' features
4. Return top-N restaurants similar to the user's profile, not seen by the user before

## Feature Extraction / Text Vectorization



business_id	text	token	nostopwrđ	idf_vec	word_vec
bfr-vJvrjdOJaWsXGJgzPA	Attention allergy sufferer s: Claims to have a ...	[attention, allergy, sufferers, claims, to, ha...	[attention, allergy, sufferers, claims, nut, f...	(2.32949090224, 0.946347870074, 0.824754840176...	[-0.0949216104917, -0.0313844011, 0.0271677521...
DI2vgi5W_nbe-A97D0zgfA	I don't understand previous review. I went the...	[i, don, t, understand, previous, review, i, w...	[understand, previous, review, went, three, ti...	(0.325045242174, 0.141952180511, 0.11782212002...	[-0.0657136221337, 0.0038411388924, 0.02664112...
65ZGMedBm7TBpWv6fzH2_Q	Food here is always fresh and healthy definite...	[food, here, is, always, fresh, and, healthy, ...	[food, always, fresh, healthy, definitely, gre...	(0.0541742070289, 0.0, 0.117822120025, 0.44249..	[-0.00367327128618, -0.0228931885972, -0.00643...

# TF-IDF

is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. The TF-IDF value increases proportionally to the number of times a word appears in the document, but is often offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general

# Word2Vec

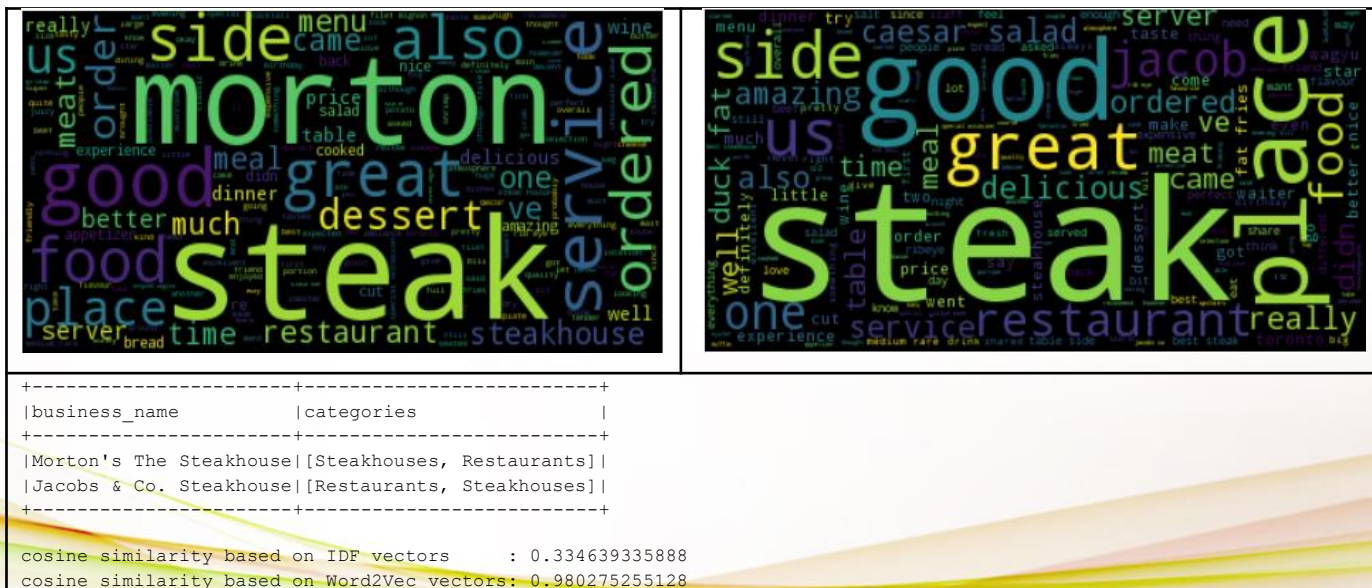
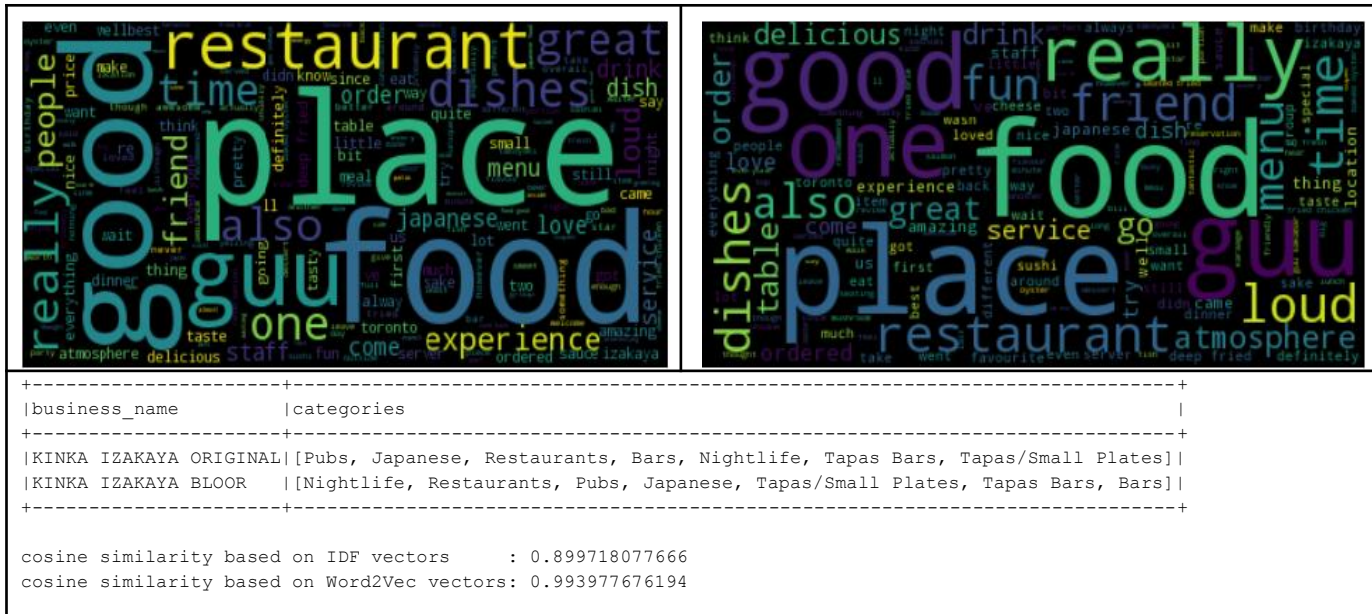
is a shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words. Word2vec takes as its input a large corpus of text and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space.

# Cosine Similarity

is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. two vectors with the same orientation have a cosine similarity of 1, two vectors at  $90^\circ$  have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude



# Cosine Similarity Results: Word2Vec Vs. TF-IDF





# Content-based Filtering Results

## User Profile:

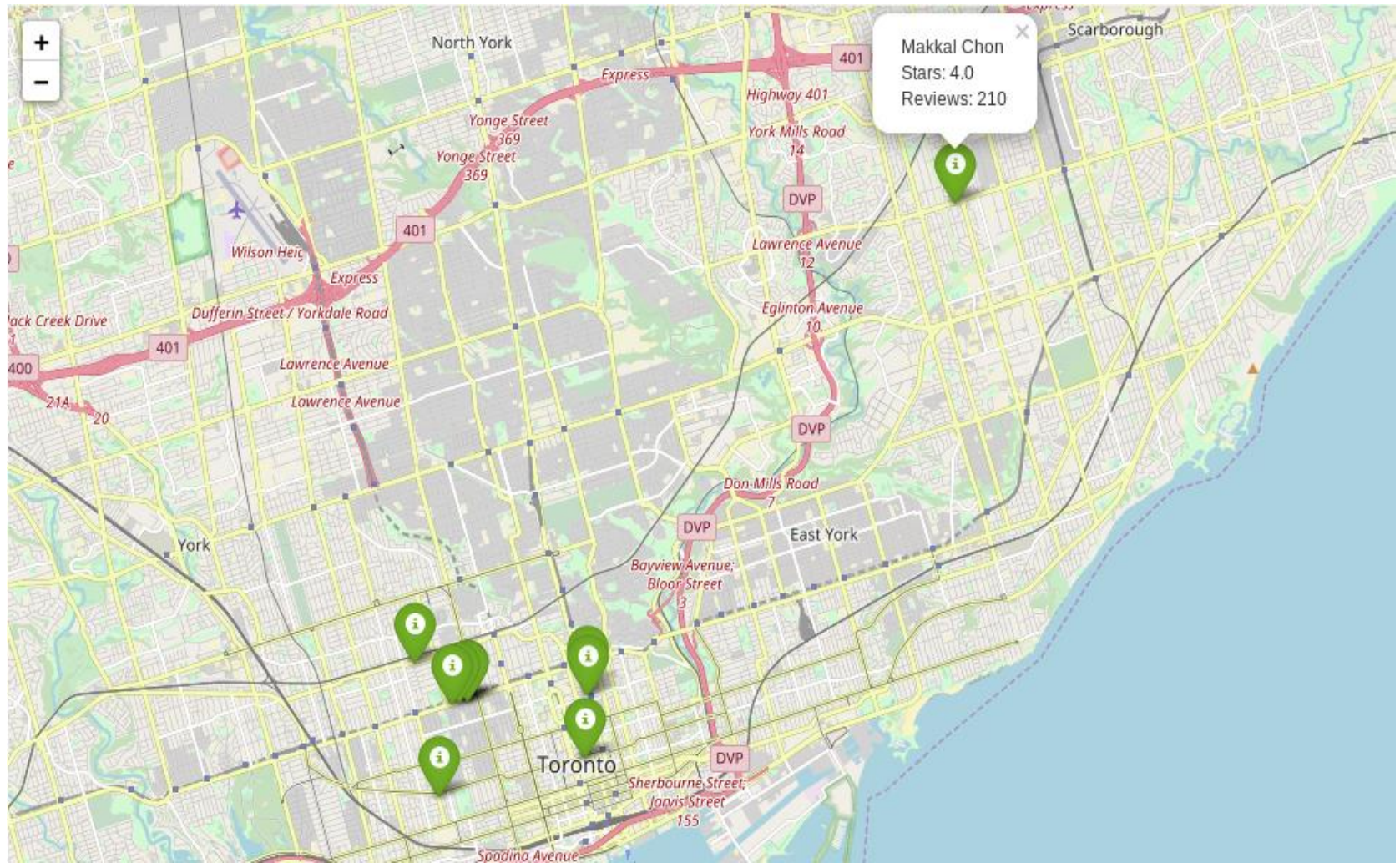
Businesses previously reviewed by the user: 'Wc5L6iuvSNF5WGBLqIO8nw'

business_id	business_name	categories
F_oPMHJrH42R67xp5eKtQA	Yummy Korean Food Restaurant	[Korean, Restaurants]
_HqZL3gK98-Q4ObAoyMlaw	Rose and Sons Swan	[Breakfast & Brunch, American (Traditional), Restaurants]
aAYl1-tsvP9Kzs7YESi8Rg	Electric Mud BBQ	[Restaurants, Sandwiches, Food, Barbeque, Desserts]
acvmtF41s5Qa1ZNadUV0Dw	Average Joe's Café	[Breakfast & Brunch, Restaurants, Food, Bagels]
M8S7poDhCIzqQx1--GB-ww	La Cubana	[Cuban, Restaurants, Mexican]

## Recommendations:

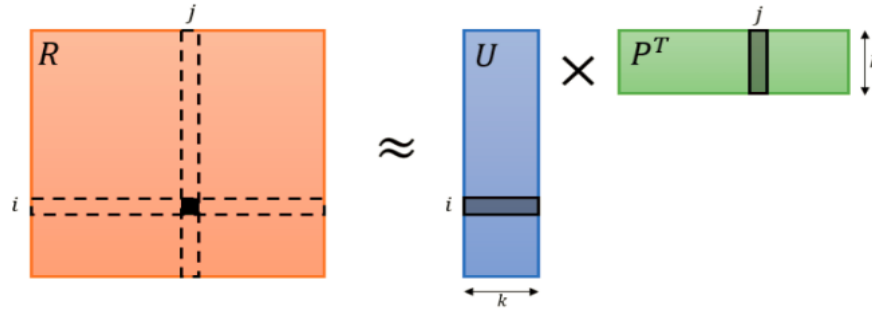
	business_id	score	business_name	categories	stars	review_count	latitude	longitude
0	rO3lZpVSoRMhhd0AEJBjlg	0.986259	Sunrise House	[Restaurants, Korean]	4.0	135	43.664068	-79.415668
1	rhyjGfqYICJoi8Zeulg6QA	0.983902	Kimchi Korea House	[Korean, Restaurants]	3.5	155	43.655256	-79.385475
2	j-Z_HAev26ZfdErMhIBuA	0.980066	Thumbs Up Korean Restaurant	[Restaurants, Korean]	4.0	56	43.664451	-79.413786
3	_MA98TVmvVIy-Xdl0poc7w	0.979998	Mom's Korean Food	[Korean, Restaurants]	3.5	62	43.664686	-79.413785
4	SNkkuchbVtUzCwyENcai_g	0.979673	Danji	[Restaurants, Chinese, Japanese, Korean]	3.5	57	43.665300	-79.384899
5	X6ZZksefmR_piQj2Gbnduw	0.975525	Paldo Gangsan	[Restaurants, Korean]	4.0	47	43.663799	-79.417393
6	oQylTvXwGikKFdcjmafKVg	0.973422	Fire on the East Side	[Southern, Restaurants, Breakfast & Brunch, Am...]	3.5	119	43.666765	-79.384836
7	WnUttoJffplgWaQGR2J2Xw	0.973141	The Saint Tavern	[Restaurants, Bars, Nightlife, Gastropubs]	3.5	121	43.649062	-79.420478
8	ShUh_MMkaVp_KXCtNjPvXA	0.973117	Universal Grill	[American (Traditional), Canadian (New), Break...]	3.5	45	43.670521	-79.426440
9	uChTCA6MsLAcIDRklpO-Fw	0.972612	Makkal Chon	[Greek, Restaurants, Korean]	4.0	210	43.744944	-79.296636

# Content-based Filtering Results - Map



# Collaborative Filtering

Rating Matrix decomposition using  
Matrix Factorization (MF)



Matrix Factorization to minimize the  
cost function:

$$J = ||R - U \times P^T||_2 + \lambda (||U||_2 + ||P||_2)$$

Alternating Least Squares (ALS) does just that. It is a two-step iterative optimization process. In every iteration it first fixes  $P$  and solves for  $U$ , and following that it fixes  $U$  and solves for  $P$

# Collaborative Filtering Results

Rating matrix no. of rows: 276,887

userId	businessId	rating
23561	872	5.0
6268	872	4.0
8646	872	4.0
531	4253	3.0
2217	4253	5.0

Collaborative Filtering Recommendations for the user: 'ZWD8UH1T7QXQr0Eq-mcWYg'

	business_id	rating	business_name	categories	stars	review_count	latitude	longitude
0	LcIgUIWaJJwtOfPoPWcmBg	4.570364	Soupe Shoppe	[Restaurants, Street Vendors, Food, Soup, Food...]	5.0	4	43.651425	-79.404123
1	mpDcuUs6dB5uBsYVKDWCNQ	4.527902	Druxy's Famous Deli	[Restaurants, Sandwiches, Delis, Breakfast & B...]	4.0	4	43.648235	-79.379525
2	1VAsBosvx02jpvUxiKvmg	4.490123	The Dumpling Shop	[Restaurants, Specialty Food, Chinese, Dim Sum...]	4.5	11	43.767971	-79.401363
3	9GLN1xfck07CKfNfejKCwg	4.438345	T-Sushi	[Food, Restaurants, Sushi Bars, Food Delivery ...]	5.0	13	43.644745	-79.390892
4	vAz5pelrjwkpMDo_OHCDag	4.414823	Kuya Willie's Kainan	[Breakfast & Brunch, Filipino, Restaurants]	3.5	3	43.759288	-79.310866
5	y9yeMK6N0UINVEC I3ljz3Q	4.401293	Hot Dog Stand	[Hot Dogs, Restaurants]	4.0	3	43.681236	-79.377222
6	XXa5R1IJSvNrbo8InhNliQ	4.399106	Toronto Star Food Buidling	[Food, Fast Food, Restaurants]	4.5	3	43.632265	-79.420313
7	LijIU7K-0SPXPtYFQixamQ	4.392767	Magic Oven	[Food Stands, Sandwiches, Restaurants, Indian]	5.0	3	43.652294	-79.405521
8	fxRcHzovnRyWh_WMdQoNOQ	4.377005	Taj Restaurant	[Restaurants, Russian, Mediterranean]	5.0	4	43.696764	-79.446227
9	2H5EaBEreDzzP7sPmD_oDQ	4.362335	Vila Verde	[Restaurants, Event Planning & Services, Portu...]	4.0	4	43.651243	-79.410631

# Friends Recommendations

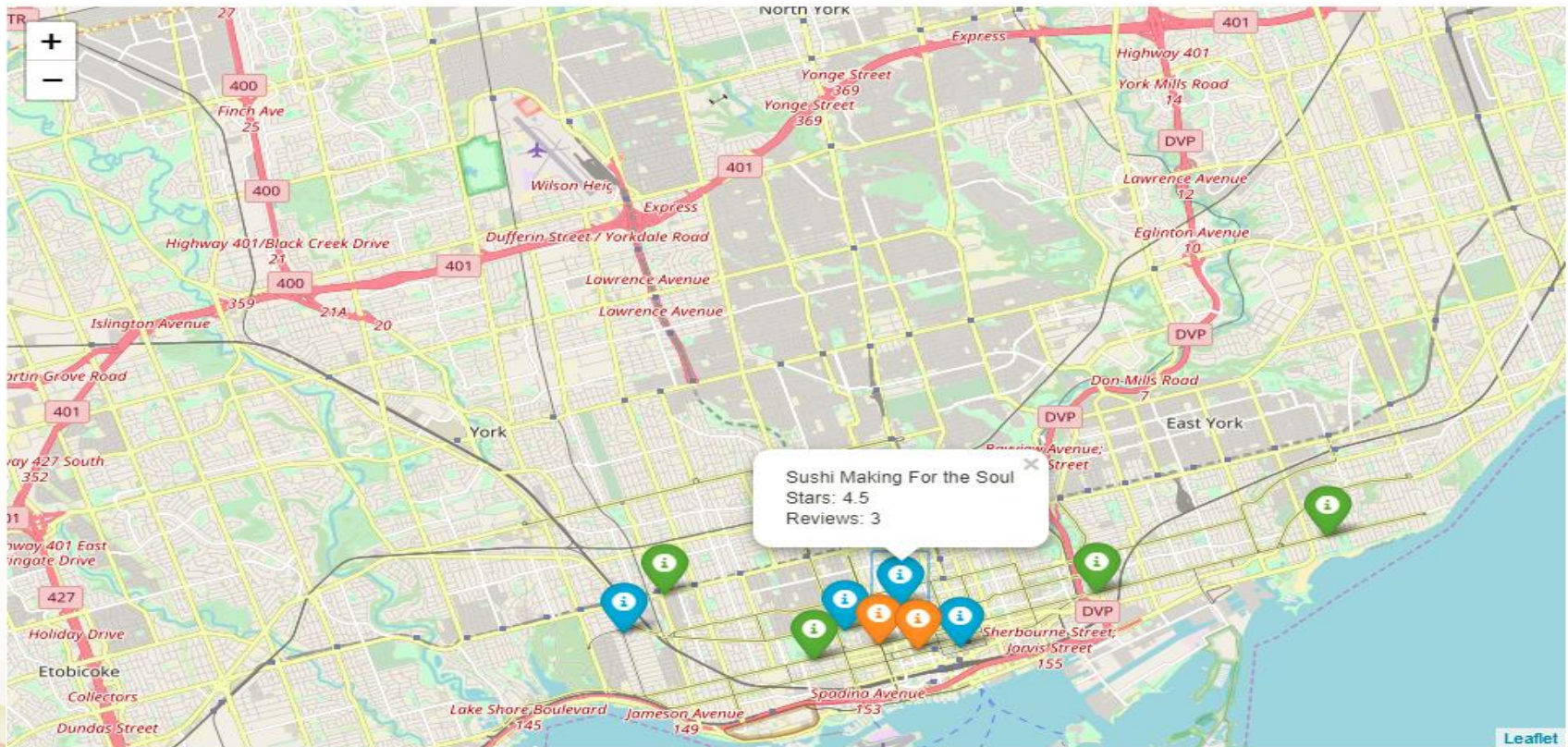
- Assumption: user trusts the agreement of his friends
- Recommended restaurants being top ranked by the user's friends

id	business_id	4_5_stars_count	business_name	categories	stars	review_count	latitude	longitude
0	SGP1jf6k7spXkgwBhiUVw	5	Kekou Gelato House	[Food, Restaurants, Ice Cream & Frozen Yogurt,...	4.5	332	43.655983	-79.392686
1	kOFDVcnj-8fd3dolpCQ06A	5	Mildred's Temple Kitchen	[Comfort Food, Event Planning & Services, Vegetarian...	4.0	472	43.639911	-79.420424
2	0a2O150ytxrDjDzXNfRWkA	4	Miku Toronto	[Sushi Bars, Restaurants, Seafood, Japanese]	4.0	384	43.641235	-79.377370
3	G6EkDTXZ6zMUovg7JTG4YQ	3	Vietnam Noodle Star	[Restaurants, Vietnamese, Noodles]	3.5	148	43.804603	-79.287842
4	RwRNR4z3kY-4OsFqigY5sw	3	Uncle Tetsu's Japanese Cheesecake	[Desserts, Japanese, Restaurants, Bakeries, Food]	3.5	806	43.655969	-79.384013
5	Yv4P4qUwd7F-qQ4Y4eDIJQ	3	Han Ba Tang	[Nightlife, Pubs, Lounges, Korean, Asian Fusion...	3.5	213	43.762928	-79.411511
6	dTuT_G3Zp79RZmnF3oxfiA	3	The Bier Markt	[Belgian, Nightlife, Bars, Gastropubs, Canadiana...]	3.0	197	43.647095	-79.373915
7	MhiBpIBNTCAm1Xd3WzRzjQ	3	Messini Authentic Gyros	[Mediterranean, Sandwiches, Greek, Restaurants...]	3.5	372	43.677691	-79.350536
8	9_CGhHMz8698M9-Pkvf0CQ	2	Little Coxwell Vietnamese & Thai Cuisine	[Vietnamese, Thai, Restaurants]	4.0	109	43.696175	-79.329092
9	ofw8aDSEg1HoQdmCgvLtaQ	2	The Pie Commission	[Canadian (New), Fast Food, Food, Do-It-Yourself...]	4.5	183	43.623881	-79.512074




# Hybrid Recommendation Engine

- Gather Content-based, Collaborative, and Friends recommendations
- Return mix of recommendation based on ratio
- Has extra methods for data and ML models loading, ML models training and saving, data transformation etc.





# Next Steps

- ✓ Enhance user profile creation for content-based filtering using weighted average of items most liked by the user
  - ✓ Perform ALS model hyper-parameters tuning on a high end hardware to enhance the RMSE value and hence model performance
  - ✓ Introduce evaluation methods and metrics to assess the performance of various recommender types and different hybrid strategies
- 
- The bottom of the slide features a decorative graphic consisting of several overlapping, wavy lines in shades of yellow, orange, and light green, creating a sense of motion and flow.