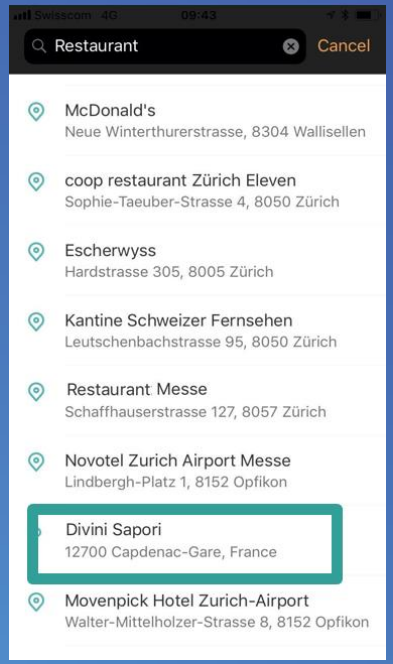
Collaborative Filtering based POI recommendations and local optimization in query context space

Soojung Hong
Lead Research Engineer @ Platform Research

March 29, 2019

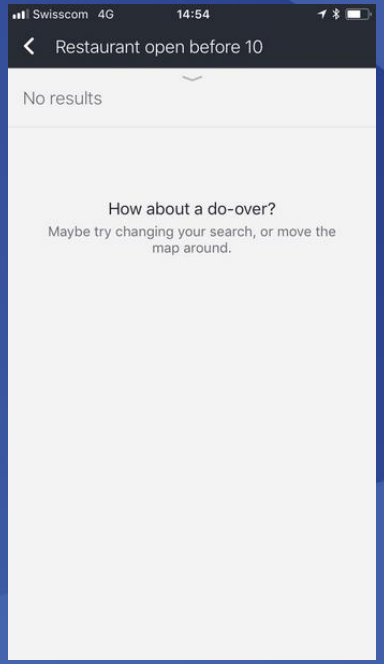
NO answers in recommendation systems

- But obvious bad recommendations



Query location : Zurich, Switzerland

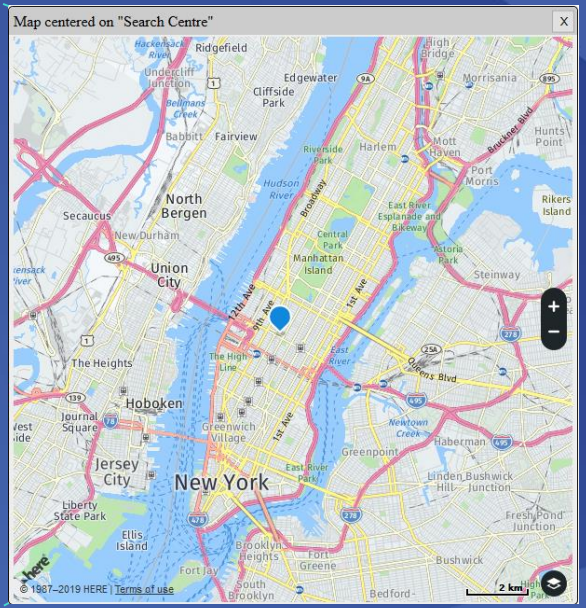
Query : ‘restaurant’



Compound query

Does this museum recommendation make sense for you?

result_index	name_chosen
1	New Museum
2	Merchant's House Museum
3	International Center of Photography Museu
4	Center for Jewish History
5	Afa Gallery
6	Theodore Roosevelt Birthplace
7	Rubin Museum of Art
8	Deitch Projects
9	Ukrainian Museum
10	Museum of Illusions
11	Artists Space
12	Margo Feiden Galleries
13	KGB Spy Museum
14	Fusion Arts Museum
15	American Artist Professional League
16	Fitzgerald Fine Arts
17	Fortnight Institute
18	Museum of Beautiful People
19	TMJ Arts Collective
20	Eva Presenhuber
21	Judd Foundation
22	Leslie+Lohman Prince Street Project
23	Museum of Modern Art
24	Metropolitan Museum of Art, New York
25	Solomon R Guggenheim Museum
26	Whitney Museum of American Art
27	Frick Collection
28	Brooklyn Museum of Art
29	Morgan Library and Museum
30	Neue Galerie New York
31	New York Hall Of Science
32	Children's Museum Of The Arts
33	New York Transit Museum
34	Studio Museum in Harlem
35	Museum of the City of New York
36	Hayden Planetarium
37	Liberty Science Center
38	Discovery Times Square
39	Jewish Museum
40	American Folk Art Museum

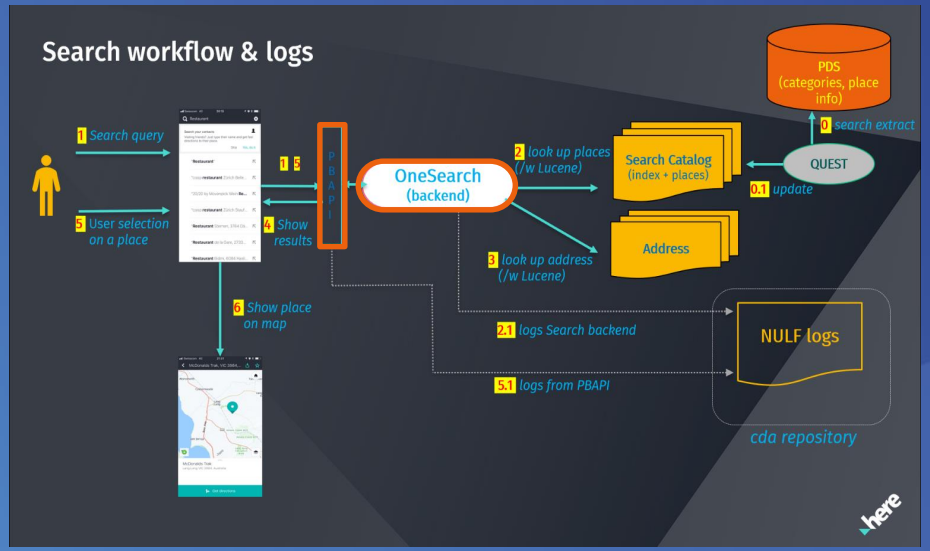


Recommended POIs on a query “museum” in New York City

Contents

- Collaborative Filtering based POI recommendations
  - Alternating Least Square method (Matrix Factorization)
  - Cross-validation, Hyper-parameter tuning, Feature engineering

- Local optimization in query context space



Three common approaches in recommendation systems

- Content based recommendation (CBR)
- Collaborative Filtering (CF)
- Hybrid model which use both CBR & CF

Contents based recommendation

- Focus on products (in our case POIs) rather than users
- Represent products (POIs) as feature vectors
- Feature vector is a set of attribute values

Place = [geo-location, place name, category, city name, language]

$X^{(i)}$ = [(40.73073, -73.996), "Metropolitan Museum of Art", museum, New York, EN]

- User feedbacks (e.g. rate) $Y^{(i)}$ constitute training data

$\widehat{y^{(i)}} = \theta x^{(i)} = \sum_{j=1}^{d} \theta_j \, x_j{}^{(i)}$ (d is total number of features for place)

To estimate parameter θ, minimize the cost function $J(\theta)$ using stochastic gradient descent

$$J(\theta) = \sum_{i \in D} ((\widehat{y^{(i)}} - \theta x^{(i)})^2/2) + \lambda \|\theta\|^2$$

Disadvantage of Contents based recommendation

- To make regression method more successful in prediction
  ➔ need rich feature vector
  ➔ high dimensional feature vector for place

- To avoid overfitting
  ➔ we must have more rating on places during training

➔ Not many ratings collected by HERE map

Collaborative Filtering

- Using user data rather than product (POI) data

- Nearest Neighbor Prediction
  - ➔ Find other users who did similar selection on certain products
  - ➔ Limitation : Assumption that at least there are some users who already have chosen the products (places) and provided rate

Similarity between user *a* and *b*

$$\mathrm{sim}(a,b) = \mathrm{corr}(a,b) = \frac{\sum_{j \in CR(a,b)} (Y_{aj} - \bar{Y}_a)(Y_{bj} - \bar{Y}_b)}{\sqrt{\sum_{j \in CR(a,b)} (Y_{aj} - \bar{Y}_a)^2} \sqrt{\sum_{j \in CR(a,b)} (Y_{bj} - \bar{Y}_b)^2}}$$

Prediction of rate on product *i* by user *a*

$$\hat{Y}_{ai} = \bar{Y}_a + \frac{\sum_{b \in KNN(a,i)} \mathrm{sim}(a,b)(Y_{bi} - \bar{Y}_b)}{\sum_{b \in KNN(a,i)} |\mathrm{sim}(a,b)|}$$

Cartoon from The New Yorker

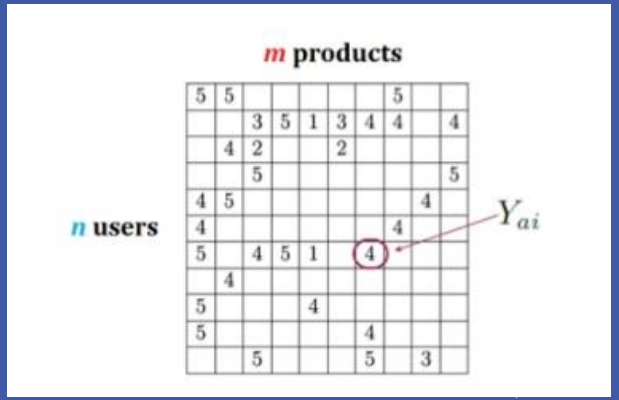
Matrix Factorization

Handles **whole users and products** rather than each individual
Solve the problem more mathematical manner using linear regression

Ultimately, find matrix X which has minimum

$$cost\ function \sum_{ai \in D} \frac{\left(y^{(i)} - Y_{ai}\right)^2}{2} + \frac{\lambda}{2} \sum_{ai} Y_{ai}^2$$

Y_{ai} : rating from observed data
$y^{(i)}$: predicted rate value

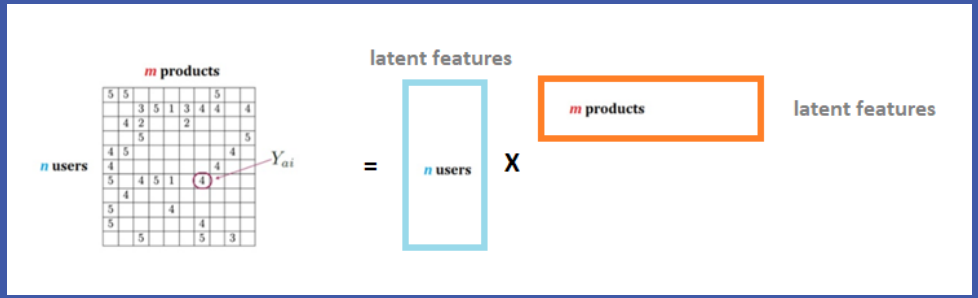


Matrix Factorization = Alternating Least Square method

Data : Matrix X is very sparse (matrix sparsity > 50%)
Goal : Approximate the solution matrix X as $\hat{X}$
ALS method :

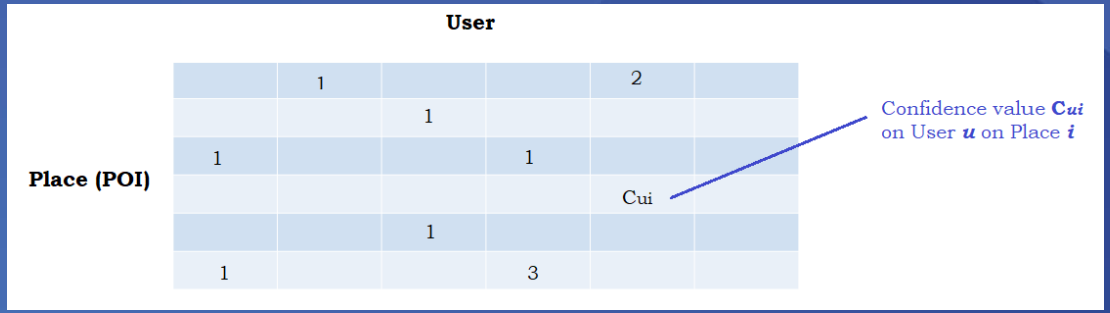
1. Factorize as a product of two smaller matrices $X = UV^T$
2. Find V is product feature vectors (product x latent features) with random init. of U
3. Find U is user feature vectors (user x latent features) with iteration of approximating V

➔ Iterate to minimize the cost function *J(U, V)*



$$J(U,V) = \sum_{(a,i)\in D} (Y_{ai} - [UV^T]_{ai})^2/2 + \frac{\lambda}{2}\sum_{a=1}^{n}\sum_{j=1}^{k} U_{aj}^2 + \frac{\lambda}{2}\sum_{i=1}^{m}\sum_{j=1}^{k} V_{ij}^2$$
$$= \sum_{(a,i)\in D} (Y_{ai} - u^{(a)} \cdot v^{(i)})^2/2 + \frac{\lambda}{2}\sum_{a=1}^{n} \|u^{(a)}\|^2 + \frac{\lambda}{2}\sum_{i=1}^{m} \|v^{(i)}\|^2$$

Y_{ai} as Confidence value C_{ui}



Assumption : a click on a place implies interest on POI
calculate C_{ui} instead of using number of clicks on POIs (places)

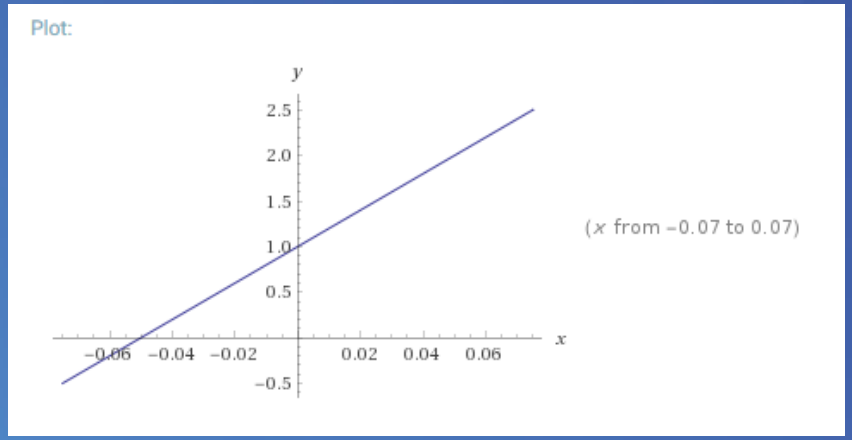
confidence value C_{ui} :

$$C_{ui} = 1 + \alpha * r_{ui}$$

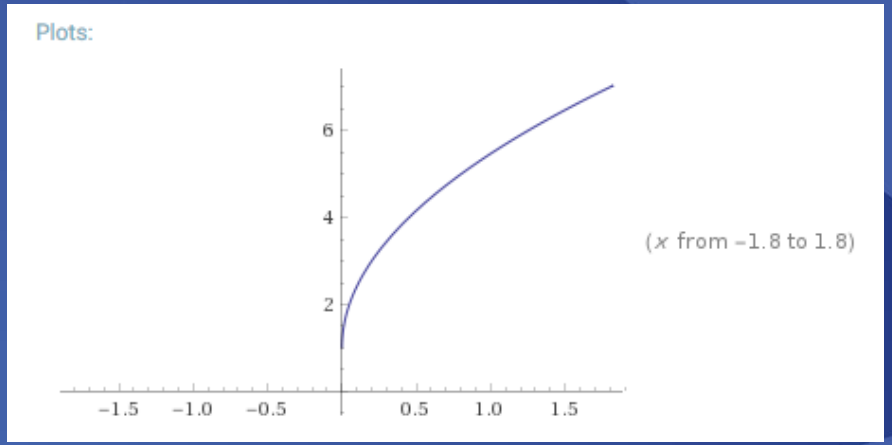
α *is weight on occurrence*
r_{ui} *is occurence of interest* $(i.e. number\ of\ clicks\ on\ POI)$

Reference paper : *"Collaborative Filtering for Implicit Feedback Datasets"* http://yifanhu.net/PUB/cf.pdf

Confidence function : Interest grows linearly or saturated?



$$C_{ui} = 1 + 20 * r_{ui}$$



$$C_{ui} = 1 + \sqrt{20 * r_{ui}}$$

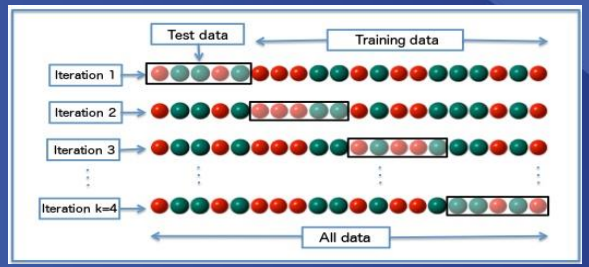
Training and Test data

Training data : Chicago 1 year (2017) user's click data on POIs
Test data : 20% of original data that we know the confidence value

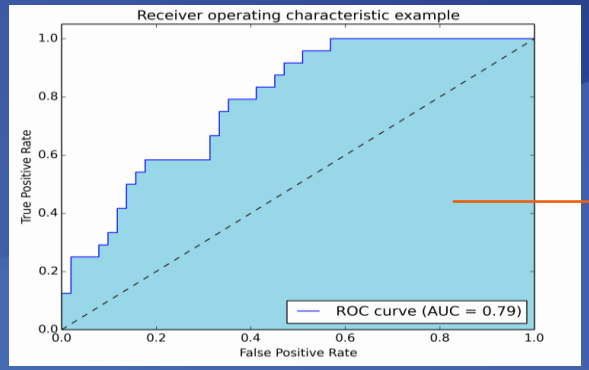
User

Place (POI)						
		(1)			2	(1)
			1			
	1			1		
					Cui	
			1			
	1			(3)		

(1) *Test data*



Cross-validation

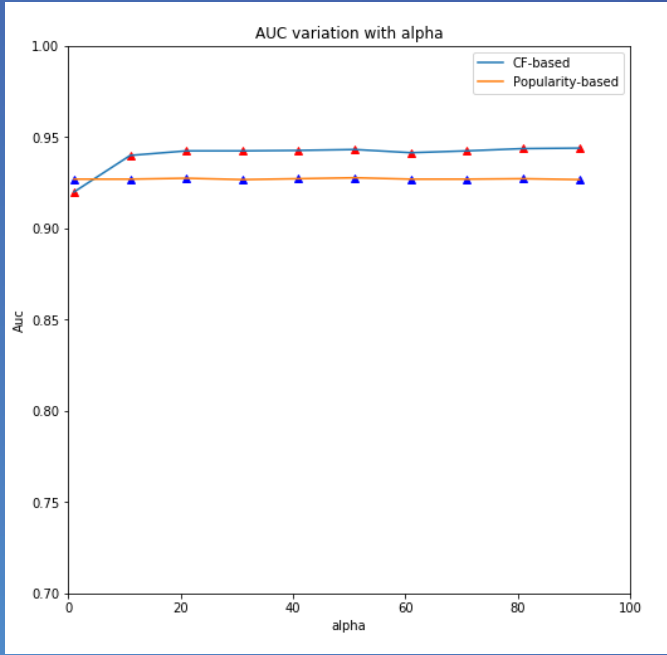


AUC

Cross-validation on alpha α

Confidence function $C_{ui} = 1 + \alpha * r_{ui}$

Cross-validation on α value

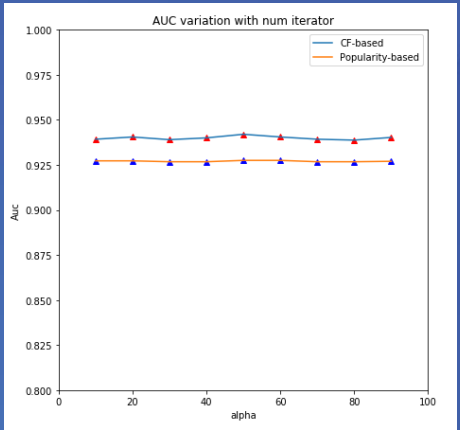


alpha	Collaborative filtering AUC	Popularity baseline AUC
1	0.920	0.927
10	0.940	0.927
20	0.943	0.927
30	0.943	0.927
40	0.943	0.927
50	0.943	0.927
60	0.943	0.928
70	0.942	0.927
80	0.943	0.927
90	0.944	0.927
100	0.944	0.927

*** Popularity-based approach : Get sum of item interactions of user to place and predict that user select the place where has no confidence value with same as the most popular item**

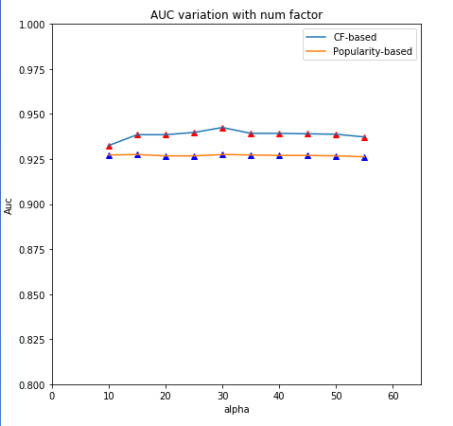
Cross-validation on # of latent features, # of ALS iterations

of latent features



Num Iteration	Collaborative Filtering AUC	Popularity based AUC
10	0.94	0.93
20	0.94	0.93
30	0.94	0.92
40	0.94	0.93
50	0.94	0.92
60	0.93	0.93
70	0.93	0.92
80	0.93	0.92
90	0.94	0.92
100	0.94	0.92

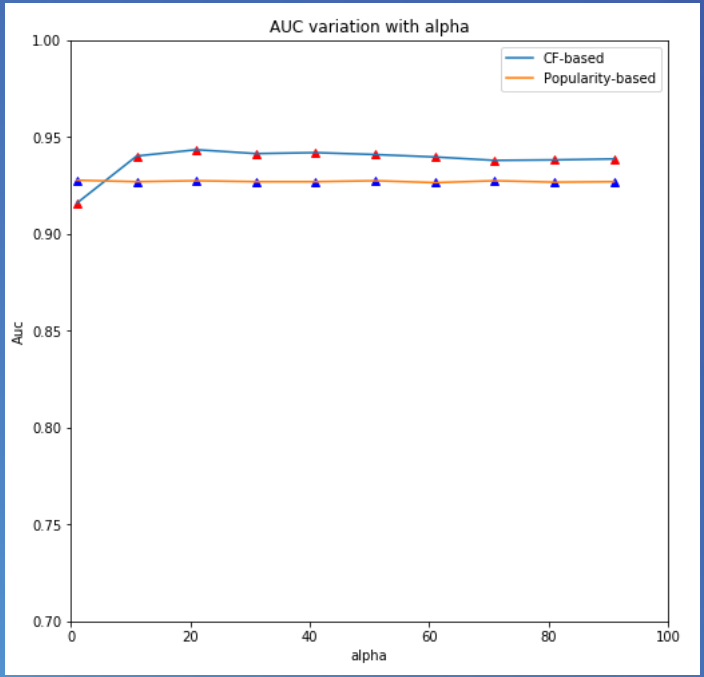
of ALS iterations



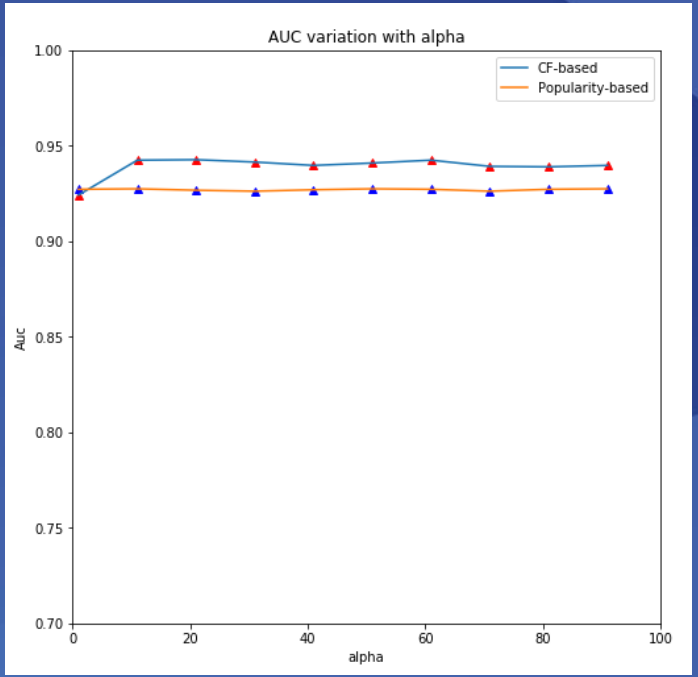
Num latent factors	Collaborative Filtering AUC	Popularity based AUC
10	0.93	0.97
15	0.94	0.92
20	0.94	0.92
25	0.94	0.93
30	0.94	0.92
35	0.93	0.93
40	0.93	0.92
45	0.93	0.92
50	0.94	0.92
55	0.93	0.93
60	0.94	0.92

Various Confidence Value

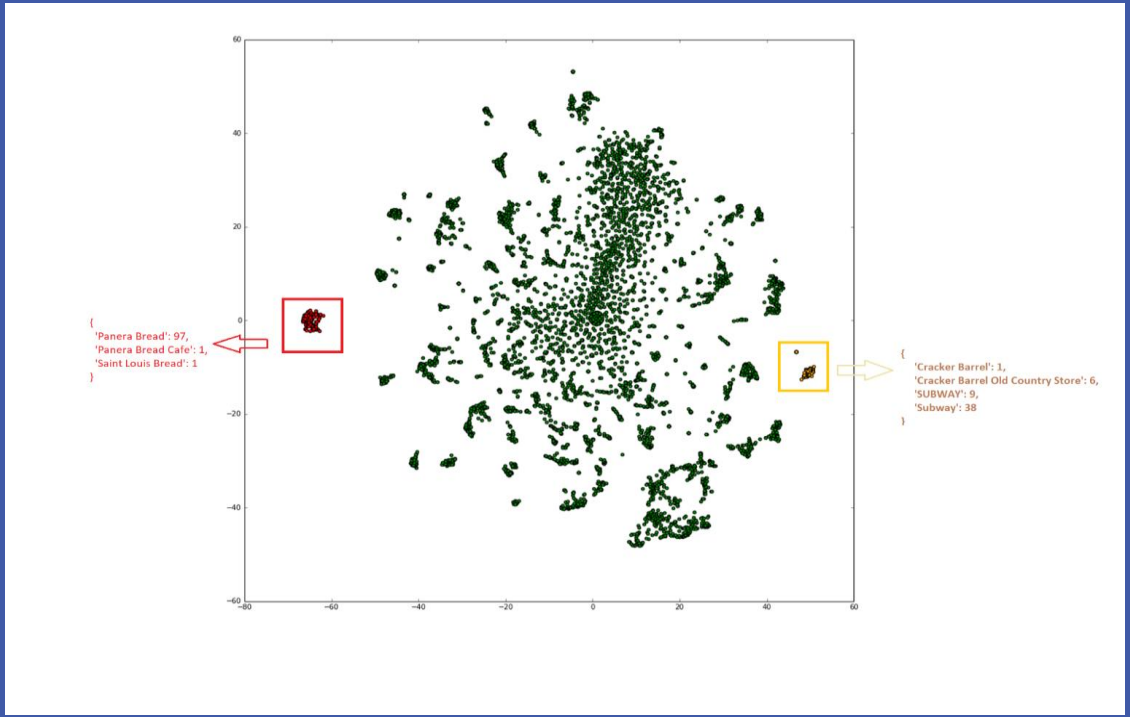
$$C_{ui} = 0 \quad (r_{ui} < threshold)$$
$$C_{ui} = \sqrt{\alpha * r_{ui}} \quad (r_{ui} \geq threshold)$$



$$C_{ui} = 0 \quad (r_{ui} < threshold)$$
$$C_{ui} = 1 + \sqrt{\alpha * r_{ui}} \quad (r_{ui} \geq threshold)$$

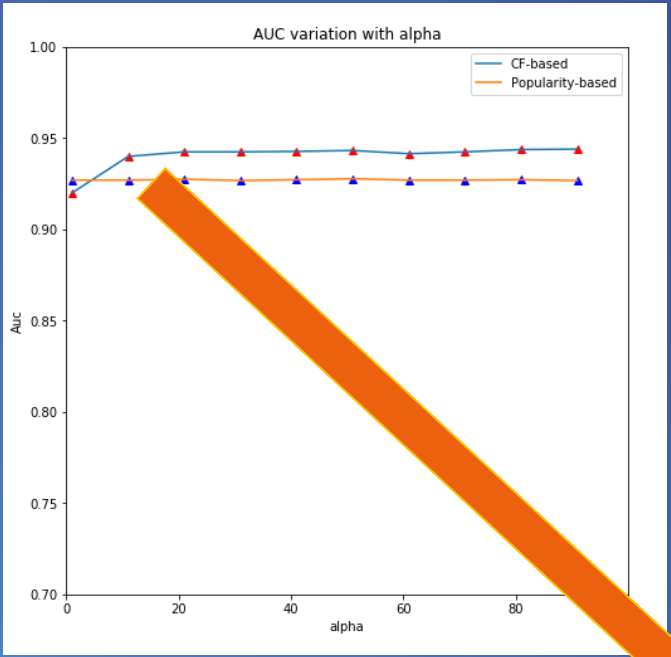


POI-latent feature vector into 2-dimensional space using t-SNE (T-distributed Stochastic Neighbor Embedding)



➔ *similar POIs form clusters*

Data duplication & Performance failure



(user a, POI i, click type_select)
(user a, POI i, click type_drive)
(user a, POI i, click type_save)

(User a , POI i)

AUC value 0.67

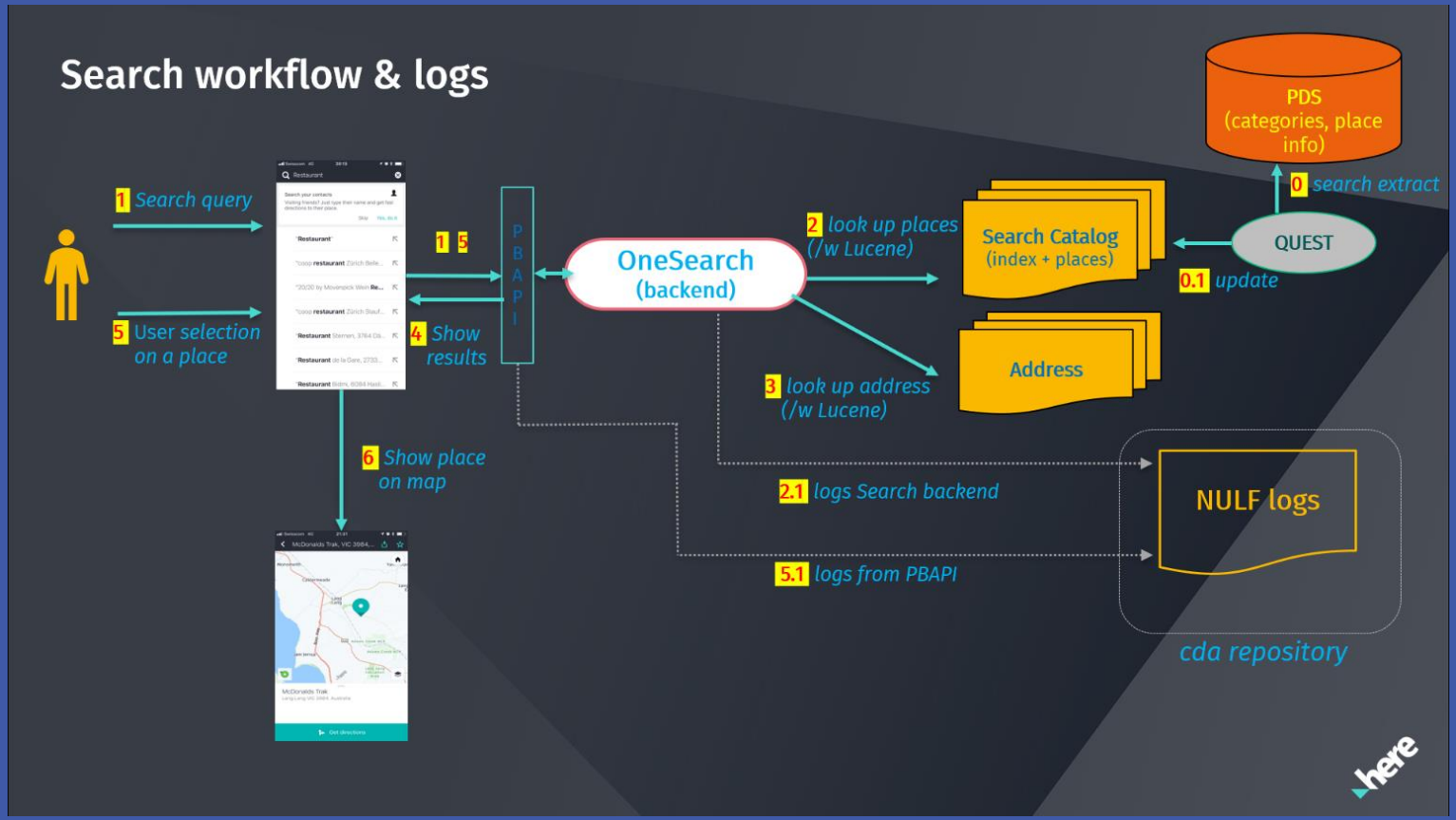
Re-think about personalization

- Can we give personalized recommendation using collaborative filtering?
- By definition, Collaborative Filtering means we borrow other users preference

Personalized  Collaborative Filtering

- Beside, we don’t know much about user
- One source we can guess about user is his QUERY

Local optimization on final recommendations



Guess What?

Top 20 frequent queries in New York City, Berlin, Chicago

'McDonald's'	13767
'Starbucks'	4922
'Dunkin' Donuts'	4779
'Panera Bread'	4774
'food'	3675
'Chipotle'	3584
'Restaurant'	3421
'Portillo's Hot Dogs'	3387
'Whole Foods'	3159
'BURGER KING'	2981
'7-Eleven'	2713
'restaurants'	2572
'Chick-fil-A'	2498
'Portillo's'	2262
'Jimmy John's'	2239
'Taco Bell'	2207
'pizza'	2141
'Lou Malnati's Pizzeria'	2134

(A)

"food"	3353
"cafe"	2982
"bar"	2603
"türkisches restaurant"	2600
"Restaurants"	1630
"restaurant"	1393
"sushi"	1325
"restaurants"	1181
"pizza berlin"	478
"Nearby German"	408
"lokal"	370
"französisches Restaurant"	364
"pizza"	359
"italian restaurant"	357
"retaurant"	304
"vegan berlin"	300
"gaststätte in geltow"	270
"burger"	252
"beer"	224
"drinks"	222

(B)

'McDonald's'	11295
'Starbucks'	7299
'Whole Foods'	6647
'Restaurant'	5772
'Dunkin' Donuts'	4903
'pizza'	4783
'food'	4617
'Wendy's'	4415
'BURGER KING'	3768
'Panera Bread'	3559
'7-Eleven'	3328
'restaurant'	3322
'Chipotle'	2980
'SUBWAY'	2807
'starbucks'	2579
'restaurants'	2351
'Taco Bell'	2208
'White Castle'	2093
'Chinese food'	2080
'Domino's Pizza'	1931

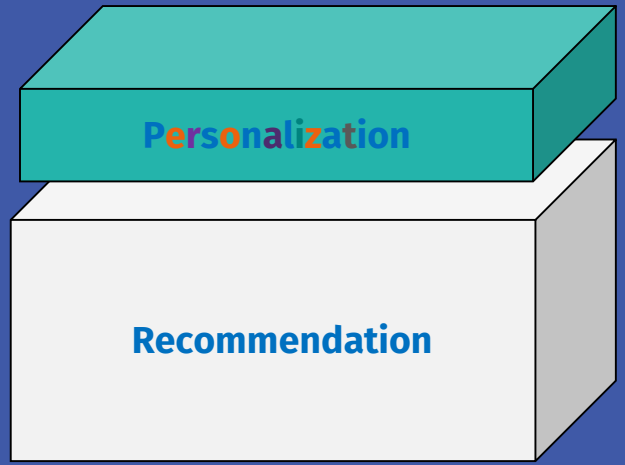
(C)

Hint : query in [ ] is not a name of restaurant chain

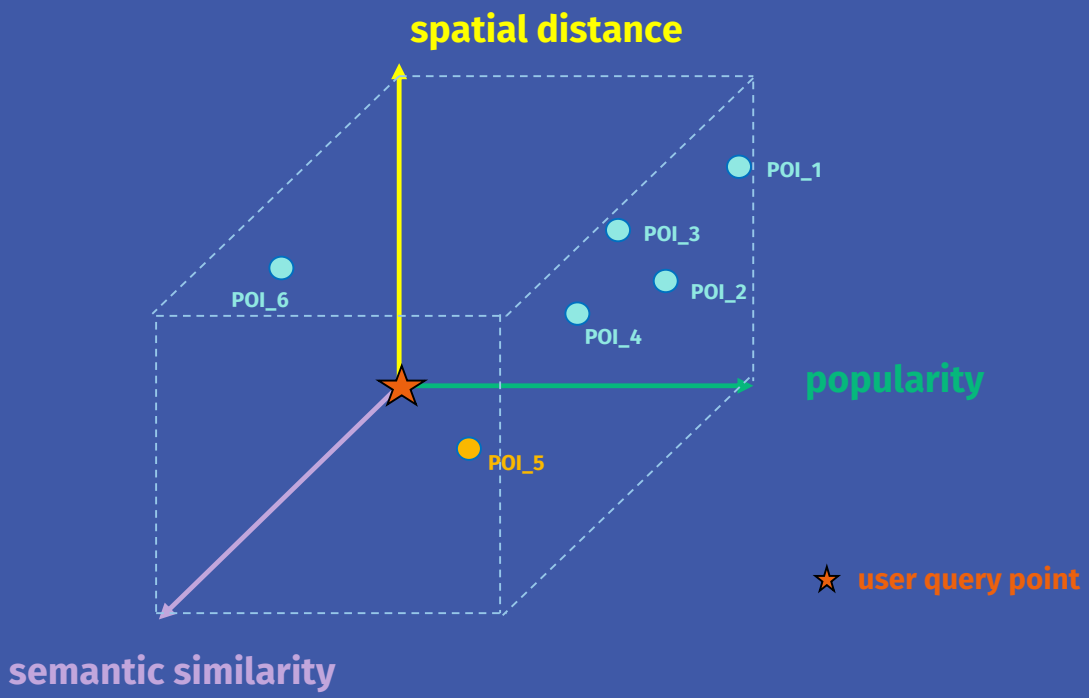
here

Personalization through local optimization

Personalization through local optimization



POI representation in 3-dimensional context space



Dataset : final recommendation places

recommendation POIs given query ‘Grocery’ (data source : Chicago NULF log 2017)

userID	ppid		query	place name	place location	query location
fe989b63-e71f-455e-989b-63e71fb55eda	840dp3m2-2c50c24b302d40138161559ffaf67e82	RECO	1494539929732	'Grocery' 'Grocery' en-MX Whole Foods Market	41.74872 -88.20033 0 10000000.0	41.804887 -88.090918
fe989b63-e71f-455e-989b-63e71fb55eda	840dp3mg-0d6c1cbb294b4761852857809a5b2d54	RECO	1494539929732	'Grocery' 'Grocery' en-MX Whole Foods Market	41.83993 -88.09903 1 10000001.0	41.804887 -88.090918
fe989b63-e71f-455e-989b-63e71fb55eda	840dp3m4-b27d5248eb6a4da098de879fc0c2505b	RECO	1494539929732	'Grocery' 'Grocery' en-MX European Deli & Subs	41.6896104 -88.1272419 2 10000002.0	41.804887 -88.090918
fe989b63-e71f-455e-989b-63e71fb55eda	8403fv6k-a6449fb7f8280ac617010654cdac56ad	RECO	1494539929732	'Grocery' 'Grocery' en-MX Peter Rubi	41.6024155763 -88.2022271844 3 10000003.0	41.804887 -88.090918
fe989b63-e71f-455e-989b-63e71fb55eda	8408lxx5-2ebe2840102d04a16278e86a241a1e3e	RECO	1494539929732	'Grocery' 'Grocery' en-MX Naked Sprout Organics Market & Juice Bar	41.5886448088 -88.057705297 4 10000004.0	41.804887 -88.090918
fe989b63-e71f-455e-989b-63e71fb55eda	840jx7ps-44727cf0a6c9065a13885bbc2c1c1080	RECO	1494539929732	'Grocery' 'Grocery' en-MX Shannon's Butcher & Deli Shoppe	41.88682 -88.01966 5 10000005.0	41.804887 -88.090918
fe989b63-e71f-455e-989b-63e71fb55eda	840aabd1-4c1bc554a3970736f95fbc0c485af48d	RECO	1494539929732	'Grocery' 'Grocery' en-MX Mediterranean Oasis Grocery & Carry Out	41.7405372 -88.1233448 6 10000006.0	41.804887 -88.090918
fe989b63-e71f-455e-989b-63e71fb55eda	840dp3qw-7932ace41d7b4b2dac64c465a2c056b3	RECO	1494539929732	'Grocery' 'Grocery' en-MX Asti Italian Foods	41.96107 -87.96363 7 10000007.0	41.804887 -88.090918
fe989b63-e71f-455e-989b-63e71fb55eda	840dp3mj-d4721cac99994f4c97c0f1350e6039dc	RECO	1494539929732	'Grocery' 'Grocery' en-MX Celina's Deli & Pantry	41.66683 -88.00046 8 10000008.0	41.804887 -88.090918
fe989b63-e71f-455e-989b-63e71fb55eda	840dp3my-6ea7a2d9d0bc4f84ac19f93545acae06	RECO	1494539929732	'Grocery' 'Grocery' en-MX Standard Market Grill Westmont	41.81022 -87.96686 9 10000009.0	41.804887 -88.090918
fe989b63-e71f-455e-989b-63e71fb55eda	840jx7ps-504e317e225a07c60d253887aa4ec75c	RECO	1494539929732	'Grocery' 'Grocery' en-MX Mariano's Fresh Market	41.91953 -87.93978 10 10000010.0	41.804887 -88.090918
fe989b63-e71f-455e-989b-63e71fb55eda	840dp3mw-3355e92e386e42ab82dca66500c64317	RECO	1494539929732	'Grocery' 'Grocery' en-MX Lassak Market & Deli	41.75959 -87.9432 11 10000011.0	41.804887 -88.090918
fe989b63-e71f-455e-989b-63e71fb55eda	840aabd1-de5436280000652d9725296371e9c9c	RECO	1494539929732	'Grocery' 'Grocery' en-MX Standard Market Coffee & Tea	41.7649568 -88.1845081 12 10000012.0	41.804887 -88.090918
fe989b63-e71f-455e-989b-63e71fb55eda	840dp3mm-c94c56b474184a198b8ab1fe452fca6d	RECO	1494539929732	'Grocery' 'Grocery' en-MX Café Smilga	41.73619 -88.01403 13 10000013.0	41.804887 -88.090918
fe989b63-e71f-455e-989b-63e71fb55eda	840jx7ps-4cbed0692fcb060ca77327be2cd26be7	RECO	1494539929732	'Grocery' 'Grocery' en-MX Mang Juan Phillipino Food Mart	41.77692 -87.97471 14 10000014.0	41.804887 -88.090918
fe989b63-e71f-455e-989b-63e71fb55eda	840dp3q5-670623f9924b4588b5c96284fd324d04	RECO	1494539929732	'Grocery' 'Grocery' en-MX Mariano's Fresh Market	41.859736 -88.105809 15 10000015.0	41.804887 -88.090918
fe989b63-e71f-455e-989b-63e71fb55eda	840jx7ps-9b03dc9f089909c40e1d142820e96236	RECO	1494539929732	'Grocery' 'Grocery' en-MX Pete's Fresh Market	41.86044 -87.97523 16 10000016.0	41.804887 -88.090918
fe989b63-e71f-455e-989b-63e71fb55eda	840dp3m2-c72dfb9bd61f4fcba1a25b9662a4977d	RECO	1494539929732	'Grocery' 'Grocery' en-MX Which Wich	41.74713 -88.2012 17 10000017.0	41.804887 -88.090918
fe989b63-e71f-455e-989b-63e71fb55eda	840dp3mw-9684770bd1f74b83a56f165b44a7943f	RECO	1494539929732	'Grocery' 'Grocery' en-MX Whole Foods Market	41.77349 -87.95198 18 10000018.0	41.804887 -88.090918
fe989b63-e71f-455e-989b-63e71fb55eda	840aabd1-6254bfc24295066854e0968d1e450d39	RECO	1494539929732	'Grocery' 'Grocery' en-MX Whole Foods Market	41.8942358 -87.9626287 19 10000019.0	41.804887 -88.090918
fe989b63-e71f-455e-989b-63e71fb55eda	840dp3m8-0c0433f0bd3d43d89d50eeb3f270cac9	RECO	1494539929732	'Grocery' 'Grocery' en-MX Dominick's	41.75933 -88.24166 20 10000020.0	41.804887 -88.090918
fe989b63-e71f-455e-989b-63e71fb55eda	840jx7ps-047cc0d7a3df03c5b252e92fe362d6ea	RECO	1494539929732	'Grocery' 'Grocery' en-MX Walmart Supercenter	41.74582 -88.19782 21 10000021.0	41.804887 -88.090918
fe989b63-e71f-455e-989b-63e71fb55eda	840dp3mt-e102fd7333a2468bb72672cd99c54c8e	RECO	1494539929732	'Grocery' 'Grocery' en-MX Walmart Supercenter	41.7492852 -87.999651 22 10000022.0	41.804887 -88.090918
fe989b63-e71f-455e-989b-63e71fb55eda	8403fv6k-fdce54a2c3c90783706de5c522a492a4	RECO	1494539929732	'Grocery' 'Grocery' en-MX Chicago Industrial Food Ingredient	41.794884 -88.124838 23 10000023.0	41.804887 -88.090918
fe989b63-e71f-455e-989b-63e71fb55eda	840dp3kb-397198cfaff4474d910bd44a03bbdb96	RECO	1494539929732	'Grocery' 'Grocery' en-MX Walmart Plainfield - S Route 59	41.65377 -88.20637 24 10000024.0	41.804887 -88.090918
fe989b63-e71f-455e-989b-63e71fb55eda	840jx7ps-f20940d1cb890d7c284ef348d0dc8cbf	RECO	1494539929732	'Grocery' 'Grocery' en-MX Wheaton Food Mart & Tobacco	41.85889 -88.08226 25 1000000nan	41.804887 -88.090918
fe989b63-e71f-455e-989b-63e71fb55eda	8403fv6k-866308ab94290509777b52a237c1454a	RECO	1494539929732	'Grocery' 'Grocery' en-MX Fotis & Son Imports Chicago	41.988174 -87.944006 26 1000000nan	41.804887 -88.090918

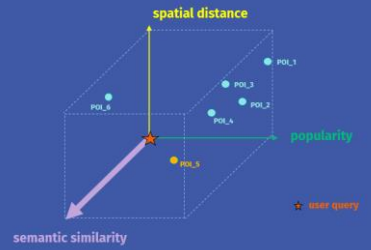
POI = [semantic similarity, spatial distance, popularity]

here

Semantic similarity – Semantic distance between query and POIs

POIs

1. No place description
2. inconsistent categories and contradictory food type

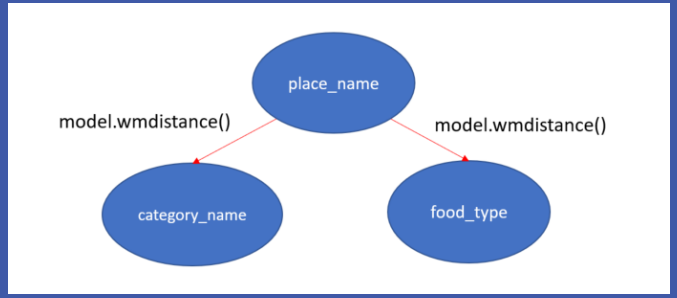


ppid_int	place_name	category_names	food_type
878	7-Eleven	[" 'Convenience Store'", " 'Restaurant'", " 'ATM'"]	[]
8045	7-Eleven	[" 'Petrol-Gasoline Station'", " 'Convenience Store'", " 'Restaurant'"]	[]
3241	7-Eleven	[" 'Convenience Store'", " 'Deli'", " 'Restaurant'"]	[" 'Sandwich'"]
3245	7-Eleven	[" 'ATM'", " 'Convenience Store'", " 'Restaurant'"]	[]
14056	7-Eleven	[" 'Convenience Store'", " 'Restaurant'", " 'ATM'"]	[]
9086	7-Eleven	[" 'Specialty Store'", " 'Convenience Store'", " 'Fast Food'", " 'Restaurant'"]	[" 'Pizza'"]
8947	7-Eleven	[" 'Convenience Store'", " 'Restaurant'"]	[" 'Pizza'", " 'Sandwich'"]
3363	7-Eleven	[" 'Petrol-Gasoline Station'", " 'Convenience Store'", " 'Restaurant'"]	[]
8036	7-Eleven	[" 'Convenience Store'", " 'Restaurant'", " 'ATM'"]	[]
11224	7-Eleven	[" 'Convenience Store'", " 'Restaurant'"]	[]
6172	7-Eleven	[" 'Convenience Store'", " 'Casual Dining'", " 'Restaurant'", " 'ATM'"]	[" 'International'"]
11226	7-Eleven	[" 'Convenience Store'", " 'Restaurant'"]	[]
8043	7-Eleven	[" 'Convenience Store'", " 'Restaurant'"]	[]
8045	7-Eleven	[" 'Petrol-Gasoline Station'", " 'Convenience Store'", " 'Restaurant'"]	[]

name_chosen	result_index	distance	category_names	food_type
New Rebozo Chicago	0	1.415389709	[" 'Restaurant'"]	[" 'Mexican'"]
Buzz Bait Taqueria	1	0.306916712	[" 'Fast Food'", " 'Taqueria'", " 'Food Market-Stall'", " 'Restaurant'"]	[" 'Mexican'", " 'Seafood'"]
Velvet Taco Chicago	2	0.736410394	[" 'Fast Food'", " 'Restaurant'"]	[" 'American'", " 'Mexican'"]
Seoul Taco	3	1.267746047	[" 'Casual Dining'", " 'Restaurant'"]	[" 'Mexican'", " 'Asian'", " 'Korean'"]
Taco Joint River North	4	1.548860881	[" 'Casual Dining'", " 'Fast Food'", " 'Taqueria'", " 'Restaurant'"]	[" 'Mexican'"]
Chipotle	5	0.610077149	[" 'Fast Food'", " 'Restaurant'"]	[" 'Mexican'"]
Adobo Grill Old Town	6	0.419497398	[" 'Casual Dining'", " 'Restaurant'", " 'Bar or Pub'"]	[" 'American-Southwestern'", " 'Mexican'", " 'Vegetarian'", " 'Grill'"]
North & Clark CafÃ©	7	0.519407544	[" 'Casual Dining'", " 'Restaurant'", " 'Coffee Shop'"]	[" 'American'"]
Adobe Gila's	8	0.942252066	[" 'Restaurant'"]	[" 'American'", " 'American-Southwestern'", " 'Mexican'"]
Foodlife	9	1.36662062	[" 'Casual Dining'", " 'Restaurant'"]	[" 'American'", " 'Mexican'", " 'Chinese'", " 'Pizza'", " 'International'"]
Downtown Dogs	10	1.357212909	[" 'Fast Food'", " 'Restaurant'"]	[" 'American'", " 'Mexican'", " 'Hot Dogs'"]
Blue Agave Tequila Bar & Restaurant	11	0.808963288	[" 'Casual Dining'", " 'Restaurant'", " 'Bar or Pub'"]	[" 'American'", " 'American-Southwestern'", " 'Mexican'"]
Sedgwick's Bar & Grill	12	1.214224529	[" 'Casual Dining'", " 'Restaurant'", " 'Bar or Pub'"]	[" 'American'", " 'American-Southwestern'", " 'Mexican'", " 'Grill'"]
Zia	13	1.285569986	[" 'Restaurant'"]	[" 'Mexican'"]
Dublin's Bar & Grill	14	0.808963288	[" 'Casual Dining'", " 'Nightlife-Entertainment'", " 'Restaurant'", " 'Bar or Pub'"]	[" 'Mexican'", " 'Irish'", " 'Grill'"]
Salpicon	15	0.206274781	[" 'Fine Dining'", " 'Restaurant'"]	[" 'Mexican'"]
Flaco's Tacos	16	1.327053537	[" 'Casual Dining'", " 'Restaurant'"]	[" 'Mexican'", " 'Chicken'"]
Nacional 27	17	1.42418577	[" 'Casual Dining'", " 'Dancing'", " 'Restaurant'"]	[" 'Mexican'", " 'Cuban'", " 'Spanish-Tapas'", " 'South American'", " 'Latin American'", " 'International'"
Tallboy Taco	18	1.404815148	[" 'Fast Food'", " 'Restaurant'", " 'Bar or Pub'"]	[" 'Mexican'"]
Chipotle	19	1.788856131	[" 'Fast Food'", " 'Restaurant'"]	[" 'Mexican'"]
Su Casa Mexican Restaurant	20	1.698143879	[" 'Casual Dining'", " 'Restaurant'"]	[" 'American-Southwestern'", " 'Mexican'"]
Chipotle	21	0.465271672	[" 'Fast Food'", " 'Restaurant'"]	[" 'Mexican'"]
Patron's Hacienda	22	1.457809211	[" 'Casual Dining'", " 'Dancing'", " 'Restaurant'", " 'Bar or Pub'", " 'Take Out and Delivery Only'"]	[" 'Mexican'"]
Downtown Pizza	23	0.258072038	[" 'Fast Food'", " 'Restaurant'"]	[" 'Pizza'"]
Taco King	24	0.47842907	[" 'Restaurant'"]	[" 'Mexican'"]
Burrito Bistro	26	1.244444824	[" 'Restaurant'"]	[" 'Mexican'"]

Recommended POIs on query ‘Mexican restaurants near my location’

word2vec, doc2vec for measuring semantic similarity

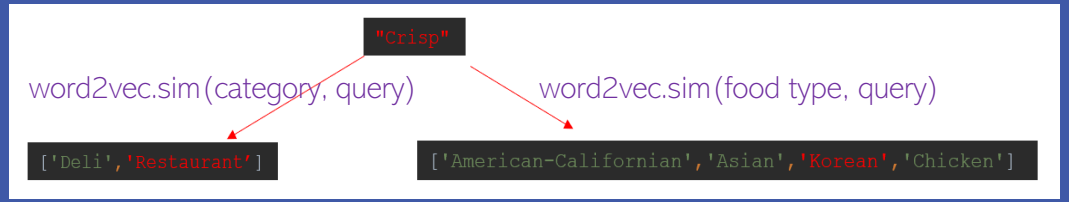


POI description = preposition (from food types) + noun (from categories)

[Example]

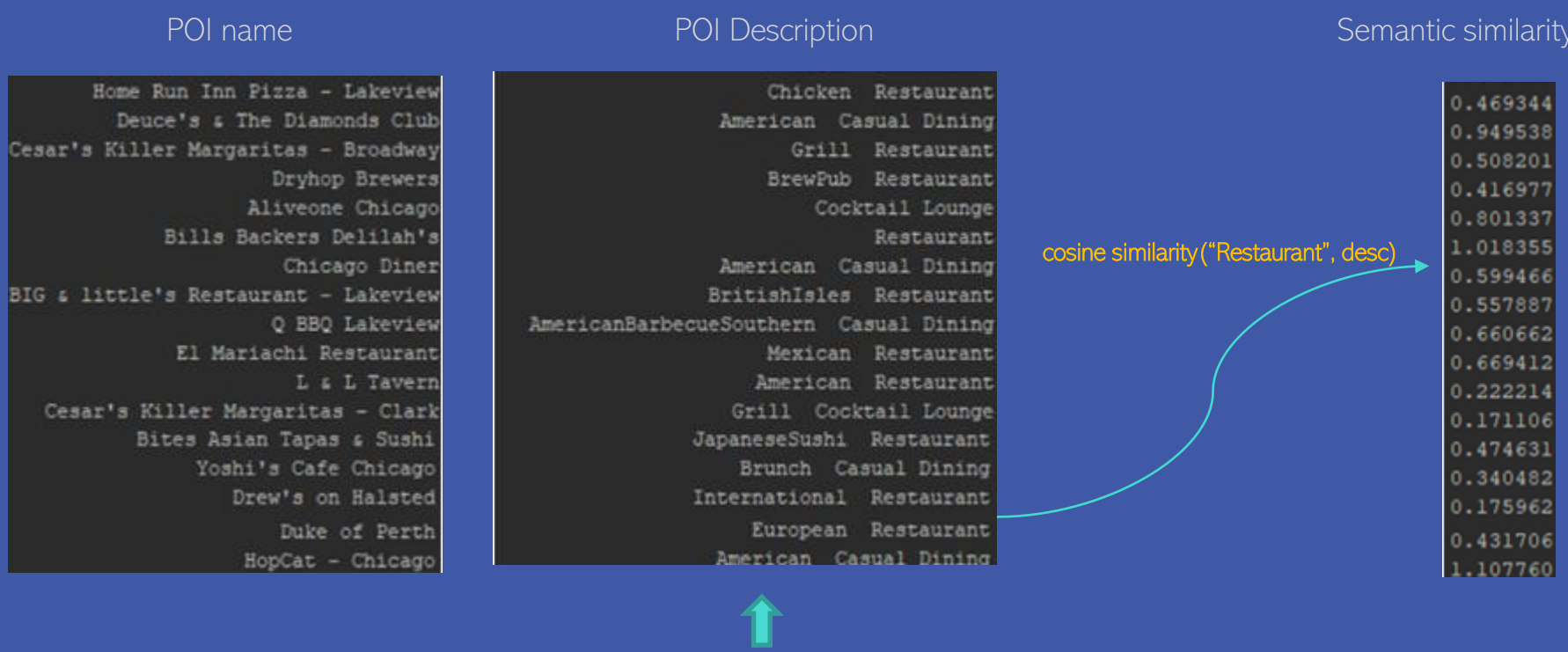
query_raw	result_name	place_name	category_names	food_type
restaurant	840dp3wt-92cd8484bd004c478c658f8e81ab6edf	Vegetarian Express	[" 'Casual Dining'", " 'Food-Beverage Specialty Store'", " 'Fast Food'", " 'Restaurant'"]	[" 'Vegan'", " 'Vegetarian'"]
restaurant	840dp3wt-b08ea990ec684846ab81c9aeba45d8eb	Crisp	[" 'Deli'", " 'Restaurant'"]	[" 'American-Californian'", " 'Asian'", " 'Korean'", " 'Chicken'"]

"vegetarian food-beverage specialty store"



"Korean Restaurant"

Measuring semantic similarity using doc2vec



POI description = preposition (from food types) + noun (from categories)

Semantic distance

Similarity value : similarity between two document vectors
Semantic distance : the shorter, more similar

scaled_similarity_val
1.192506
4.768913
0.504470
3.458025
1.643256
1.807787
2.205364
3.238897
2.739563
2.102435
2.673313
0.000000
5.000000
3.119741
1.759413
2.259980

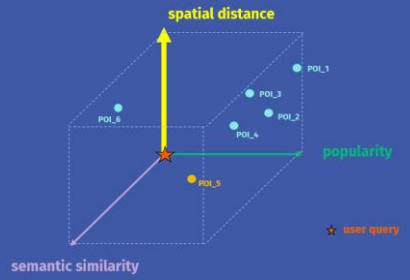
$$normalized\_sim = (5.0 - 0.0)\frac{Sim - Sim_{min}}{Sim - Sim_{min}} + 0.0$$

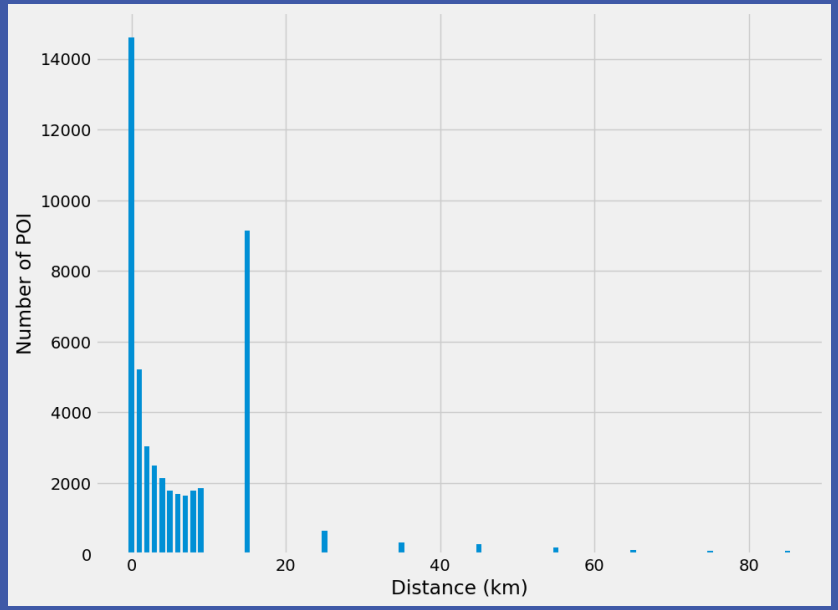
$$MAX_{sim} - normalized\_sim$$

semantic_distance
3.973011
0.243044
4.607105
1.599590
3.311137
3.199646
2.733324
1.727418
2.201356
2.853906
2.258425
5.000000
0.000000
1.803167
3.195022

Spatial distance to query location

- queries asked Chicago 2017
  “stores **near** me” , “library **near** me”, “**nearest** gas station” , “**Nearest** Airport”
  “african restaurant **near** me”, “mosque **near** me Chicago”, “best restaurant **near** me”
  “mexican restaurants **near** my location”, “lodging **near** arlington heights, IL”





query : 'pizza near me'	Place name	Distance
	Guerrero's Pizza	0.871357
	Barraco's Pizza	0.149527
	Pizza Tango	1.055590
	Domino's	1.402527
	Little Caesars	1.279179
	Del Campo Pizzeria	1.096107
	3050 W Taylor St	0.944542
	Lou Malnati's Pizzeria	6.354933
	Giordano's	7.266272
	Pequod's Chicago Pizza	7.348856
	Chicago Pizza And Oven Grinder Company	8.403900
	UNO Pizzeria & Grill	6.980681
	Piece Brewery and Pizzeria	5.817650
	Lou Malnati's Pizzeria	7.493320
	Pizano's Pizza & Pasta Frozen Division	6.586690
	The Original Gino’s East	7.441764
	Guthrie's Tavern	10.067587
	Lou Malnati's Pizzeria - West Loop	4.293820
	Pizano's Pizza & Pasta	7.242685
	Pizzeria Ora - Chicago Style Pizza	6.565961
	My Pie Pizza	6.589319
	Osteria la Madia	6.691985
	Luke's Italian Beef	5.747071
	Pompei Taylor Street	2.988695
	Simone's Bar	4.066894
	Fatpour Tap Works	4.994860

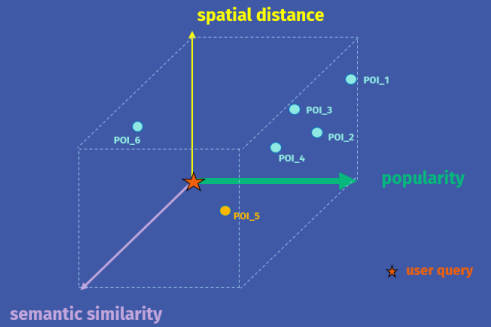
Popularity

- Obtain multiple sources of rates on POIs (e.g. TripAdvisor, Yelp, Facebook)
- Normalize rate values into range (0.0, 5.0)

$$normalized\_r = (5.0 - 0.0)\frac{r - r_{min}}{r_{max} - r_{min}} + 0.0$$

- Inversed rate value : max_rate – normalized rate
- Average rate of Inverse rate value

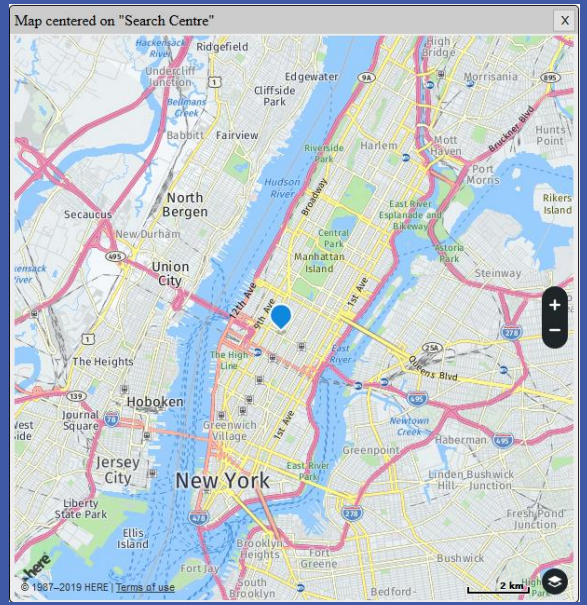
$$\frac{1}{n}\sum_{i=1}^{n}(MAX_{rate} - normalized\_r)$$



Query : “Museum”
Viewport : New York City

q_lat	q_lon	distance_	result_index	name_chosen	lat	lon	category_names
40.7307	-73.996	960	1	New Museum	40.7224	-73.993	Other Museum, Art Museum, Other Landmark Attraction
40.7307	-73.996	458	2	Merchant's House Museum	40.7277	-73.9923	Other Museum, History Museum, Other Landmark-Attraction
40.7307	-73.996	878	3	International Center of Photography Museu	40.7231	-73.9934	Art Museum, Other Museum, Other Landmark-Attraction
40.7307	-73.996	825	4	Center for Jewish History	40.738	-73.9939	Other Museum, History Museum, Tourist Attraction, Historical Monument, Other Landmark-Attraction
40.7307	-73.996	1057	5	Afa Gallery	40.7222	-74.0016	Art Museum, Other Museum, Gallery, Other Landmark-Attraction
40.7307	-73.996	1065	6	Theodore Roosevelt Birthplace	40.7387	-73.989	Other Museum, History Museum, Tourist Attraction, Historical Monument, Other Landmark-Attraction
40.7307	-73.996	1055	7	Rubin Museum of Art	40.7401	-73.9977	Other Museum, Art Museum, Other Landmark-Attraction
40.7307	-73.996	1114	8	Deitch Projects	40.722	-74.0026	Other Shop, Art Museum, Gallery, Historical Monument, Other Landmark-Attraction
40.7307	-73.996	627	9	Ukrainian Museum	40.7277	-73.9898	Other Museum, Art Museum, Other Landmark-Attraction
40.7307	-73.996	1155	10	Museum of Illusions	40.7397	-74.003	Children's Museum, Art Museum, Other Museum, Other Landmark-Attraction
40.7307	-73.996	1118	11	Artists Space	40.7218	-74.002	Other Museum, Book Store, Art Museum, Other Theater, Music and Culture, Gallery
40.7307	-73.996	221	12	Margo Feiden Galleries	40.7325	-73.9949	Art Museum, Other Museum, Gallery, Other Landmark-Attraction
40.7307	-73.996	1088	13	KGB Spy Museum	40.7396	-74.0014	History Museum, Other Museum, Other Landmark-Attraction
40.7307	-73.996	1091	14	Fusion Arts Museum	40.722	-73.9902	Art Museum, Other Museum, Other Landmark-Attraction
40.7307	-73.996	417	15	American Artist Professional League	40.7343	-73.9946	Art Museum, Gallery
40.7307	-73.996	1093	16	Fitzgerald Fine Arts	40.722	-74.002	Art Museum, Other Museum, Gallery, Other Landmark-Attraction
40.7307	-73.996	639	17	Fortnight Institute	40.7266	-73.9907	Art Museum
40.7307	-73.996	899	18	Museum of Beautiful People	40.7239	-73.9902	Art Museum
40.7307	-73.996	944	19	TMJ Arts Collective	40.7231	-74.0009	Art Museum, Gallery
40.7307	-73.996	513	20	Eva Presenhuber	40.7268	-73.9929	Art Museum, Other Museum, Gallery, Other Landmark-Attraction
40.7307	-73.996	862	21	Judd Foundation	40.7234	-73.9995	Art Museum, Other Museum, Gallery, Other Landmark-Attraction
40.7307	-73.996	663	22	Leslie+Lohman Prince Street Project	40.7257	-74.0001	Art Museum, Gallery
40.7307	-73.996	3730	23	Museum of Modern Art	40.7611	-73.9772	Tourist Attraction, Cinema, Art Museum, Other Museum, Other Theater, Music and Culture, Other Landm
40.7307	-73.996	6004	24	Metropolitan Museum of Art, New York	40.7784	-73.9625	Art Museum, Gallery
40.7307	-73.996	6591	25	Solomon R Guggenheim Museum	40.783	-73.9591	Other Museum, Art Museum, Residential Area-Building, Other Landmark-Attraction
40.7307	-73.996	1488	26	Whitney Museum of American Art	40.7395	-74.0094	Other Museum, Art Museum, Library, Tourist Attraction, Gallery, Other Landmark-Attraction
40.7307	-73.996	5088	27	Frick Collection	40.7709	-73.9672	Other Museum, Art Museum, Other Theater, Music and Culture, Residential Area-Building, Gallery, Othe
40.7307	-73.996	7155	28	Brooklyn Museum of Art	40.6715	-73.9628	Other Museum, Art Museum, Gallery, Historical Monument, Other Landmark-Attraction
40.7307	-73.996	2366	29	Morgan Library and Museum	40.7491	-73.9818	Other Museum, Art Museum, Library, Other Landmark-Attraction
40.7307	-73.996	6377	30	Neue Galerie New York	40.7813	-73.9603	Other Museum, Book Store, Gift, Antique and Art, History Museum, Art Museum, Other Landmark-Attract
40.7307	-73.996	12226	31	New York Hall Of Science	40.7471	-73.8525	Other Museum, Science Museum, Tourist Attraction, Other Landmark-Attraction
40.7307	-73.996	1154	32	Children's Museum Of The Arts	40.7205	-73.9986	Children's Museum, Tourist Attraction, Other Museum
40.7307	-73.996	4504	33	New York Transit Museum	40.6905	-73.99	History Museum, Other Theater, Music and Culture, Other Government or Community Facility, Tourist Att
40.7307	-73.996	9559	34	Studio Museum in Harlem	40.8085	-73.9475	Other Museum, Art Museum, Tourist Attraction, Other Landmark-Attraction
40.7307	-73.996	7808	35	Museum of the City of New York	40.7925	-73.952	Other Museum, History Museum, Other Theater, Music and Culture, Other Landmark-Attraction
40.7307	-73.996	5766	36	Hayden Planetarium	40.7797	-73.9734	Other Museum, Science Museum, Historical Monument, Other Landmark-Attraction
40.7307	-73.996	5611	37	Liberty Science Center	40.7078	-74.0553	Other Museum, Cinema, Science Museum, Other Theater, Music and Culture, Tourist Attraction, Other La
40.7307	-73.996	3092	38	Discovery Times Square	40.7577	-73.9873	Other Museum, History Museum
40.7307	-73.996	6880	39	Jewish Museum	40.7852	-73.9573	Other Museum, Art Museum, Other Theater, Music and Culture, Tourist Attraction, Gallery, Other Landm
40.7307	-73.996	4887	40	American Folk Art Museum	40.7733	-73.9817	Other Museum, Art Museum, Tourist Attraction, Other Landmark-Attraction

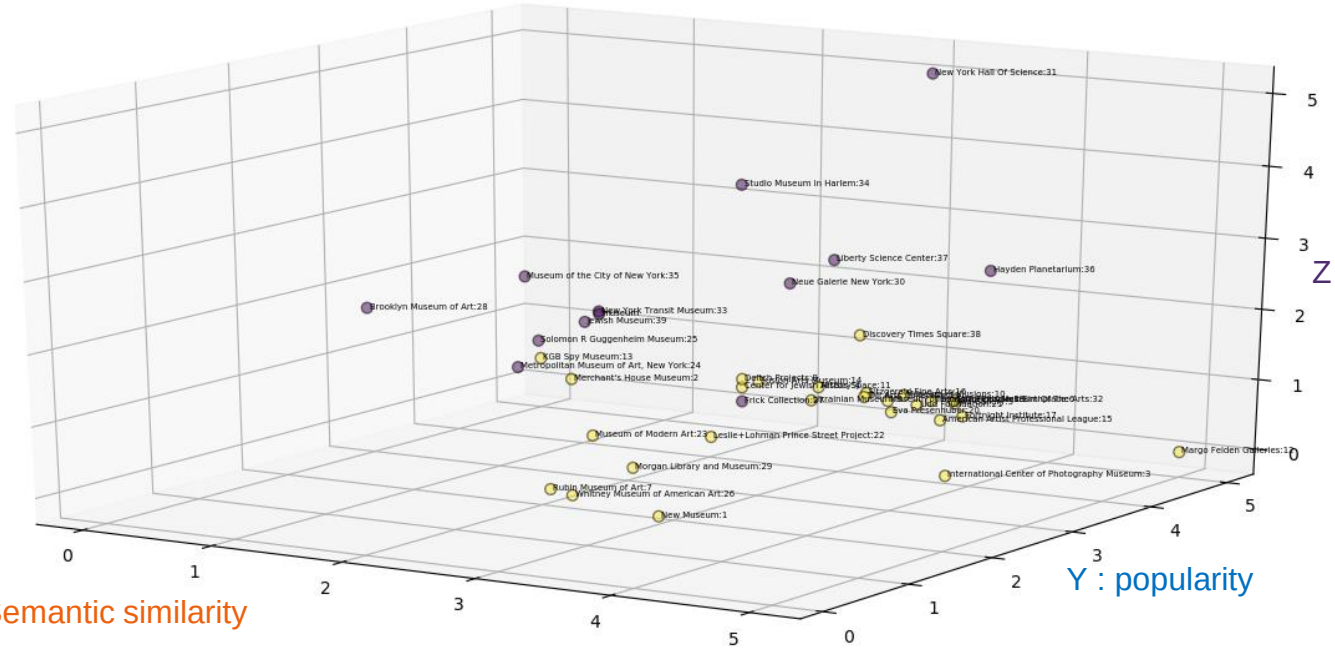
recommendation places on query 'museum' in NYC



Query location

POI = [semantic similarity, spatial distance, popularity]

Recommendation POIs (NYC museums) in query context space



X : Semantic similarity

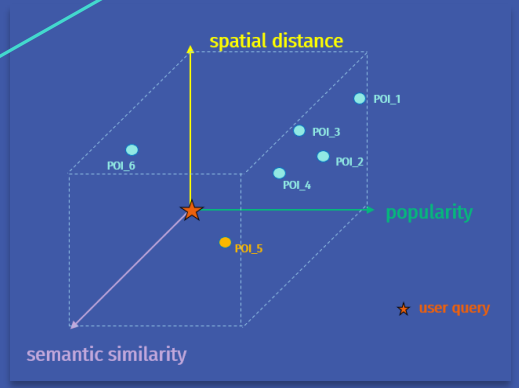
Y : popularity

Z : spatial distance

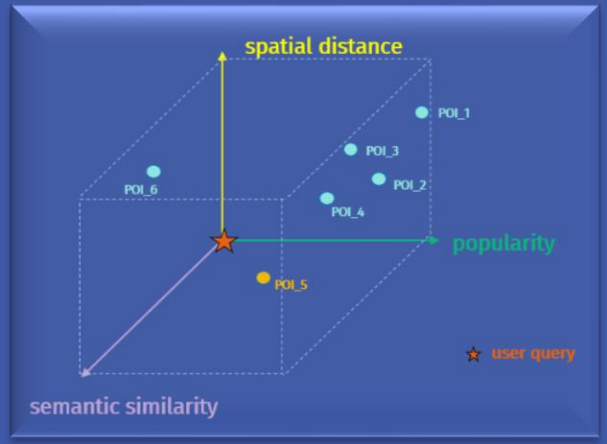
K-means clustering

Recommendation re-ordered by contextual distance in semantic-spatial-popularity space

distance_	result_index	name_chosen
960	1	New Museum
458	2	Merchant's House Museum
878	3	International Center of Photography Museu
825	4	Center for Jewish History
1057	5	Afa Gallery
1065	6	Theodore Roosevelt Birthplace
1055	7	Rubin Museum of Art
1114	8	Deitch Projects
627	9	Ukrainian Museum
1155	10	Museum of Illusions
1118	11	Artists Space
221	12	Margo Feiden Galleries
1088	13	KGB Spy Museum
1091	14	Fusion Arts Museum
417	15	American Artist Professional League
1093	16	Fitzgerald Fine Arts
639	17	Fortnight Institute
899	18	Museum of Beautiful People
944	19	TMJ Arts Collective
513	20	Eva Presenhuber
862	21	Judd Foundation
663	22	Leslie+Lohman Prince Street Project
3730	23	Museum of Modern Art
6004	24	Metropolitan Museum of Art, New York
6591	25	Solomon R Guggenheim Museum
1488	26	Whitney Museum of American Art
5088	27	Frick Collection
7155	28	Brooklyn Museum of Art
2366	29	Morgan Library and Museum
6377	30	Neue Galerie New York
12226	31	New York Hall Of Science
1154	32	Children's Museum Of The Arts
4504	33	New York Transit Museum
9559	34	Studio Museum in Harlem
7808	35	Museum of the City of New York
5766	36	Hayden Planetarium
5611	37	Liberty Science Center
3092	38	Discovery Times Square
6880	39	Jewish Museum
4887	40	American Folk Art Museum



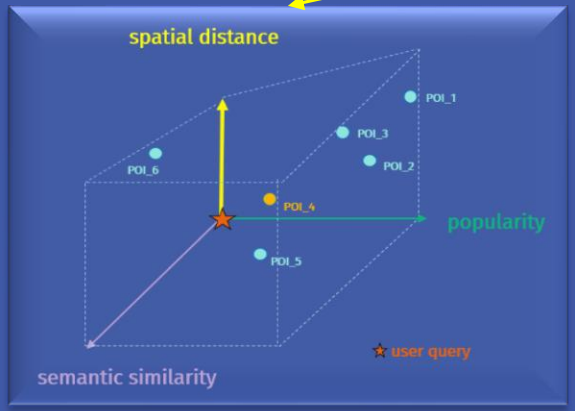
original_rank	name_chosen	distance_in_3d
7	Rubin Museum of Art	2.960503585
26	Whitney Museum of American Art	3.328606081
28	Brooklyn Museum of Art	3.49653933
29	Morgan Library and Museum	3.708922969
23	Museum of Modern Art	3.812016077
1	New Museum	3.84352818
24	Metropolitan Museum of Art, New York	3.973314957
40	American Folk Art Museum	4.045054129
25	Solomon R Guggenheim Museum	4.085124956
35	Museum of the City of New York	4.171340831
33	New York Transit Museum	4.214433757
22	Leslie+Lohman Prince Street Project	4.219482346
39	Jewish Museum	4.262505434
27	Frick Collection	4.865965489
2	Merchant's House Museum	5.007465215
13	KGB Spy Museum	5.013039959
30	Neue Galerie New York	5.216194487
4	Center for Jewish History	5.269443445
8	Deitch Projects	5.305872099
14	Fusion Arts Museum	5.324920236
3	International Center of Photography Museur	5.383101961
9	Ukrainian Museum	5.464678703
11	Artists Space	5.500076398
19	TMJ Arts Collective	5.653880834
16	Fitzgerald Fine Arts	5.658006157
20	Eva Presenhuber	5.739949411
18	Museum of Beautiful People	5.749962701
38	Discovery Times Square	5.751233977
34	Studio Museum in Harlem	5.756837282
10	Museum of Illusions	5.795685979
21	Judd Foundation	5.846686848
15	American Artist Professional League	5.950175595
37	Liberty Science Center	5.950658188
6	Theodore Roosevelt Birthplace	5.964634827
5	Afa Gallery	5.991657094
32	Children's Museum Of The Arts	6.00467816
17	Fortnight Institute	6.052078377
36	Hayden Planetarium	6.618170235
12	Margo Feiden Galleries	7.071067812
31	New York Hall Of Science	7.700919545

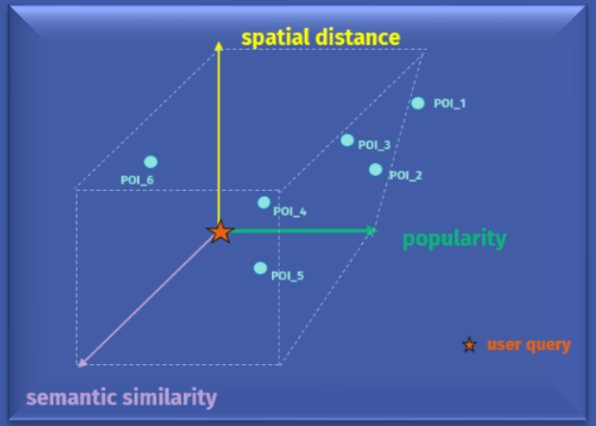


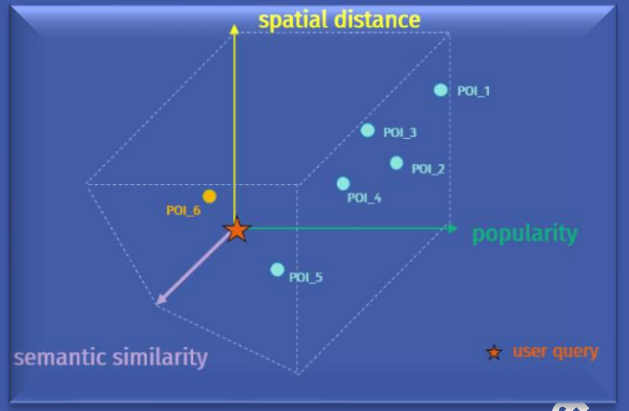
shorter distance

more popular

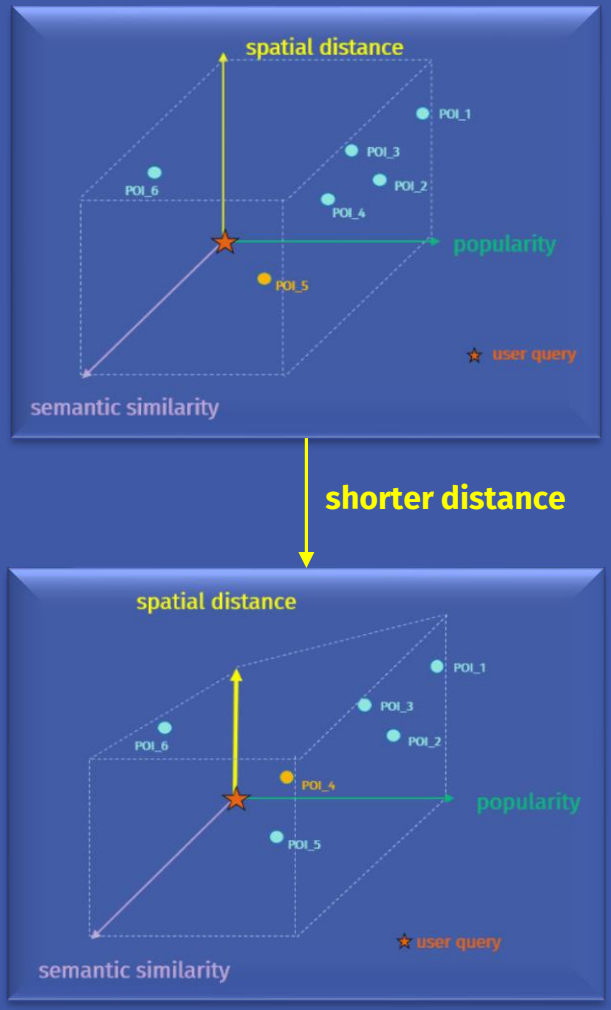
query semantic context





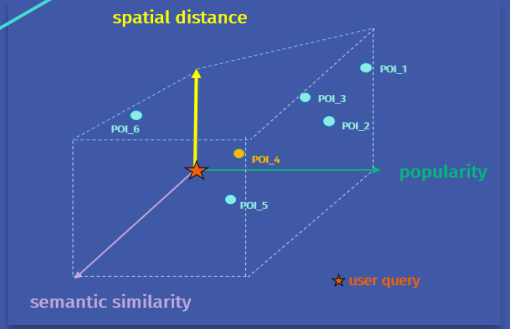


Query “museum near me”



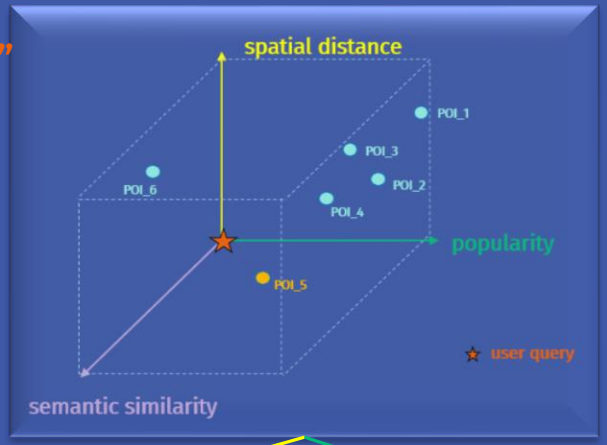
Weight on spatial distance

distance_	result_index	name_chosen
960	1	New Museum
458	2	Merchant's House Museum
878	3	International Center of Photography Museu
825	4	Center for Jewish History
1057	5	Afa Gallery
1065	6	Theodore Roosevelt Birthplace
1055	7	Rubin Museum of Art
1114	8	Deitch Projects
627	9	Ukrainian Museum
1155	10	Museum of Illusions
1118	11	Artists Space
221	12	Margo Feiden Galleries
1088	13	KGB Spy Museum
1091	14	Fusion Arts Museum
417	15	American Artist Professional League
1093	16	Fitzgerald Fine Arts
639	17	Fortnight Institute
899	18	Museum of Beautiful People
944	19	TMJ Arts Collective
513	20	Eva Presenhuber
862	21	Judd Foundation
663	22	Leslie+Lohman Prince Street Project
3730	23	Museum of Modern Art
6004	24	Metropolitan Museum of Art, New York
6591	25	Solomon R Guggenheim Museum
1488	26	Whitney Museum of American Art
5088	27	Frick Collection
7155	28	Brooklyn Museum of Art
2366	29	Morgan Library and Museum
6377	30	Neue Galerie New York
12226	31	New York Hall Of Science
1154	32	Children's Museum Of The Arts
4504	33	New York Transit Museum
9559	34	Studio Museum in Harlem
7808	35	Museum of the City of New York
5766	36	Hayden Planetarium
5611	37	Liberty Science Center
3092	38	Discovery Times Square
6880	39	Jewish Museum
4887	40	American Folk Art Museum



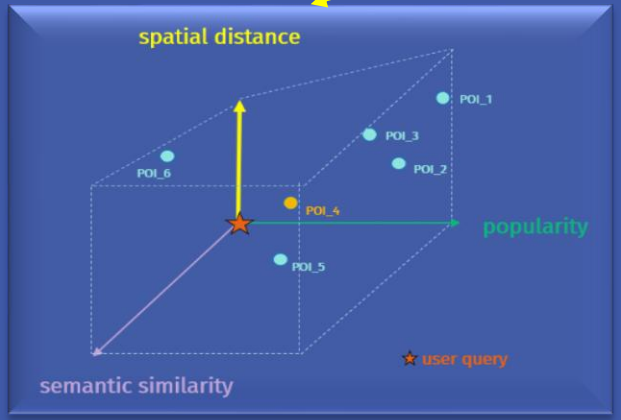
result_index	name_chosen	distance_in_3d
28	Brooklyn Museum of Art	1.944008036
35	Museum of the City of New York	2.708702253
7	Rubin Museum of Art	2.884562181
24	Metropolitan Museum of Art, New York	3.075487991
25	Solomon R Guggenheim Museum	3.215634538
26	Whitney Museum of American Art	3.227113843
39	Jewish Museum	3.331808723
23	Museum of Modern Art	3.479077001
40	American Folk Art Museum	3.537107718
29	Morgan Library and Museum	3.698730279
33	New York Transit Museum	3.813989279
1	New Museum	3.945985885
22	Leslie+Lohman Prince Street Project	4.168553815
34	Studio Museum in Harlem	4.231938417
27	Frick Collection	4.534992227
30	Neue Galerie New York	4.545722956
13	KGB Spy Museum	5.000130568
2	Merchant's House Museum	5.006074114
4	Center for Jewish History	5.256904786
8	Deitch Projects	5.276860818
14	Fusion Arts Museum	5.29431061
3	International Center of Photography Museum	5.420865365
9	Ukrainian Museum	5.448986036
11	Artists Space	5.466585435
37	Liberty Science Center	5.518249528
38	Discovery Times Square	5.613459603
19	TMJ Arts Collective	5.627264684
16	Fitzgerald Fine Arts	5.693510651
18	Museum of Beautiful People	5.704702539
20	Eva Presenhuber	5.786897195
10	Museum of Illusions	5.813412772
31	New York Hall Of Science	5.881987773
21	Judd Foundation	5.890290357
15	American Artist Professional League	5.923079719
6	Theodore Roosevelt Birthplace	5.923867409
5	Afa Gallery	5.979805427
17	Fortnight Institute	6.009715833
32	Children's Museum Of The Arts	6.031164343
36	Hayden Planetarium	6.17404459
12	Margo Feiden Galleries	7.071067812

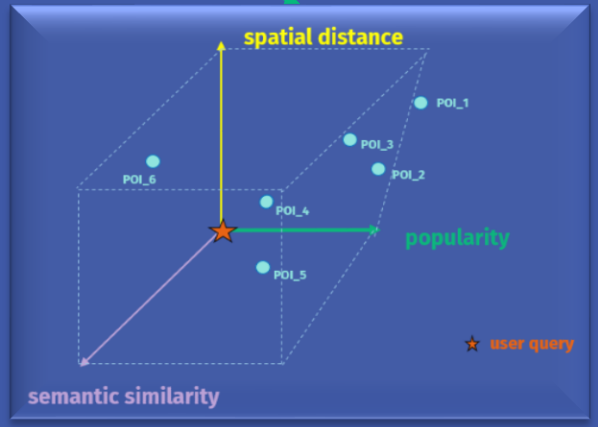
Query “best museum near my location”



shorter distance

more popular





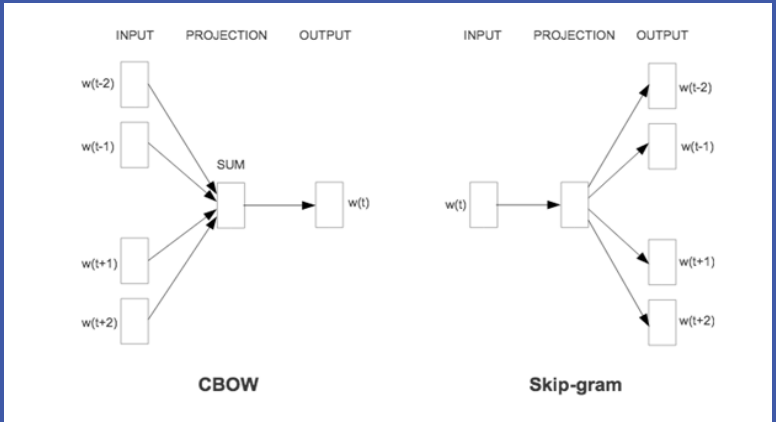
Lessons Learned

- Study your raw data carefully
- Don’t hypothesis based on personal bias
- Define what ‘good’ means and manifest technically

Vielen Dank!

Reference : word2vec

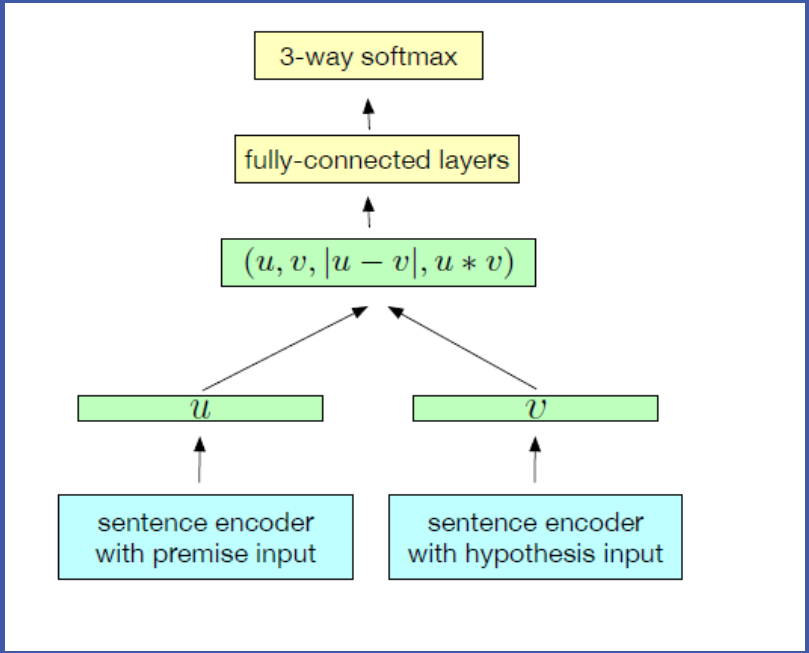
- Neural probabilistic language models train text to maximize the probability of the next word w_t (for "target") given the previous words h (for "history") in terms of a softmax function, where score(w_t,h) computes the compatibility of word w_t with the context h.
- Tensorflow word2vec trains this model by maximizing its log-likelihood on the training set



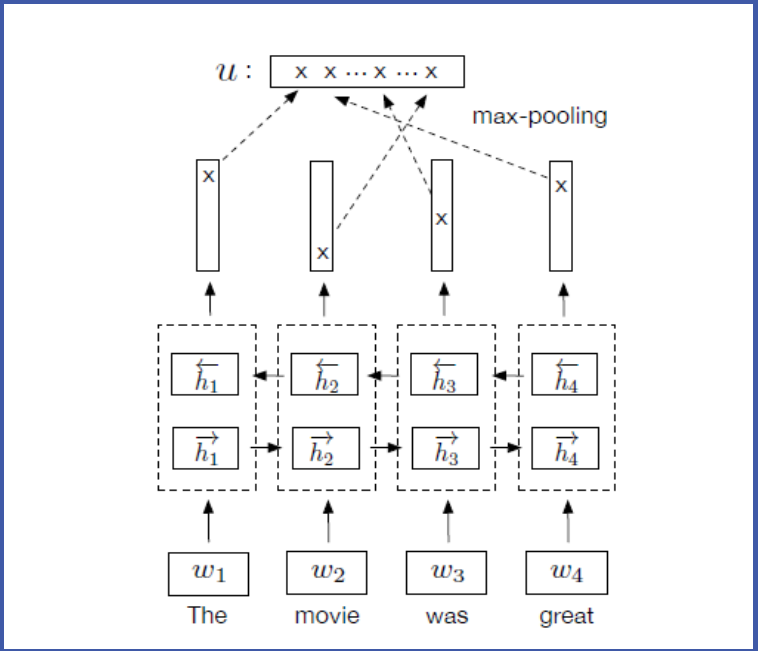
$$P(w_t|h) = \text{softmax}(\text{score}(w_t, h)) = \frac{\exp\{\text{score}(w_t, h)\}}{\sum_{\text{Word w' in Vocab}} \exp\{\text{score}(w', h)\}}$$

$$J_{\text{ML}} = \log P(w_t|h) = \text{score}(w_t, h) - \log\left(\sum_{\text{Word w' in Vocab}} \exp\{\text{score}(w', h)\}\right)$$

Reference : doc2vec



Generic NLI training scheme



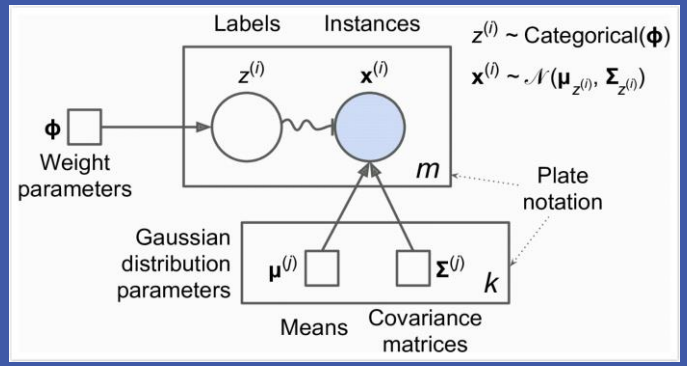
Bi-directional LSTM max-pooling network

Reference : Gaussian Mixture Model

GMM is often categorized as a clustering algorithm, fundamentally it is algorithm for density estimation.

The result of GMM fit to some data is technically not a clustering model, but a generative probabilistic model describing the distribution of the data.

The generative model of the distribution, meaning that the GMM gives us the recipe to generate new random data distributed similarly to our input.

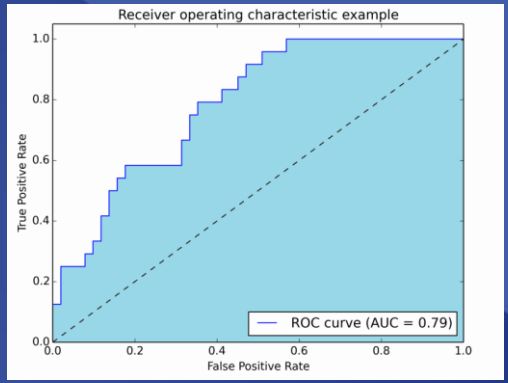


K-means vs GMM

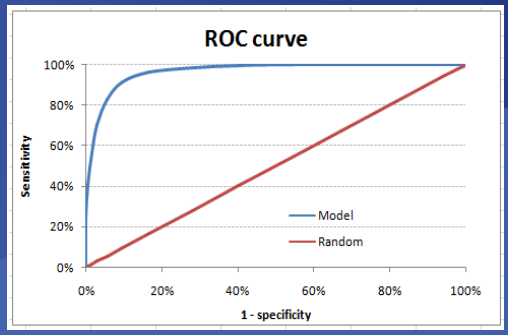
	K-means algorithm	Gaussian Mixture Model
time	Fast running time	Depends on data
dimensionality	Better for high-dimensional data	Difficult to initialization of cluster since GMM use all components access to
Assignment	Hard assignment (good for spherical shaped data), Lack of probabilistic cluster assignment, its lack of flexibility in cluster shape	Soft assignment (Good for complex geometric shaped data), (data point belong to cluster A with probability 20% and to cluster B with probability 80%)
Interpretation	Easy to interpret	Difficult to interpret
Minimization method	Find k to minimize $(x-\mu_k)^2$	Find k to minimize $\frac{(x-\mu_k)^2}{\sigma^2}$ take the variance σ^2 into consideration

Reference : Confusion matrix and Area Under the ROC curve (AUC)

- **Accuracy** : the proportion of the total number of predictions that were correct.
- **Positive Predictive Value (Precision)** : proportion of positive cases that were correctly identified.
- **Negative Predictive Value** : proportion of negative cases that were correctly identified.
- **Sensitivity (Recall)** : proportion of actual positive cases which are correctly identified.
- **Specificity** : proportion of actual negative cases which are correctly identified.



Confusion Matrix		Target			
		Positive	Negative		
Model	Positive	a	b	*Positive Predictive Value*	a/(a+b)
	Negative	c	d	*Negative Predictive Value*	d/(c+d)
		Sensitivity	*Specificity*	**Accuracy** = (a+d)/(a+b+c+d)	
		a/(a+c)	d/(b+d)		



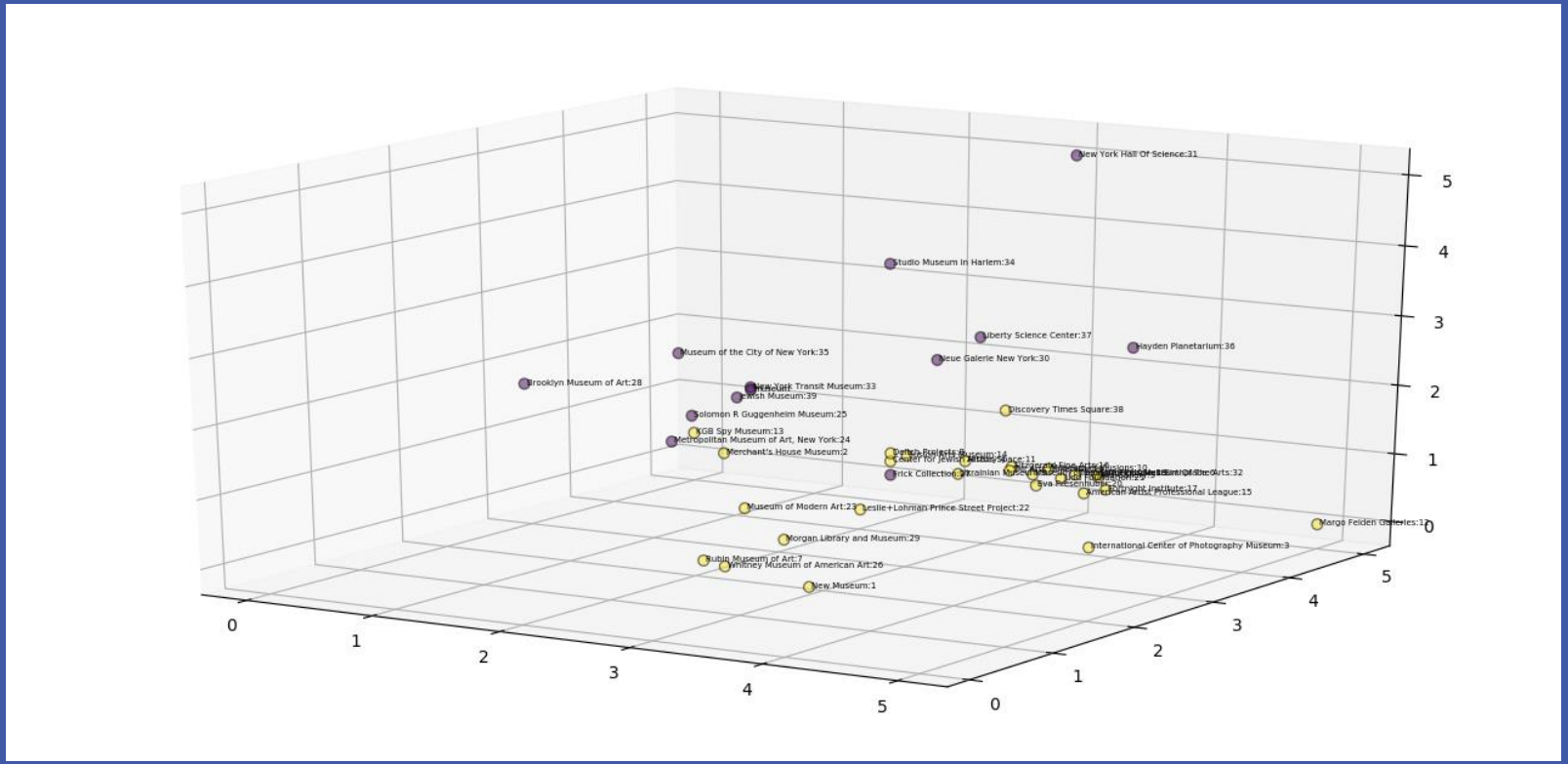
https://www.analyticsvidhya.com/blog/2016/02/7-important-model-evaluation-error-metrics/

Clustering

* Why?
  - Make two clusters and check which cluster contains query point
  - Check the characteristics of each cluster

* clustering algorithm
  - K-means algorithm : hard cluster assignment
  - Gaussian Mixture model : probabilistic (soft) cluster assignment

Clustering by k-means



Clustering by Gaussian Mixture Model