

Leveraging Latent Factors for Predicting User Preferences from Restaurant Reviews



Suman Adhya

Roll: 2018SMCS008

Registration Number: 201802050102032

School of Mathematical and Computational Sciences

Indian Association for the Cultivation of Science

Supervised by: Dr. Partha Basuchowdhuri

Contents:

- ❖ Introduction
- ❖ Data crawling
- ❖ Data summarization and visualization
- ❖ SVD
- ❖ Matrix factorization
- ❖ Generating the user's Network
- ❖ Community detection
- ❖ Community analysis by LDA
- ❖ Further Improvements
- ❖ References

Introduction

- ❖ Motivation
- ❖ Problem statement

Motivation

Hitaishi Majumder
86 Reviews , 109 Followers
Follow

18 days ago

RATED 5.0 There's nothing new to be said about the excellent service or the amazing food at Chili's. It's one of my favourite dining places in the city for a reason. They have launched a new menu called Flavours of Chili's 2.0. It's a limited menu but as usual, all the items taste amazing. Do try the roasted garlic spaghetti with chili oil. The spaghetti has a smoked flavour that gives the dish a unique touch.

Keshav Kedia
1 Review , 1 Follower
Follow

16 days ago

RATED 1.0 The service sucks
The food quality has been going down
And asking for small changes to the order like topping replacement is a crime
They are unwilling to do anything unfortunately
The quality of service at the other outlets in South City and acropolis is much better

Like 0 Comment 0 Share

<https://www.zomato.com/kolkata/chilis-grill-bar-ballygunge/reviews>

Problem Statement

We aim to divide user into groups and try to understand what are the hidden factors(latent factors) that makes an individual of a group similar to others members of the same group and dissimilar from others of different groups.

In other words, we tried to understand the latent factors for each group.



Data Collection

- ❖ Why choose Zomato?
- ❖ Data scraping
- ❖ Problem faced

Why choose Zomato?



Popular localities in and around Kolkata

Explore restaurants, bars, and cafés by locality

Park Street Area (204 places)	Sector 5, Salt Lake (303 places)	Sector 1, Salt Lake (246 places)
Ballygunge (249 places)	Southern Avenue (112 places)	New Town (617 places)
Camac Street Area (76 places)	Chinar Park (211 places)	Elgin (113 places)
Kasba (236 places)	Science City Area (39 places)	Kankurgachi (175 places)
New Market Area (128 places)	Park Circus Area (200 places)	Desapriya Park (58 places)
Golpark (50 places)	Bhawanipur (141 places)	Sector 3, Salt Lake (120 places)
Hindustan Park (37 places)	Prince Anwar Shah Road (135 places)	Gariahat (111 places)
Behala (378 places)	Esplanade (64 places)	Theatre Road (59 places)
Jadavpur (223 places)	Hatibagan (98 places)	Tollygunge (354 places)
Nagerbazar (96 places)	Lake Market Area (71 places)	Tangra (64 places)

Rich data set

What is data scraping?

Data scraping, also known as **web scraping**, is the process of importing information from a website into a spreadsheet or local file saved on your computer. It's one of the most efficient ways to get **data** from the **web**.



How to scrape data?

- A. If whole HTML(hyper text markup language) data is available in single web page i.e. if we don't need automated click we can use these python packages:
 - a. ***Requests***
 - b. ***html5lib(HTML parser)***
 - c. ***bs4(BeautifulSoup)***
 - d. ***re(regular expression)***
- B. If automated clicks are needed there is no way other than to make a crawler(bot).

We build our bot to crawl data from Zomato using ***selenium*** package.

Problem faced

- ❖ IP may get blocked
- ❖ Some time they force you to log in for verification
- ❖ Slow/unstable load speed

Data Summarization

Number of Users	185346
Number of Restaurants	6488
Number of Reviews	587709

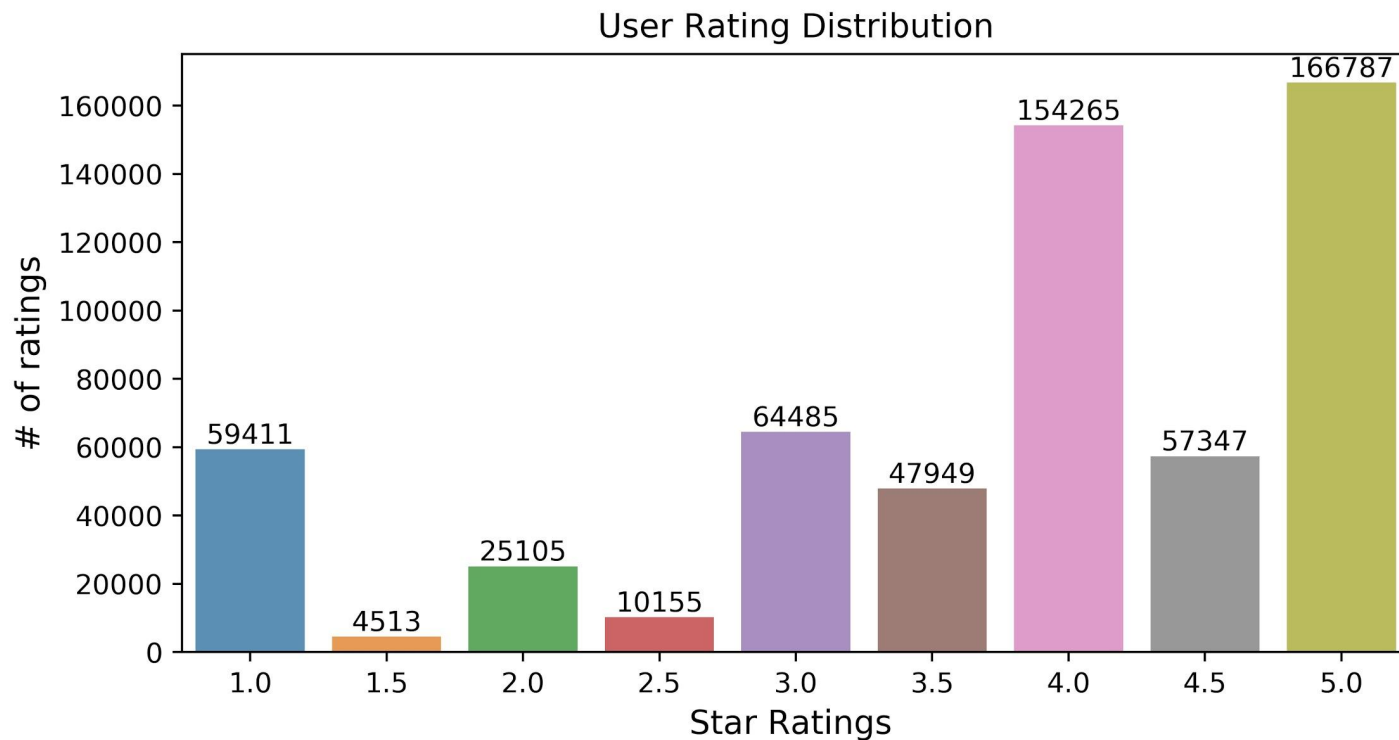
Restaurant Information:

	Link	Name	Restaurant ID	Rating	No. of Votes	Address	Cuisine	Cost for Two
0	https://www.zomato.com/kolkata/leaky-cauldron-...	Leaky Cauldron Kolkata	19089531	NEW	0	Plot 121, Baroda Avenue, Near Indian Oil Petro...	North Indian, Chinese, Continental	₹1,000
1	https://www.zomato.com/kolkata/shawarma-nation...	Shawarma Nation Kolkata	18950586	3.3	12	P 20, Nabalia Para Road, Behala, Kolkata	Lebanese, Middle Eastern, Healthy Food	₹350
2	https://www.zomato.com/kolkata/food-junction-2...	Food Junction Kolkata	19110775	NEW	0	4th Floor, 32 Kabi Guru Rabindra Path (N), Kha...	North Indian, Chinese, Mughlai	₹500
3	https://www.zomato.com/kolkata/the-shack-secto...	The Shack Kolkata	25386	3.5	89	Plot A3, EP & GP Block, Bytes Food Court, Infi...	Fast Food	₹450
4	https://www.zomato.com/kolkata/8th-day-cafe-ba...	8th Day Cafe & Bakery Kolkata	24874	4.1	972	Arcadia, 6, West Range, Mullick Bazar, Park Ci...	Cafe, Desserts	₹600

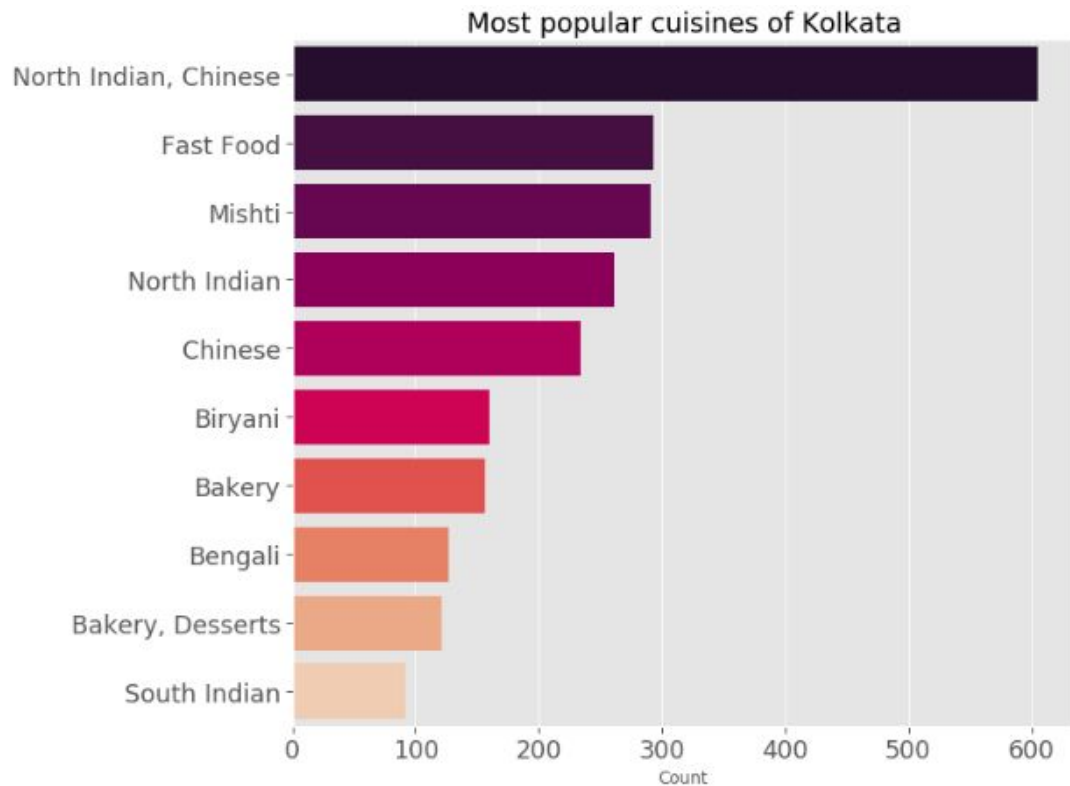
User reviews:

	Restaurant Name	Restaurant Id	User Name	User Id	User Link	User Rating	Review Time	Review Text
0	Leaky Cauldron Kolkata	19089531.0	Susmita Mondal	118650082	https://www.zomato.com/users/susmita-mondal-11...	5.0	2019-07-06 17:46:05	It's a very nice place with decor.. The people...
1	Leaky Cauldron Kolkata	19089531.0	Sayon Mukherjee	29293643	https://www.zomato.com/users/sayon-mukherjee-2...	4.0	2019-07-06 11:36:58	Really wanted to try this place and see how th...
2	Leaky Cauldron Kolkata	19089531.0	Soham Bhaduri	657702	https://www.zomato.com/users/soham-bhaduri-657702	5.0	2019-07-05 12:14:48	This place is an absolute must visit if for a ...
3	Leaky Cauldron Kolkata	19089531.0	Teena Panda	76785732	https://www.zomato.com/users/teena-panda-76785732	4.0	2019-06-30 07:42:53	The lamb was outstanding but the fire whiskey ...
4	Leaky Cauldron Kolkata	19089531.0	Upasana Ghosh	46441933	https://www.zomato.com/users/upasana-ghosh-464...	5.0	2019-06-25 21:33:27	Honestly speaking, I hadn't expected the place...

User Rating Distribution:

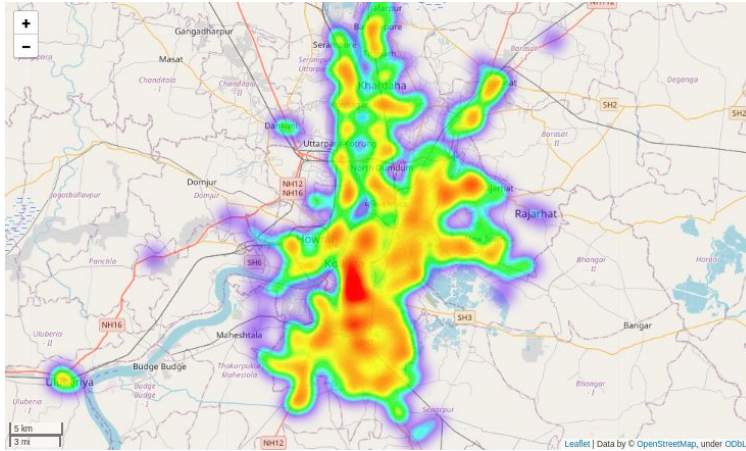


Popular Cuisines:

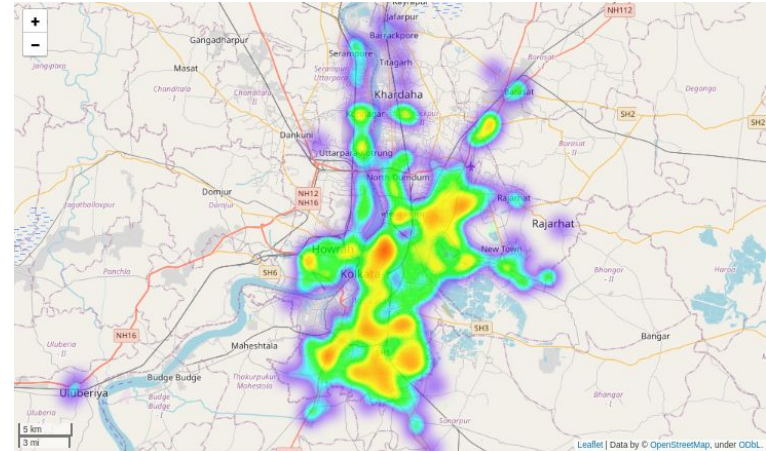


Heatmaps:

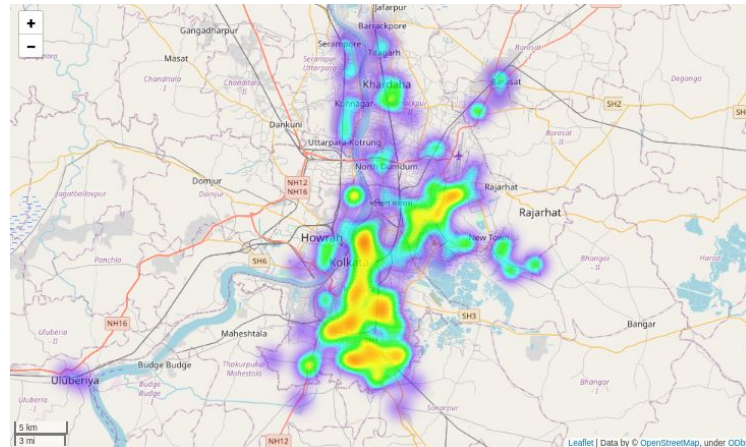
(A)



(B)



(C)



1. Heatmap of all restaurants in Kolkata
2. Heatmap of all North Indian & Chinese restaurants in Kolkata
3. Heatmap of all Fast food counters in Kolkata

SVD - Definition

$$A_{m \times n} = U_{m \times k} \Sigma_{k \times k} V_{n \times k}^T$$

A: input data matrix (m x n)

U: left singular vectors (m x k)

Σ: singular values (k x k diagonal matrix (strength of each 'concept')) k- rank of the matrix

V: right singular vector (n x k)

Relation to Eigen value decomposition

$$S = X \Lambda X^T$$

Where:

- S is a symmetric matrix
- X is eigen vectors of S , which is orthonormal for symmetric matrices.
- Λ is diagonal matrix of eigen values

CONCLUSION,

Let $A_{m \times n}$ be any matrix, then from SVD, $A = U \Sigma V^T$

$$AA^T = (U \Sigma V^T)(U \Sigma V^T)^T = U \Sigma \Sigma^T U^T \quad \leftarrow X \Lambda X^T$$

$$A^T A = (U \Sigma V^T)^T (U \Sigma V^T) = V \Sigma \Sigma^T V^T \quad \leftarrow Y \Lambda Y^T$$

Shows how to
compute SVD
from eigenvalue
decomposition

SVD - properties

It is always possible to decompose a matrix A into $A = U\Sigma V^T$

- U, Σ, V are unique
- $U_{m \times k}$ is a column orthonormal matrix; i.e. $U^T U = I_k$
- $\Sigma_{k \times k}$ is diagonal matrix, s.t. for each $\sigma_i \in \Sigma$, $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_k \geq 0$
- $V_{n \times k}$ is a column orthogonal matrix; i.e. $V^T V = I_k$

SVD-Example on user to movie rating matrix

	Matrix	Alien	Star Wars	Casablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	2	0	4	4
Jenny	0	0	0	5	5
Jane	0	1	0	2	2

Let M be a rating matrix where each row represent a user's rating for movies. By SVD we can divide M into three factors U, Σ, V^T where U is user to concept matrix, Σ is concept to concept matrix and V is movie to concept matrix.

SVD - Example on user to movie rating matrix

$$\begin{bmatrix} .13 & .02 & -.01 \\ .41 & .07 & -.03 \\ .55 & .09 & -.04 \\ .68 & .11 & -.05 \\ .15 & -.59 & .65 \\ .07 & -.73 & -.67 \\ .07 & -.29 & .32 \end{bmatrix} \begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & 1.3 \end{bmatrix} \begin{bmatrix} .56 & .59 & .56 & .09 & .09 \\ .12 & -.02 & .12 & -.69 & -.69 \\ .40 & -.80 & .40 & .09 & .09 \end{bmatrix}$$

$U \qquad \qquad \qquad \Sigma \qquad \qquad \qquad V^T$

SVD - Dimension reduction

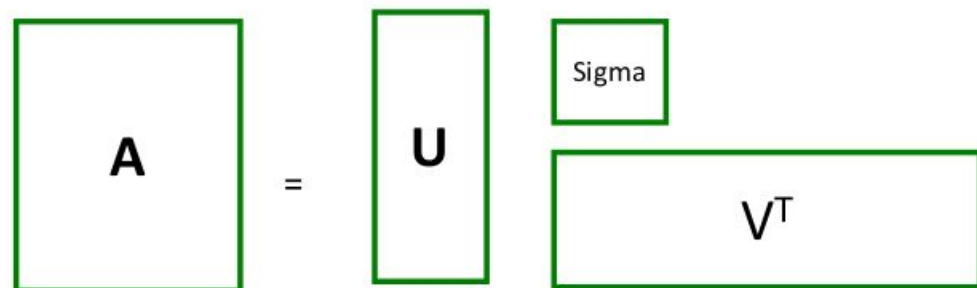
How to do dimension reduction?

⇒ Set smallest singular values to zero.

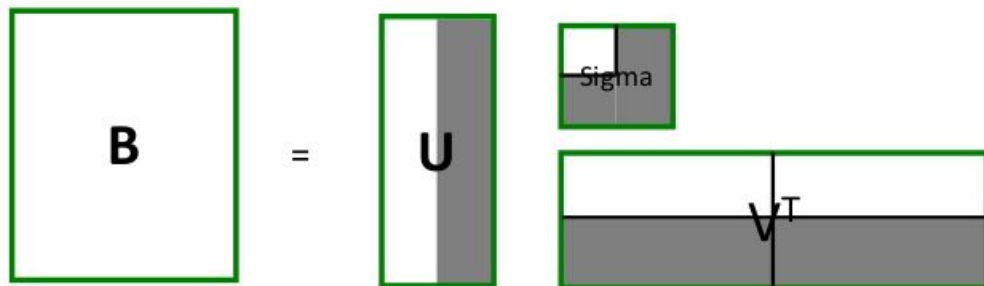
$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix} \approx \begin{bmatrix} 0.13 & 0.02 & -0.01 \\ 0.41 & 0.07 & -0.03 \\ 0.55 & 0.09 & -0.04 \\ 0.68 & 0.11 & -0.05 \\ 0.15 & -0.59 & 0.65 \\ 0.07 & -0.73 & -0.67 \\ 0.07 & -0.29 & 0.32 \end{bmatrix} \times \begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & 1.3 \end{bmatrix} \times \begin{bmatrix} 0.56 & 0.59 & 0.56 & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & -0.69 & -0.69 \\ 0.40 & -0.80 & 0.40 & 0.09 & 0.09 \end{bmatrix}$$

35

SVD - How many singular values can be set to zero ?



B is best approximation of A



**keep the few
largest singular
values to
preserve
(80-90% of
'energy')**

SVD - Dimension reduction

How to do dimension reduction?

⇒ Set smallest singular values to zero.

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix} \approx \begin{bmatrix} 0.92 & 0.95 & 0.92 & 0.01 & 0.01 \\ 2.91 & 3.01 & 2.91 & -0.01 & -0.01 \\ 3.90 & 4.04 & 3.90 & 0.01 & 0.01 \\ 4.82 & 5.00 & 4.82 & 0.03 & 0.03 \\ 0.70 & 0.53 & 0.70 & 4.11 & 4.11 \\ -0.69 & 1.34 & -0.69 & 4.78 & 4.78 \\ 0.32 & 0.23 & 0.32 & 2.01 & 2.01 \end{bmatrix} = \mathbf{A}_k$$

Eckart–Young–Mirsky theorem (for spectral norm) says that, if $\text{rank}(\mathbf{B})=k$ then, $\|\mathbf{A}-\mathbf{B}\| \geq \|\mathbf{A}-\mathbf{A}_k\|$

I.e. \mathbf{A}_k is the best approximation of \mathbf{A}

Problem with SVD:

SVD deals with missing entries as zero

Matrix factorization model

Built a recommendation that works well on the known ratings.
And hope system will also predict the unknown ratings as well.

It is an optimization problem; the optimization is to give the best rating prediction.

Goal: The lower RMSE value \Rightarrow better rating prediction

Matrix Factorization:

1		3			5			5		4	
		5	4			4			2	1	3
2	4		1	2		3		4	3	5	
	2	4		5			4			2	
	4	3	4	2						2	5
1		3		3			2			4	

R

Here R is a user to rating matrix i.e. $r_{ij} \Rightarrow$ rating of user i for item j.

R can be factorize as **$R = QP^T$** .

From SVD, **$R = U\Sigma V^T = (U)(\Sigma V^T)$**

Matrix Factorization:

1		3			5			5		4	
		5	4			4			2	1	3
2	4		1	2		3		4	3	5	
	2	4		5			4			2	
	4	3	4	2						2	5
1		3		3			2			4	

R

\approx

Q

0.1	-.4	0.2
-.5	0.6	0.5
0.2	0.3	0.5
1.1	2.1	.3
-.7	2.1	-2
.1	.7	.3

X

1.1	-.2	.3	.5	-2	-.5	.8	-.4	.3	1.4	2.4	0.9
-.8	.7	.5	1.4	.3	-1	1.4	2.9	-.7	1.2	-.1	2.3
2.1	-.4	.6	1.7	2.4	.9	-.3	.4	.8	.7	-.6	-.3

P^T

Matrix Factorization:

1		3			5			5		4	
		5	4	?		4			2	1	3
2	4		1	2		3		4	3	5	
	2	4		5			4			2	
	4	3	4	2						2	5
1		3		3			2			4	

≈

Q

0.1	-.4	0.2
-.5	0.6	0.5
0.2	0.3	0.5
1.1	2.1	.3
-.7	2.1	-2
.1	.7	.3

X

$$\hat{r}_{xi} = q_i \cdot p_x$$

$$= \sum_f q_{if} \cdot p_{xf}$$

q_i = row i of Q
 p_x = column x of P^T

R

1.1	-.2	.3	.5	-2	-.5	.8	-.4	.3	1.4	2.4	0.9
-.8	.7	.5	1.4	.3	-1	1.4	2.9	-.7	1.2	-.1	2.3
2.1	-.4	.6	1.7	2.4	.9	-.3	.4	.8	.7	-.6	-.3

P^T

Matrix Factorization:

1		3			5			5		4	
		5	4	2.4		4			2	1	3
2	4		1	2		3		4	3	5	
	2	4		5			4			2	
	4	3	4	2						2	5
1		3		3			2			4	

≈

Q

0.1	-.4	0.2
-.5	0.6	0.5
0.2	0.3	0.5
1.1	2.1	.3
-.7	2.1	-2
.1	.7	.3

x

$$\hat{r}_{xi} = q_i \cdot p_x$$

$$= \sum_f q_{if} \cdot p_{xf}$$

q_i = row i of Q
 p_x = column x of P^T


R

1.1	-.2	.3	.5	-2	-.5	.8	-.4	.3	1.4	2.4	0.9
-.8	.7	.5	1.4	.3	-1	1.4	2.9	-.7	1.2	-.1	2.3
2.1	-.4	.6	1.7	2.4	.9	-.3	.4	.8	.7	-.6	-.3

P^T

Matrix Factorization

Our task is to find P and Q s.t.-

$$\min_{P,Q} \sum_{(i,x) \in R} (r_{xi} - q_i \cdot p_x)^2$$


Only over the known ratings

But there is still an issue of overfitting.

Matrix Factorization

Regularization: Find P, Q s.t. :

$$\min_{P, Q} \underbrace{\sum_{training} (r_{xi} - q_i p_x)^2}_{\text{"error"}} + \underbrace{\left[\lambda_1 \sum_x \|p_x\|^2 + \lambda_2 \sum_i \|q_i\|^2 \right]}_{\text{"length"}}$$

$\lambda_1, \lambda_2 \dots$ user set regularization parameters

Matrix Factorization

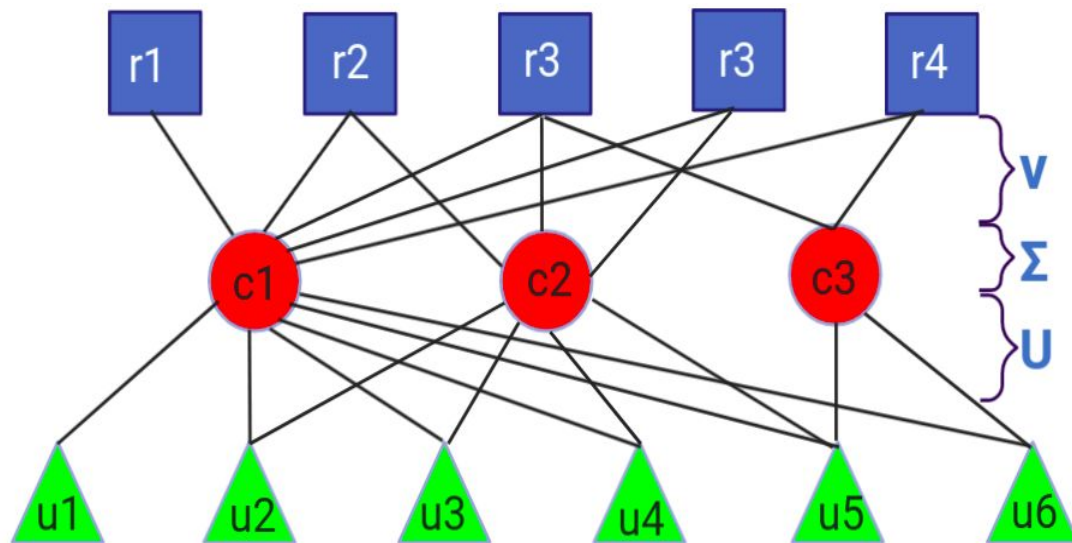
Applying this model on our data we get RMSE value = 0.8217

i.e. $\text{actual rating} \in (\text{predicted rating} - 0.8217, \text{predicted rating} + 0.8217)$

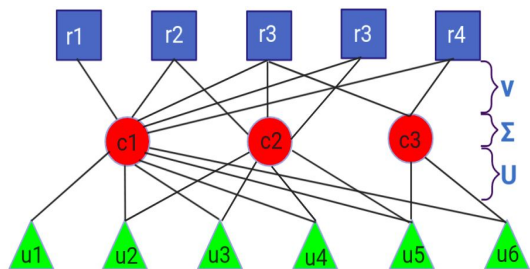
SVD - Interpretation

We can interpret output of SVD as a tripartite graph:

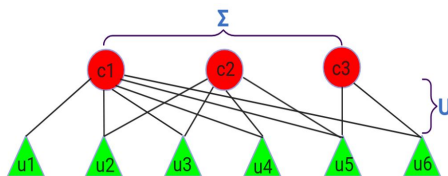
$$\mathbf{A}_T^{m \times n} = \mathbf{U}^{m \times k} \mathbf{\Sigma}^{k \times k} (\mathbf{V}^{n \times k})$$



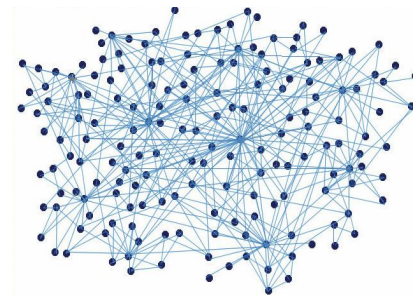
Workflow



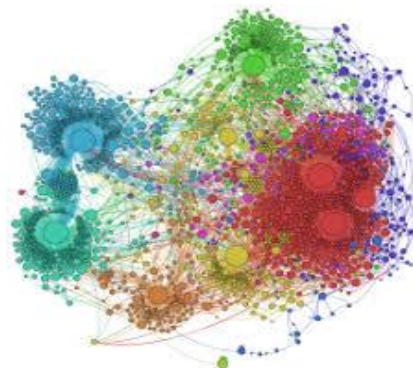
Output of SVD



Users to factors matrix



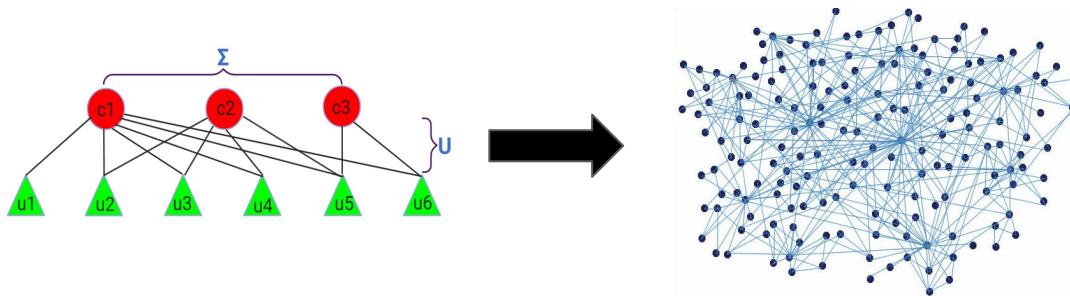
Users network



User's communities

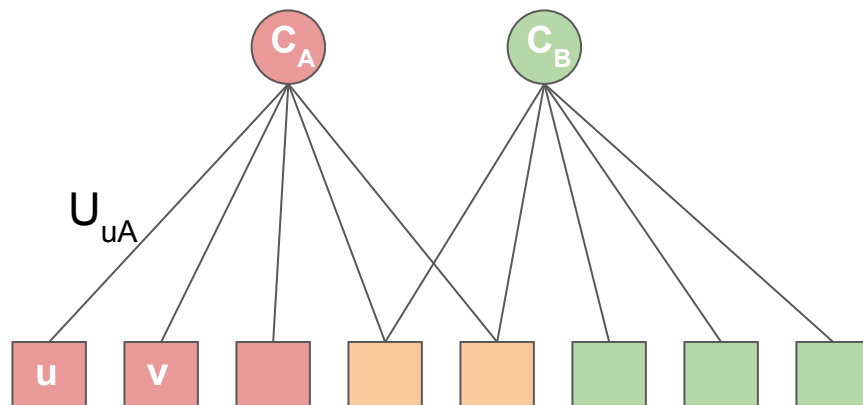
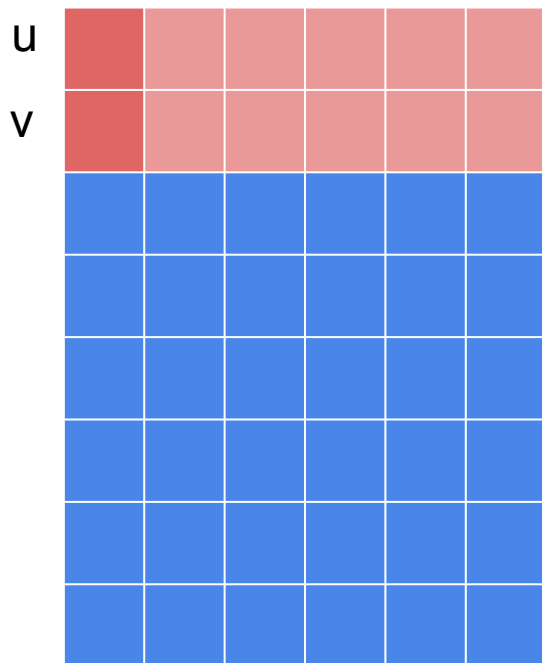


Goal: Define a model that can generate networks



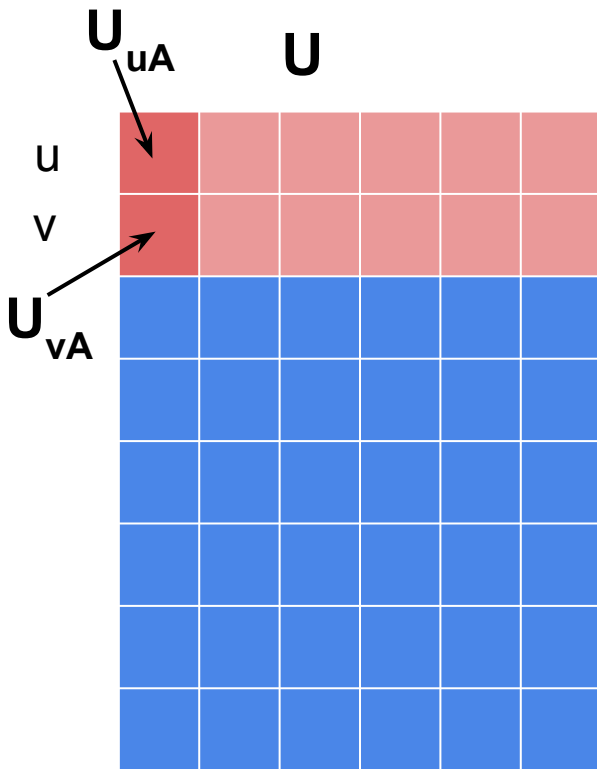
Q: Given a set of nodes, how do latent factors “generate” edges of the network?

Interpretation of matrix U



U_{uA} : user to community A weightage

Defining edge probability of our network



Edge probability for a given community A:

$$p_A(u, v) = 1 - \exp(-U_{uA}U_{vA})$$

- The edge probability is proportional to the product of weights for a given community.

NOTE: It is clear that if one of the node does not belong to the community **A** i.e. say if, $U_{uA} = 0$ or $U_{vA} = 0$
Then, $p_A(u, v) = 1 - \exp(0) = 1 - 1 = 0$

Defining edge probability of our network

Now for single community we have:

$$p_A(u, v) = 1 - \exp(-U_{uA}U_{vA})$$

Therefore for all such communities we have,

$$\begin{aligned} p(u, v) &= 1 - \exp\left(-\sum_{A \in C} U_{uA} \cdot U_{vA}\right) \\ &= 1 - \exp(-U_u \cdot U_v^T) \end{aligned}$$

Explanation of the edge probability formula:

Let, for a pair of nodes $u, v \in V$ in $G(V, E)$; this pair of nodes connected if $X_{uv} > 0$

Now consider that, nodes u, v generate an interaction of strength $X_{uv}^{(c)}$ within each community c using a **Poisson distribution** with mean $U_{uc} \cdot U_{vc}$

therefore, $X_{uv}^{(c)} \sim \text{Pois}(U_{uc} \cdot U_{vc})$

Now, $X_{uv} = \sum_c X_{uv}^c$

$X_{uv} \sim \text{Pois}(U_u \cdot U_v)$

The edge probability then, $P(X_{uv} > 0) = 1 - P(X_{uv} = 0) = 1 - \exp(-U_u \cdot U_v)$



$$P(X = 0) = \frac{\lambda^0}{0!} \exp(-\lambda) = \exp(-\lambda)$$

Example

U_u	3.1	0	2.4	0
-------	-----	---	-----	---

U_v	0	2.6	4.1	5.2
-------	---	-----	-----	-----

U_w	0	1.6	0	2.7
-------	---	-----	---	-----

Now,

➤ $U_u \cdot U_v = (2.4 \times 4.1) = 9.84$

➤ $U_v \cdot U_w = (2.6 \times 1.6) + (5.2 \times 2.7) = 18.2$

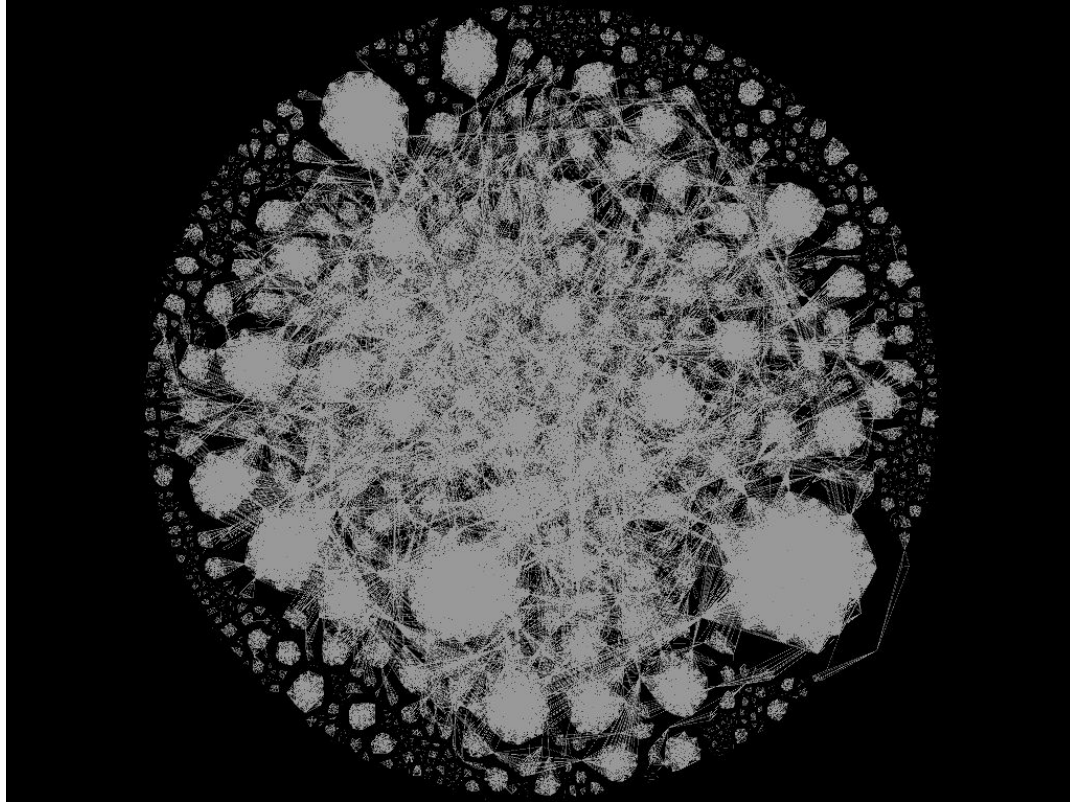
➤ $U_u \cdot U_w = 0$

Therefore,

- $p(u,v) = 1 - \exp(-9.84) = 0.999$
- $p(v,w) = 1 - \exp(-18.2) = 0.999$
- $p(u,w) = 1 - \exp(0) = 0$

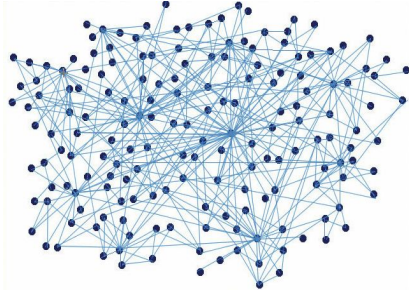
User's Network

$G(V,E)$:

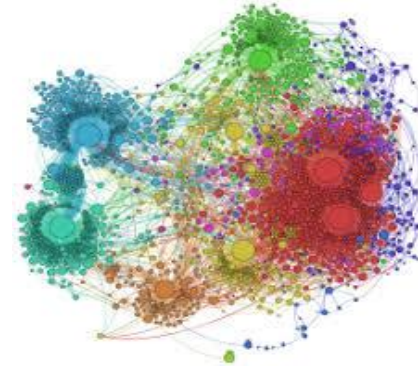


Identify communities:

How to identify the communities?



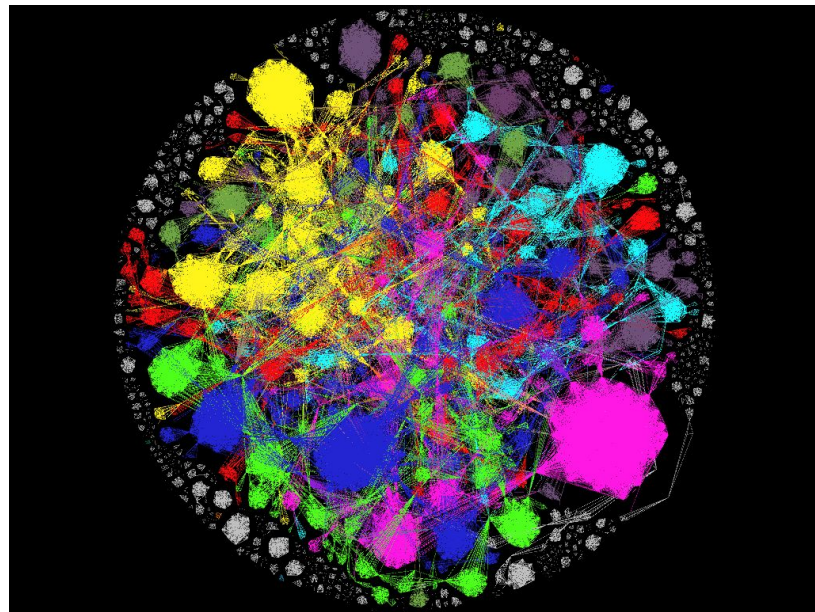
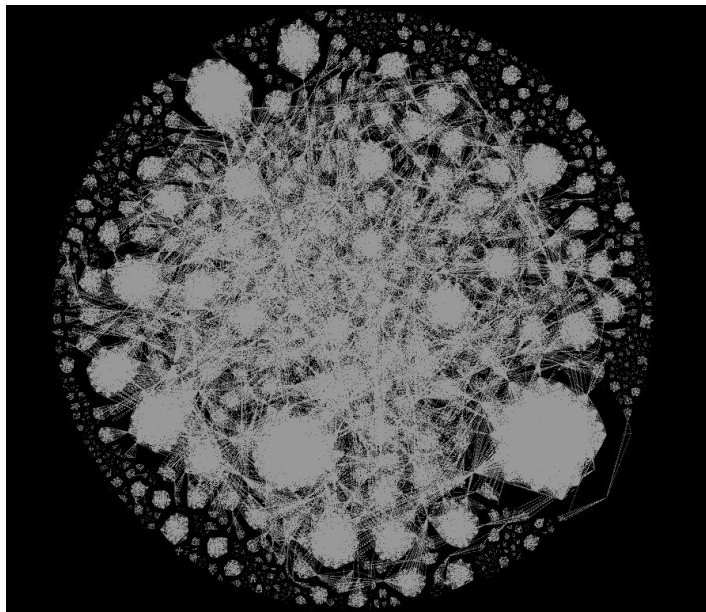
Users network



User's communities

We use Gephi's modularity detection algorithm to detect the communities

Output: we get 1326 number of communities



Analysis of the communities

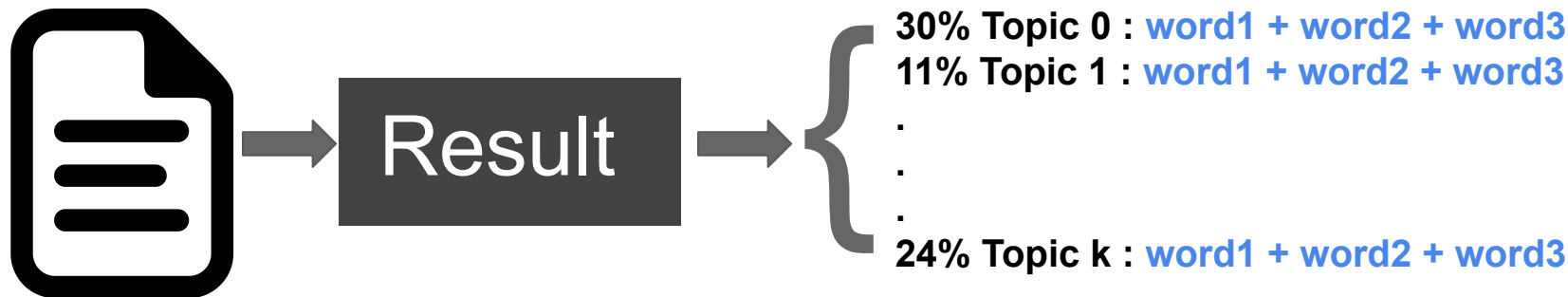
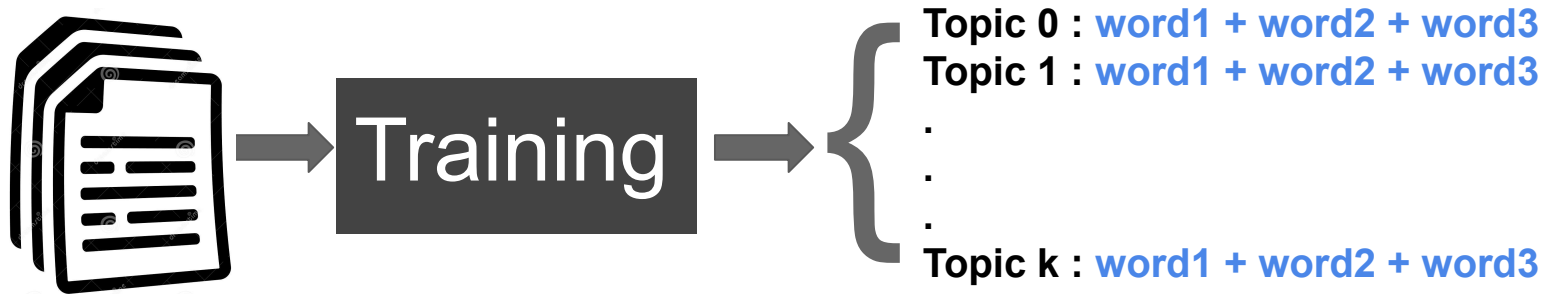
Now to analyse these user's communities we look at their text reviews.

Topic modeling on text review data

We use LDA for topic modeling.

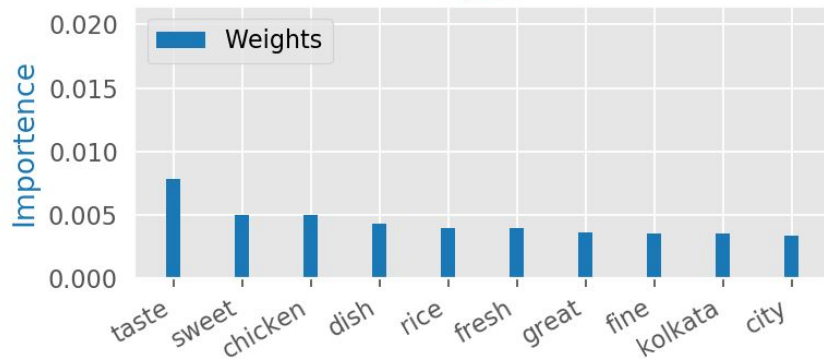
Idea- *Each document can describe as distribution of topics, and each topic can be described as distribution of words.*

How it works:

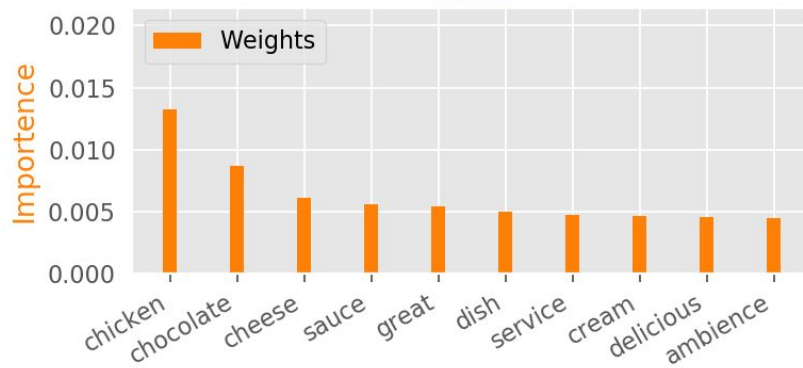


We learn our LDA model on our review corpus for 4 topics:

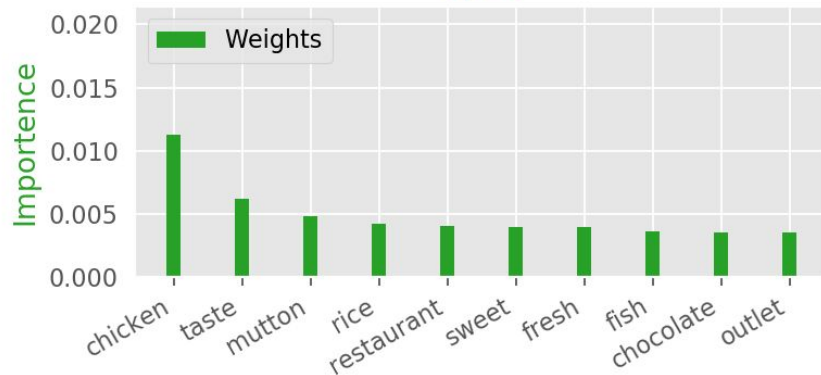
Topic: 0



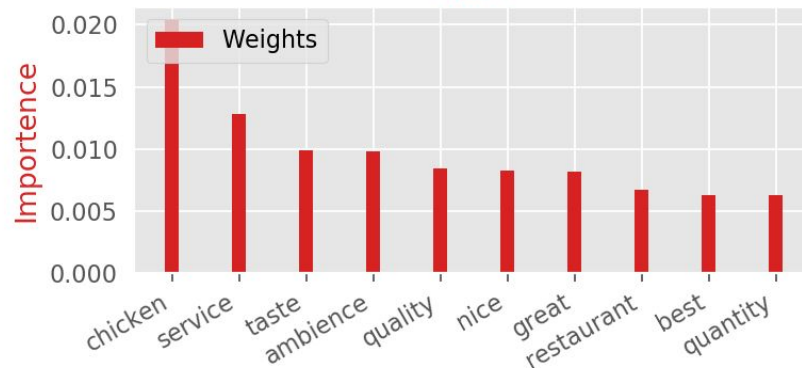
Topic: 1



Topic: 2



Topic: 3



Further Improvement:

- In similar way we can generate restaurant communities, and analyse those communities.
- This extracted latent factors can be used for further modeling, where we can learn those factors for better prediction.

References:

1. V. CHESNOKOV, *Overlapping community detection in social networks with node attributes by neighborhood influence*, 06 2017, pp. 187–203.
2. Y. KOREN, R. BELL, AND C. VOLINSKY, *Matrix factorization techniques for recommender systems*, Computer, 42 (2009), pp. 30–37.
3. J. YANG AND J. LESKOVEC, *Overlapping community detection at scale: A nonnegative matrix factorization approach*, in Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, WSDM '13, New York, NY, USA, 2013, Association for Computing Machinery, p. 587–596.
4. J. LESKOVEC, *Recommender systems: Latent factor model*, 2015.

Acknowledgements:

- Dr. Partha Basuchowdhuri
- CCRES
- Gautam Khanna(from Google)
- Dr. Sanjay Kumar Saha(prof. in Jadavpur University)
- Lab members

Thank You