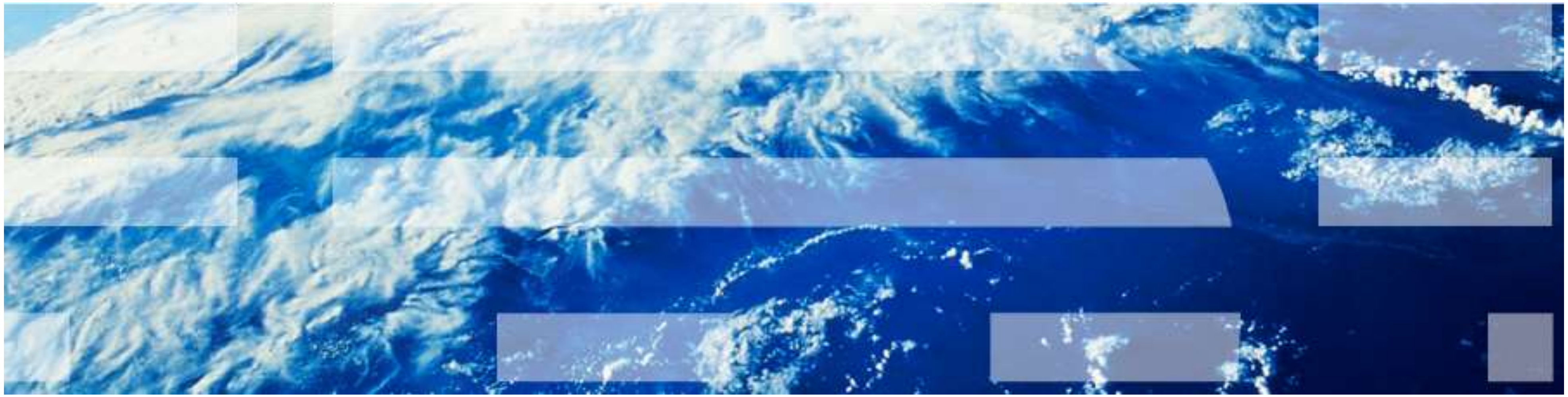
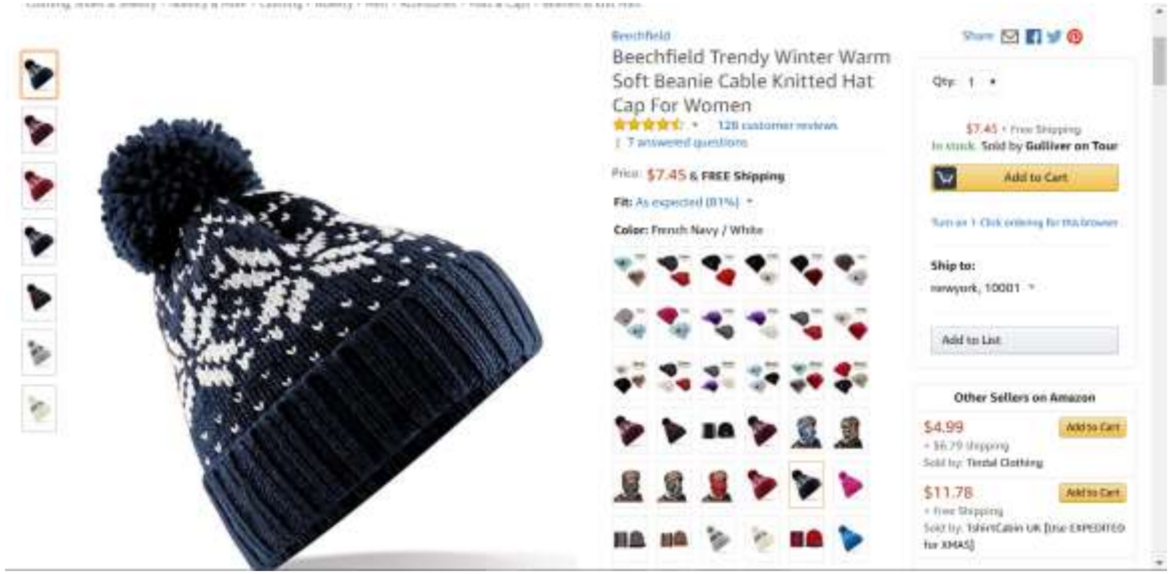
E6893 Big Data Analytics:

Amazon Recommendation

Team Members: Minhui Li (ml4026), Qiaofeng Wu (qw2235)



Motivation



Beechfield
Beechfield Trendy Winter Warm Soft Beanie Cable Knitted Hat Cap For Women
★★★★☆ 128 customer reviews | 7 answered questions

Price: $7.45 & FREE Shipping

Fit: As expected (81%)

Color: French Navy / White

Share

Qty: 1

$7.45 + Free Shipping
In stock. Sold by Gulliver on Tour

Add to Cart

Turn on 1-Click ordering for this browser

Ship to:
newyork, 10001

Add to List

Other Sellers on Amazon

$4.99 + $6.79 shipping
Sold by: Tendal Clothing
Add to Cart

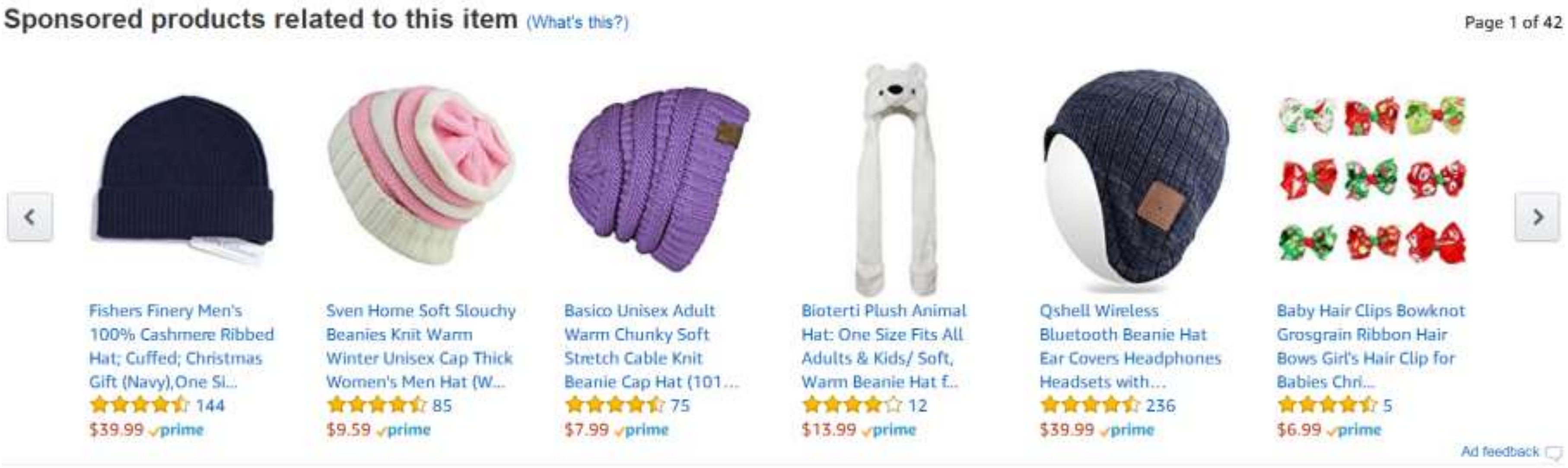
$11.78 + Free Shipping
Sold by: TshirtCabin UK [Use EXPEDITED for XMAS]
Add to Cart

Benefit customers, convenient, quickly locate;

Increase profits for sellers, a trick may temp customers to buy something that not been added to shopping list;

But current recommendation sometimes fails and the list shows unexpected things

NEW Algorithm



Sponsored products related to this item (What's this?)

Page 1 of 42

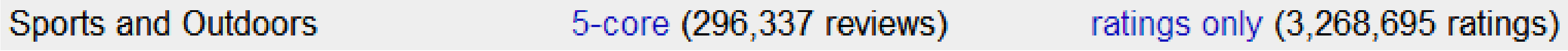
Fishers Finery Men's 100% Cashmere Ribbed Hat; Cuffed; Christmas Gift (Navy),One Si...	Sven Home Soft Slouchy Beanies Knit Warm Winter Unisex Cap Thick Women's Men Hat (W...	Basico Unisex Adult Warm Chunky Soft Stretch Cable Knit Beanie Cap Hat (101...	Bioterti Plush Animal Hat: One Size Fits All Adults & Kids/ Soft, Warm Beanie Hat f...	Qshell Wireless Bluetooth Beanie Hat Ear Covers Headphones Headsets with...	Baby Hair Clips Bowknot Grosgrain Ribbon Hair Bows Girl's Hair Clip for Babies Chri...
★★★★½ 144	★★★★½ 85	★★★★½ 75	★★★★☆ 12	★★★★½ 236	★★★★½ 5
$39.99 ✓prime	$9.59 ✓prime	$7.99 ✓prime	$13.99 ✓prime	$39.99 ✓prime	$6.99 ✓prime

Ad feedback

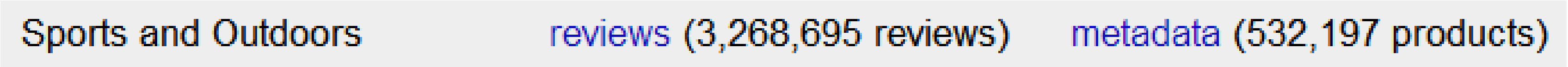
Dataset and Tools

Raw Dataset: “Sports and Outdoors”

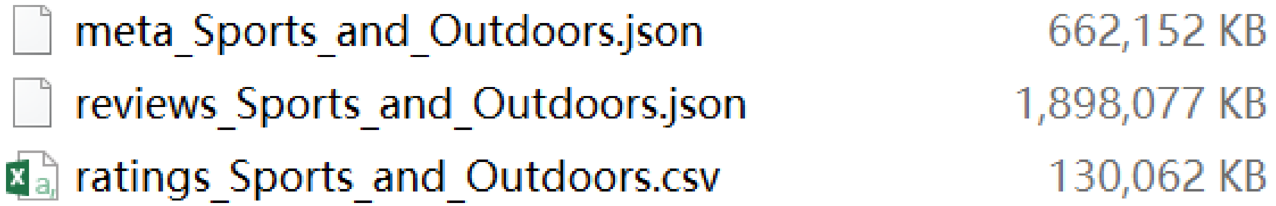
Obtained from: http://jmcauley.ucsd.edu/data/amazon/



Sports and Outdoors	5-core (296,337 reviews)	ratings only (3,268,695 ratings)



Sports and Outdoors	reviews (3,268,695 reviews)	metadata (532,197 products)



meta_Sports_and_Outdoors.json	662,152 KB
reviews_Sports_and_Outdoors.json	1,898,077 KB
ratings_Sports_and_Outdoors.csv	130,062 KB

Algorithm

prediction of unknown pairs

average known ratings for user a

coefficient

similarity matrix of uers

known ratings

average known ratings for user b

$$P_{a,\mathrm{i}} = \overline{r}_a + \alpha \sum_{b=1}^{n} su(a,b)(r_{b,i} - \overline{r}_b)$$

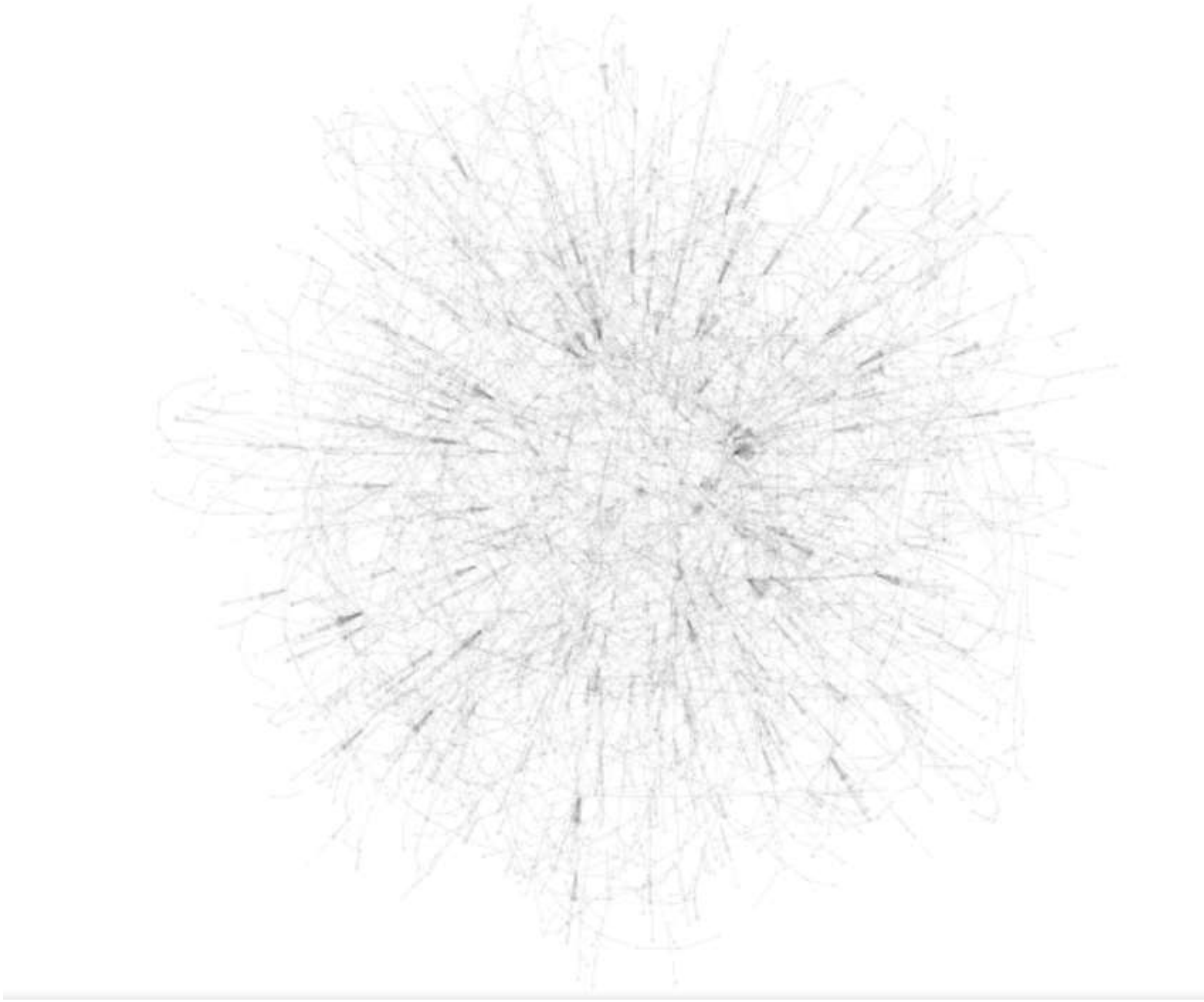
user-based

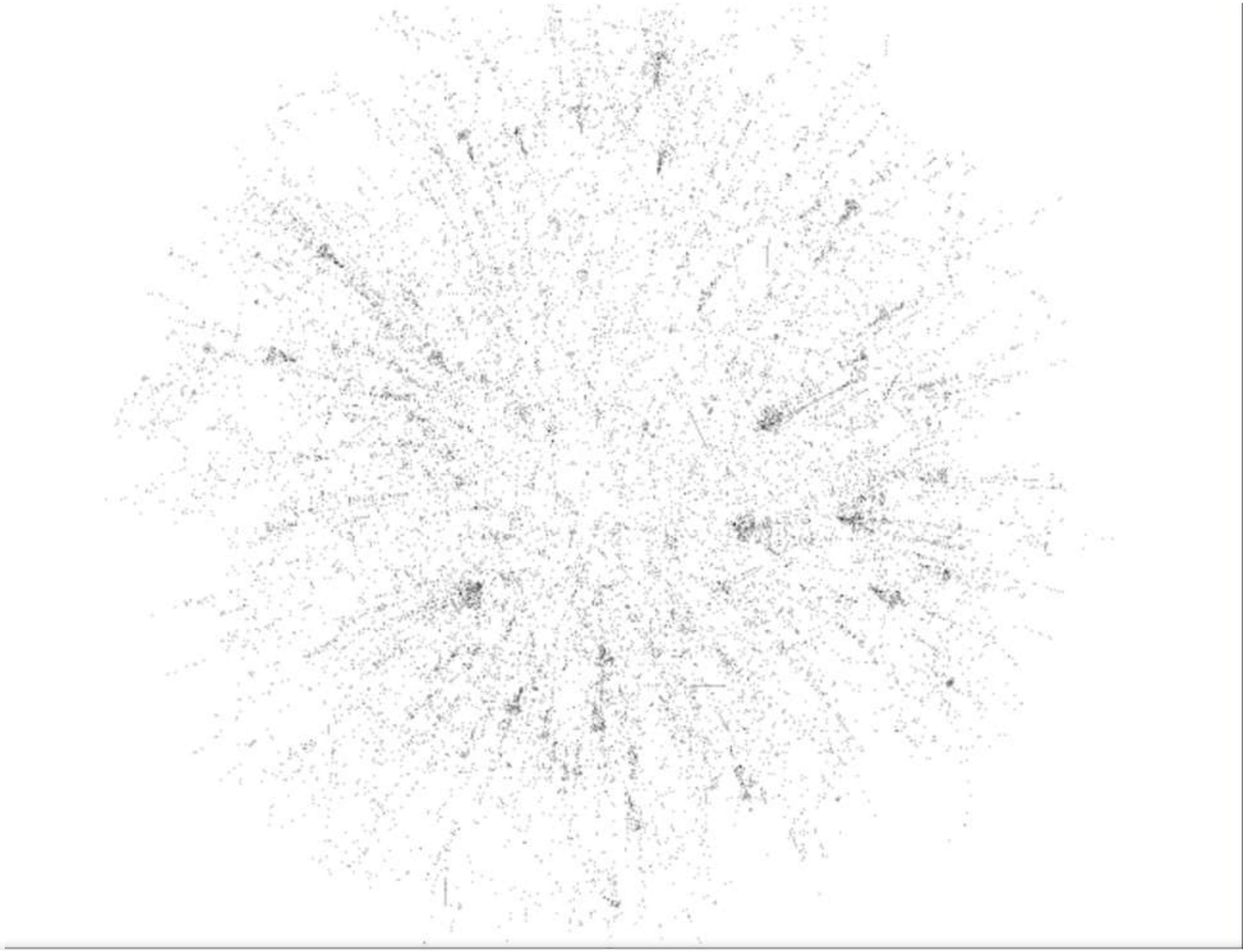
$$+\overline{r}_i + \beta \sum_{j=1}^{m} si(i,j)(r_{a,j} - \overline{r}_j)$$

item-based

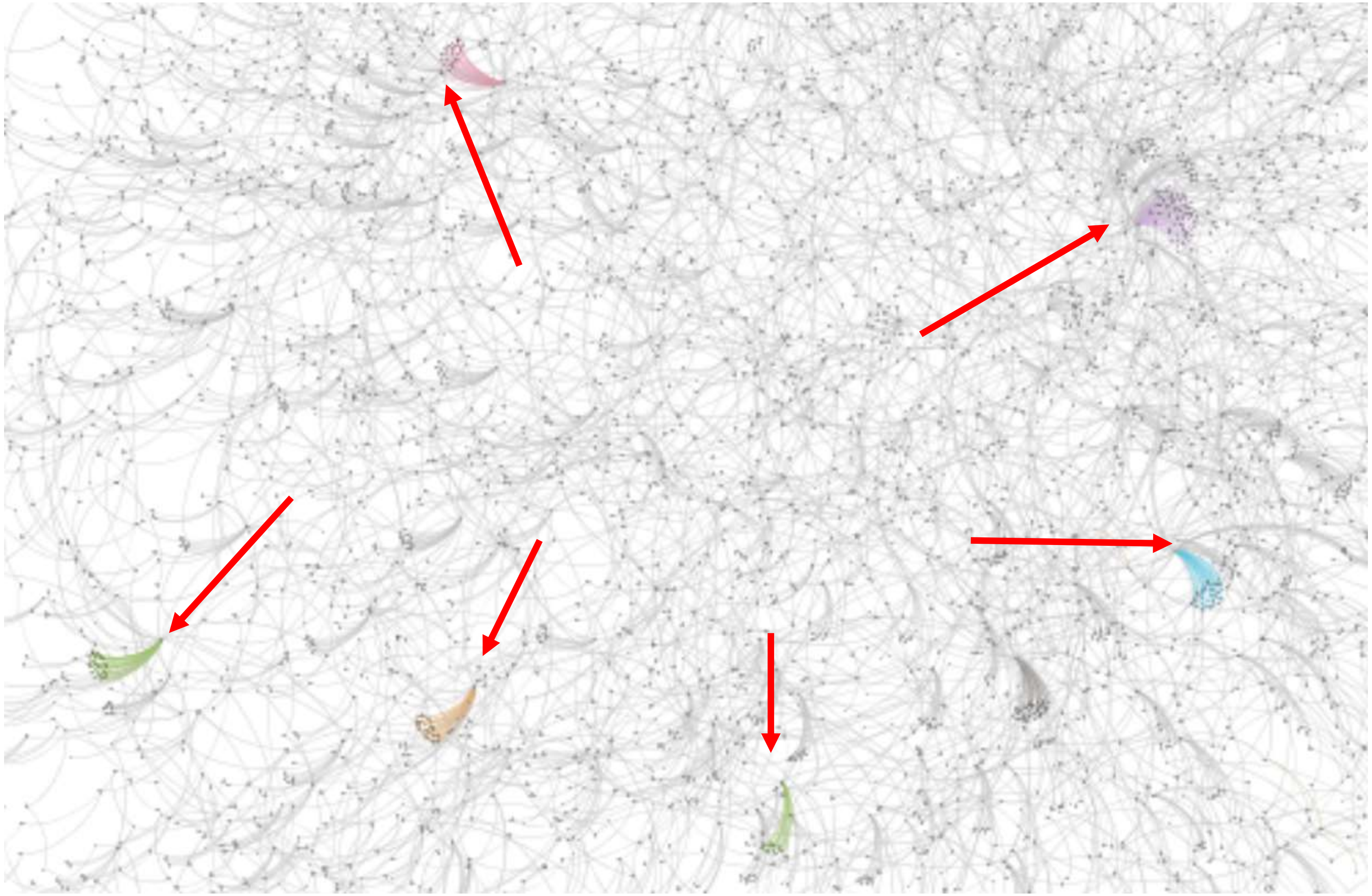
own average + coefficient*sum (likeliness*neighbor average)

Visualization





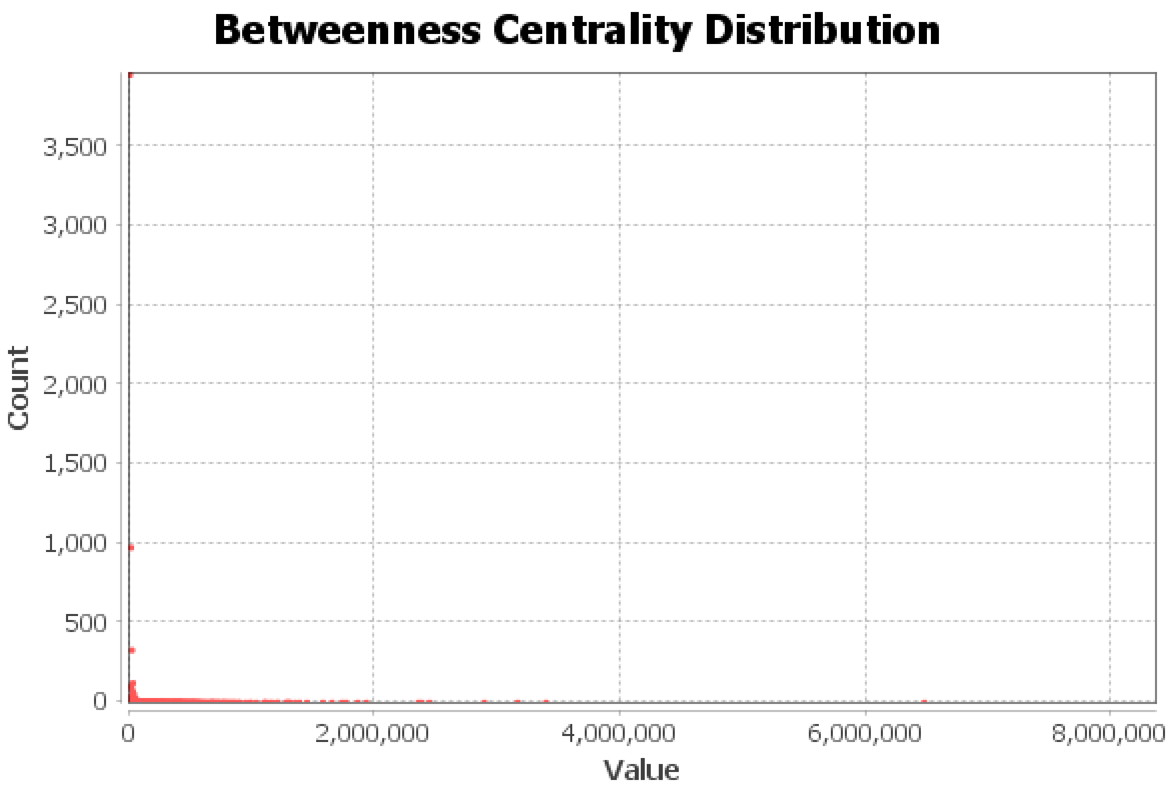
Visualization

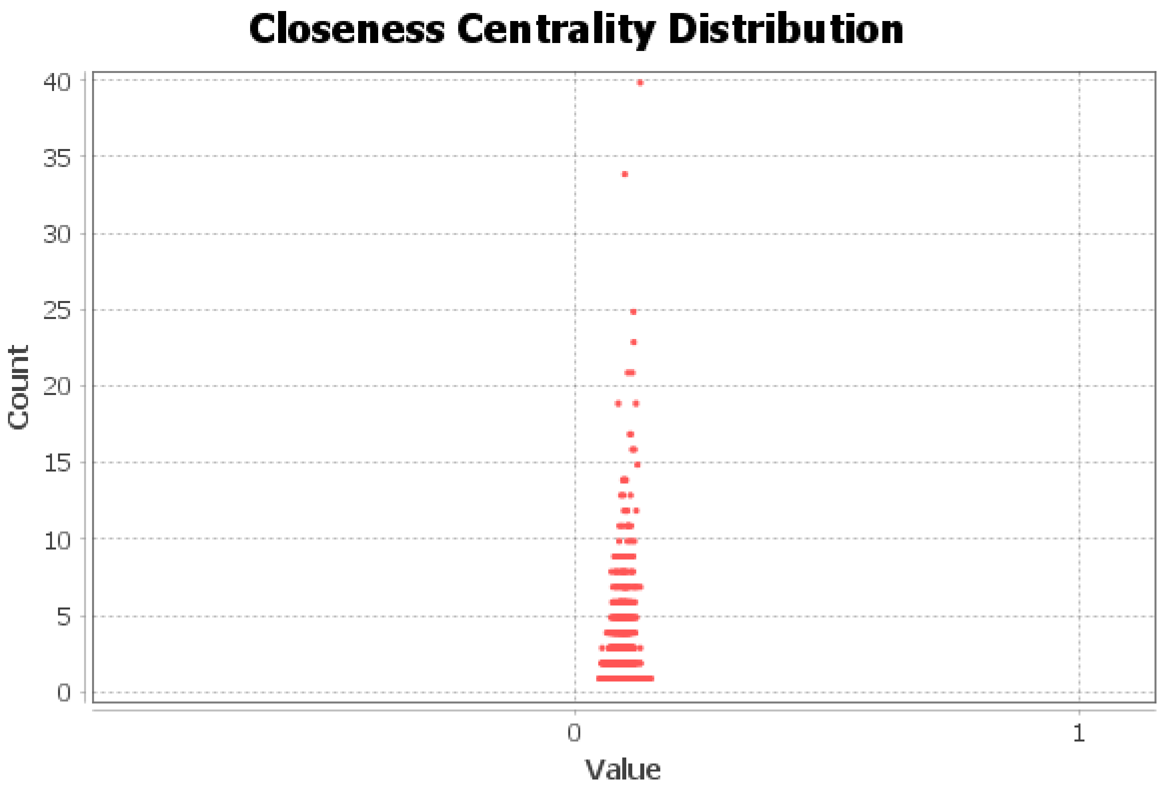


Betweenness Centrality: size

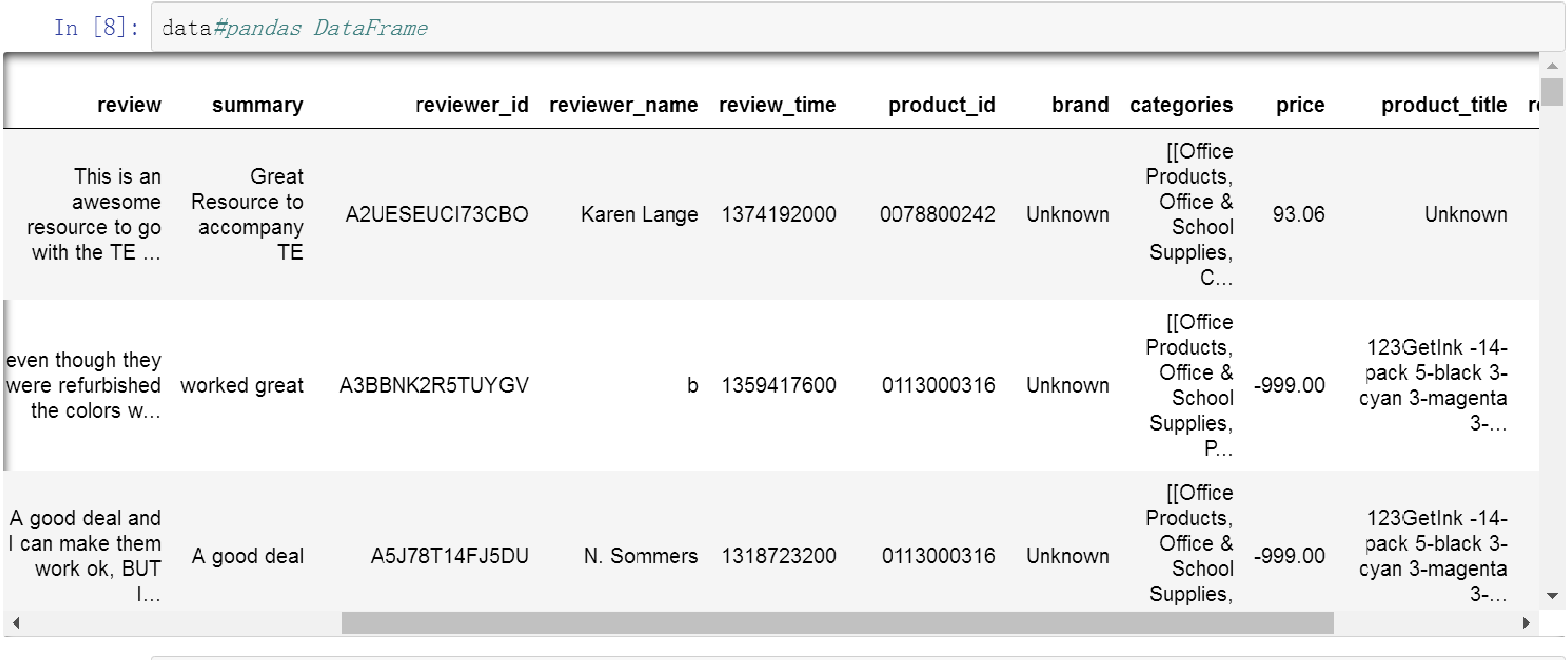
Closeness Centrality: color

Visualization





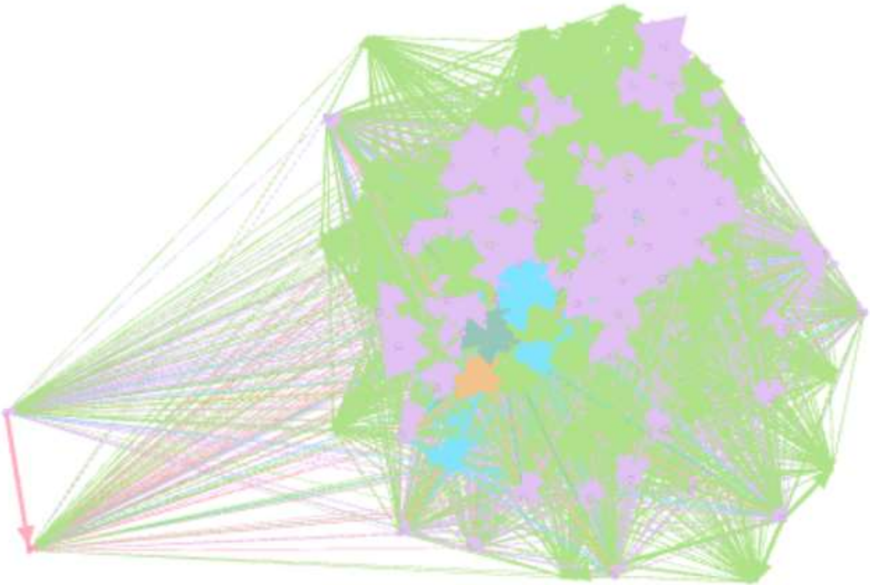
Similarity

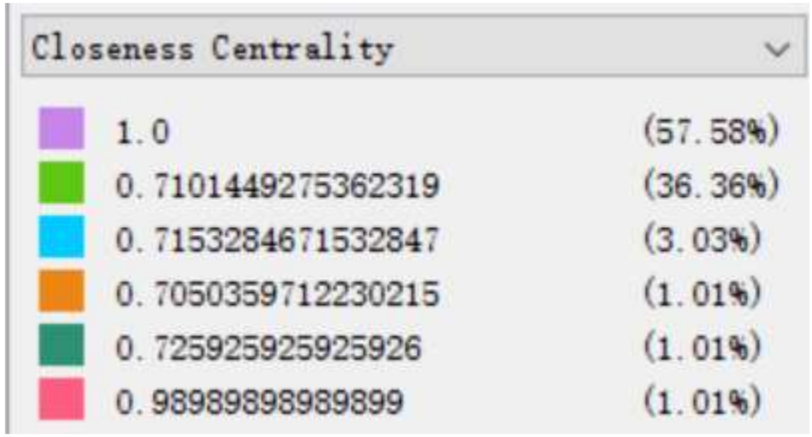


```
In [8]: data#pandas DataFrame
```

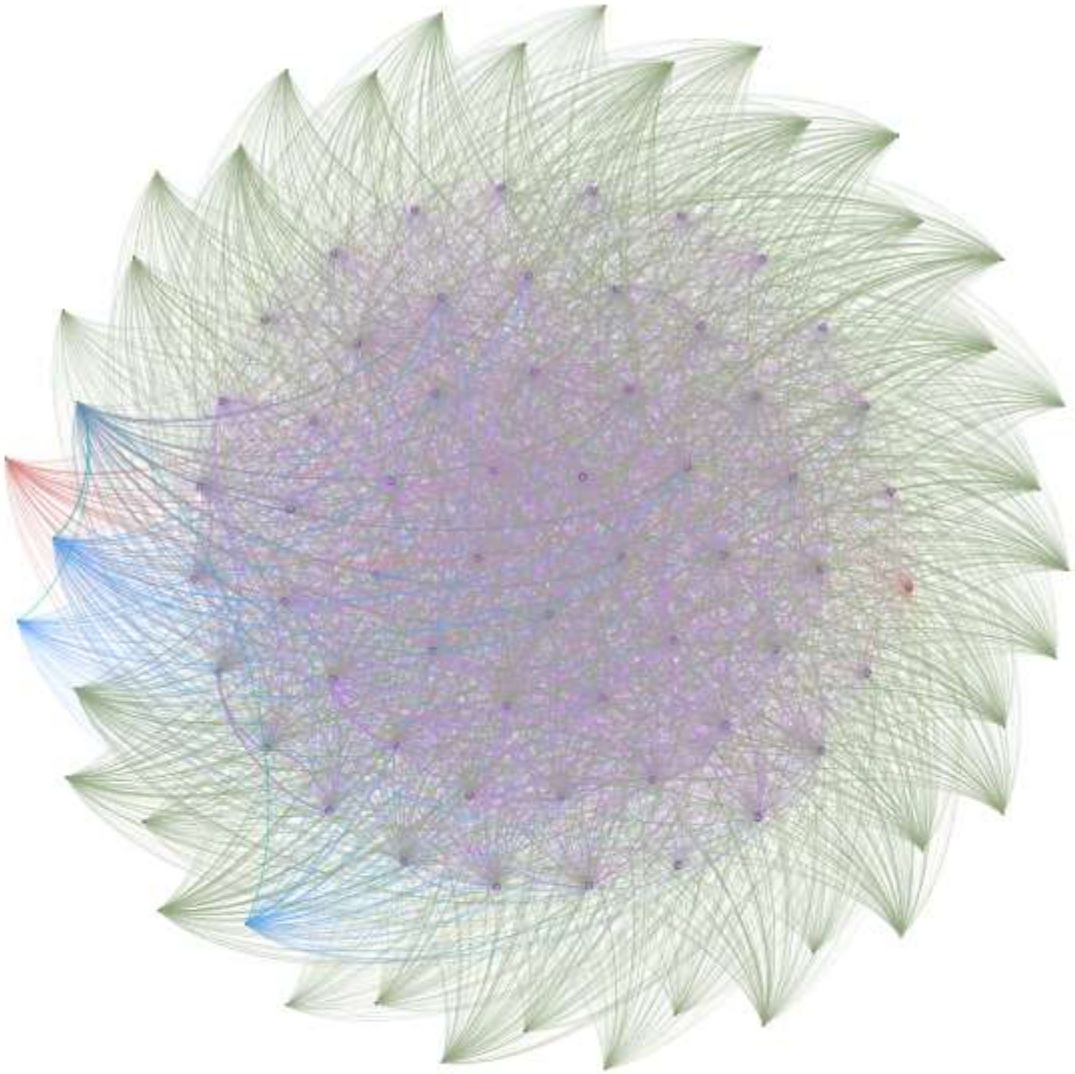
	review	summary	reviewer_id	reviewer_name	review_time	product_id	brand	categories	price	product_title	r
	This is an awesome resource to go with the TE ...	Great Resource to accompany TE	A2UESEUCI73CBO	Karen Lange	1374192000	0078800242	Unknown	[[Office Products, Office & School Supplies, C...	93.06	Unknown	
	even though they were refurbished the colors w...	worked great	A3BBNK2R5TUYGV	b	1359417600	0113000316	Unknown	[[Office Products, Office & School Supplies, P...	-999.00	123GetInk -14-pack 5-black 3-cyan 3-magenta 3-...	
	A good deal and I can make them work ok, BUT I...	A good deal	A5J78T14FJ5DU	N. Sommers	1318723200	0113000316	Unknown	[[Office Products, Office & School Supplies,	-999.00	123GetInk -14-pack 5-black 3-cyan 3-magenta 3-...	

Visualization

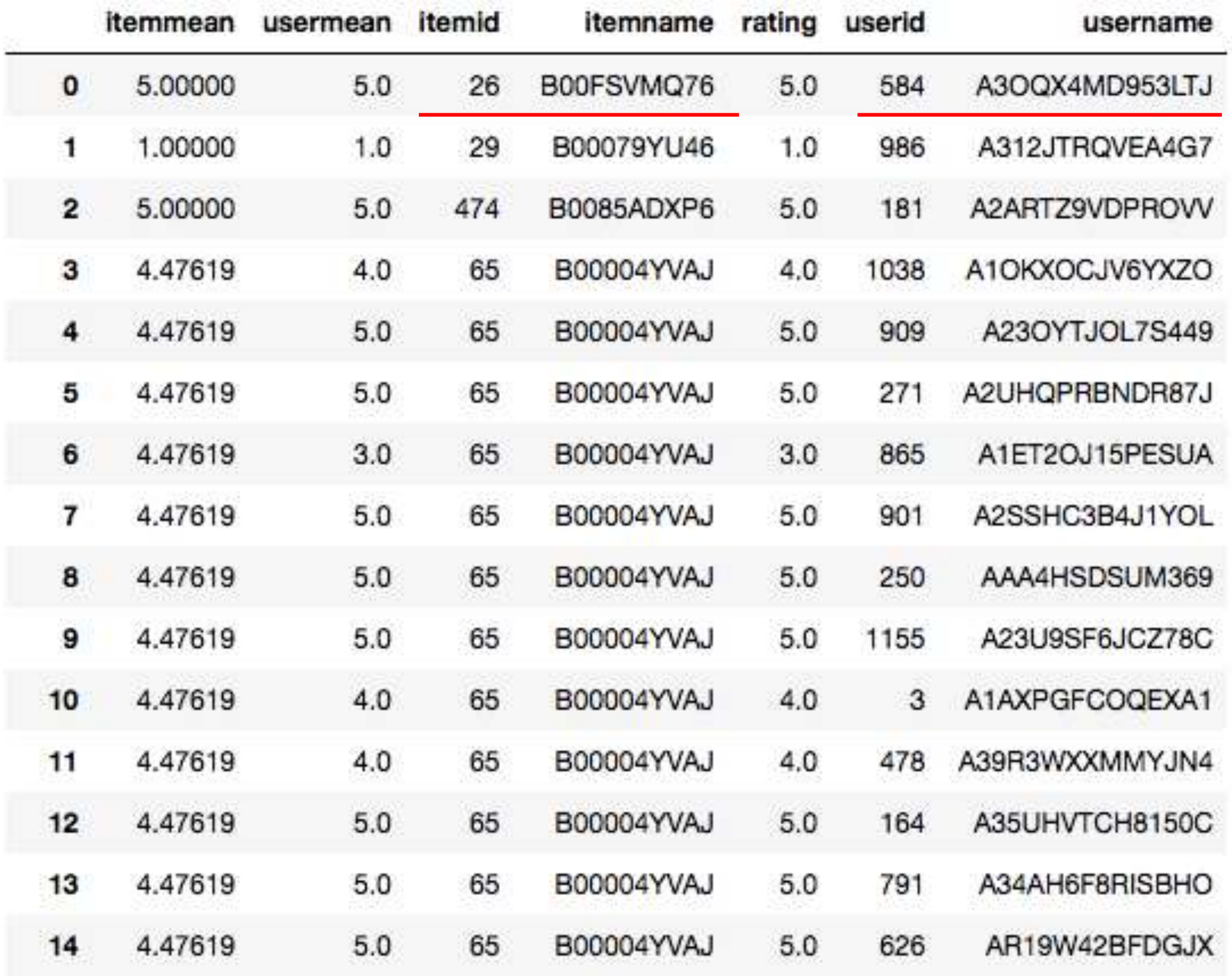




🟪	1.0	(57.58%)
🟩	0.7101449275362319	(36.36%)
🟦	0.7153284671532847	(3.03%)
🟧	0.7050359712230215	(1.01%)
🟩	0.725925925925926	(1.01%)
🟥	0.98989898989899	(1.01%)

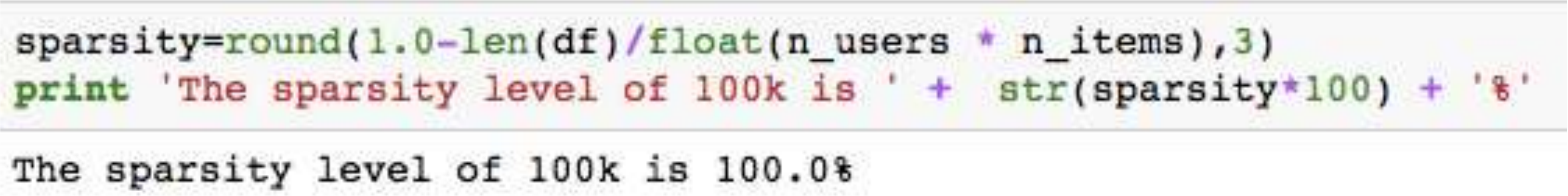


Preprocess



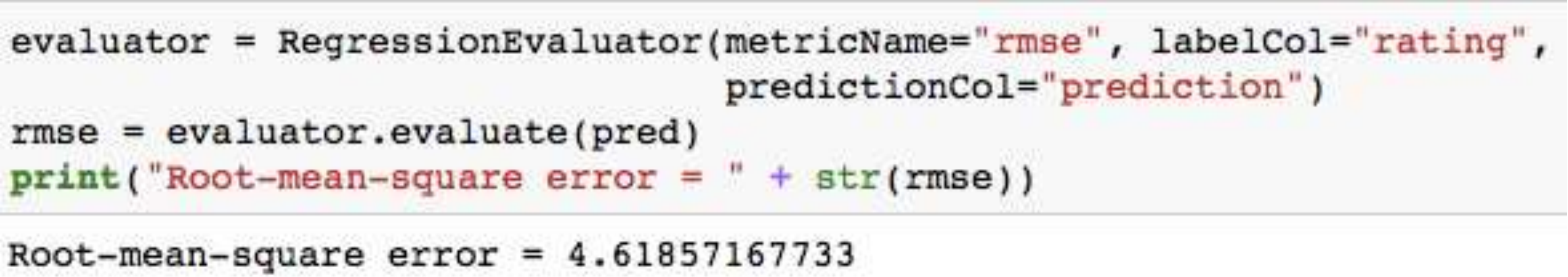
	itemmean	usermean	itemid	itemname	rating	userid	username
0	5.00000	5.0	26	B00FSVMQ76	5.0	584	A3OQX4MD953LTJ
1	1.00000	1.0	29	B00079YU46	1.0	986	A312JTRQVEA4G7
2	5.00000	5.0	474	B0085ADXP6	5.0	181	A2ARTZ9VDPROVV
3	4.47619	4.0	65	B00004YVAJ	4.0	1038	A1OKXOCJV6YXZO
4	4.47619	5.0	65	B00004YVAJ	5.0	909	A23OYTJOL7S449
5	4.47619	5.0	65	B00004YVAJ	5.0	271	A2UHQPRBNDR87J
6	4.47619	3.0	65	B00004YVAJ	3.0	865	A1ET2OJ15PESUA
7	4.47619	5.0	65	B00004YVAJ	5.0	901	A2SSHC3B4J1YOL
8	4.47619	5.0	65	B00004YVAJ	5.0	250	AAA4HSDSUM369
9	4.47619	5.0	65	B00004YVAJ	5.0	1155	A23U9SF6JCZ78C
10	4.47619	4.0	65	B00004YVAJ	4.0	3	A1AXPGFCOQEXA1
11	4.47619	4.0	65	B00004YVAJ	4.0	478	A39R3WXXMMYJN4
12	4.47619	5.0	65	B00004YVAJ	5.0	164	A35UHVTCH8150C
13	4.47619	5.0	65	B00004YVAJ	5.0	791	A34AH6F8RISBHO
14	4.47619	5.0	65	B00004YVAJ	5.0	626	AR19W42BFDGJX

Why not ALS?



```
sparsity=round(1.0-len(df)/float(n_users * n_items),3)
print 'The sparsity level of 100k is ' +  str(sparsity*100) + '%'
```

```
The sparsity level of 100k is 100.0%
```



```
evaluator = RegressionEvaluator(metricName="rmse", labelCol="rating",
                                predictionCol="prediction")
rmse = evaluator.evaluate(pred)
print("Root-mean-square error = " + str(rmse))
```

```
Root-mean-square error = 4.61857167733
```

Parameter setting

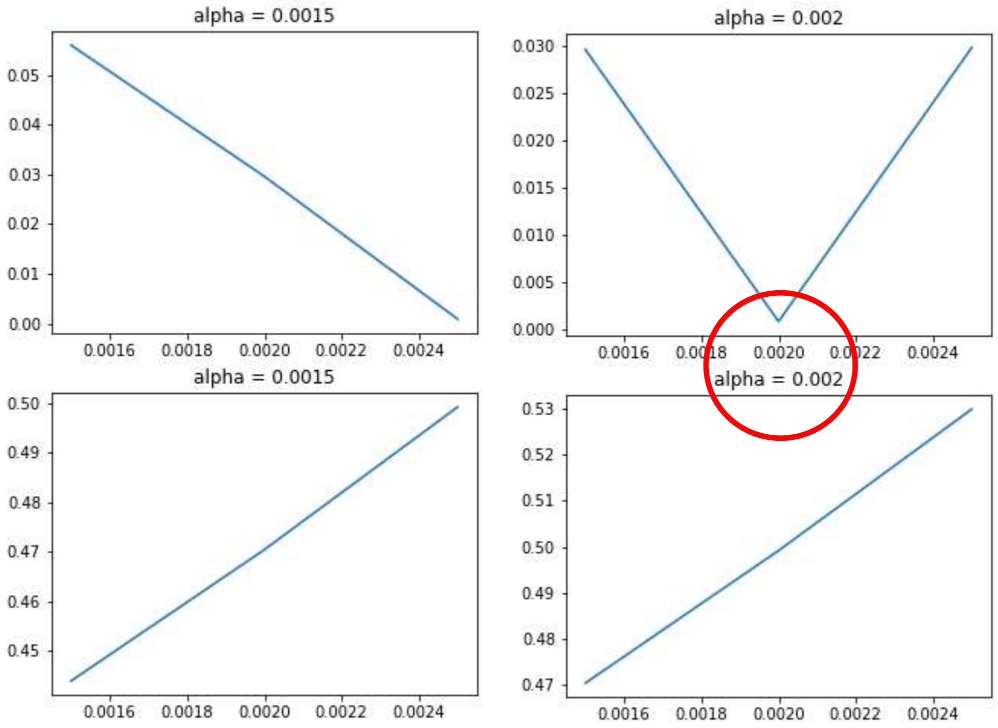
$$P_{a,\mathrm{i}} = \overline{r_a} + \alpha \sum_{b=1}^{n} su(a,b)(r_{b,i} - \overline{r}_b)$$

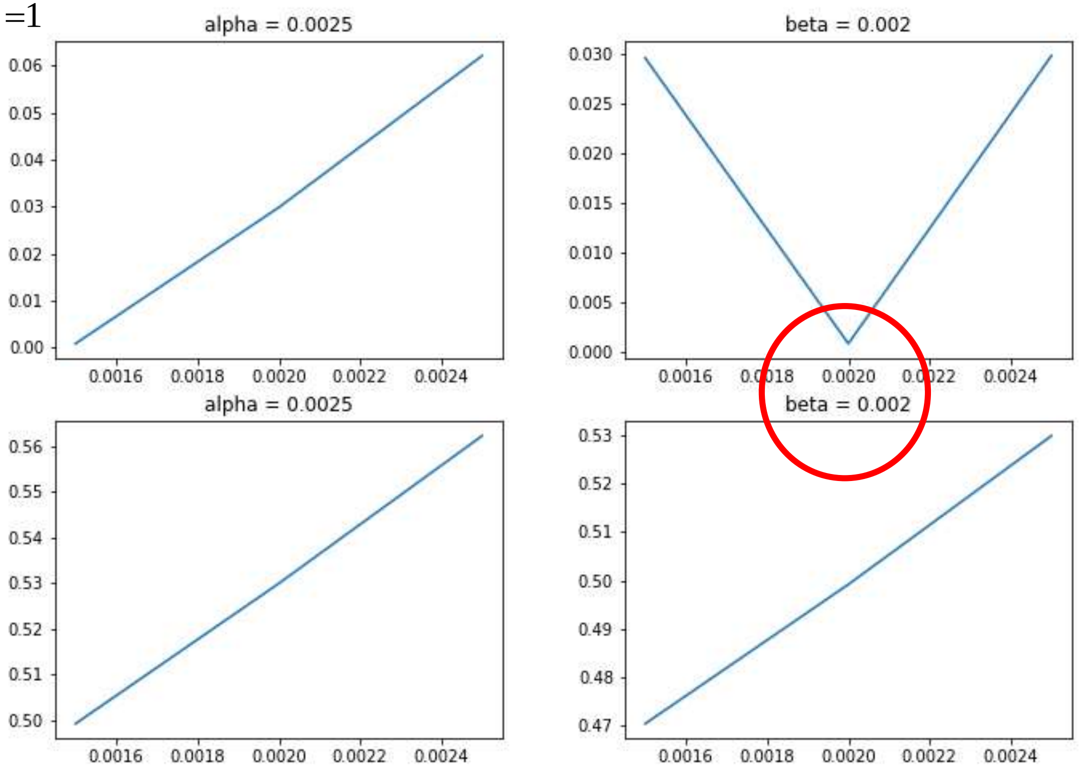
$$+\overline{r_i} + \beta \sum_{j=1}^{m} si(i,j)(r_{a,j} - \overline{r}_j)$$

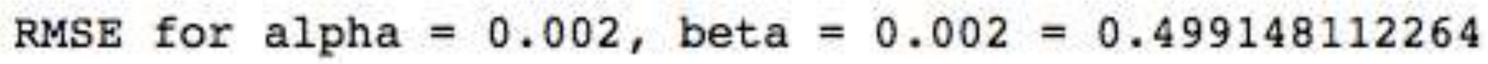
Parameter setting

$$P_{a,\mathrm{i}} = \overline{r_a} + \alpha \sum_{b=1}^{n} su(a,b)(r_{b,i} - \overline{r}_b)$$

$$+\overline{r_i} + \beta \sum_{j=1}^{m} si(i,j)(r_{a,j} - \overline{r}_j)$$

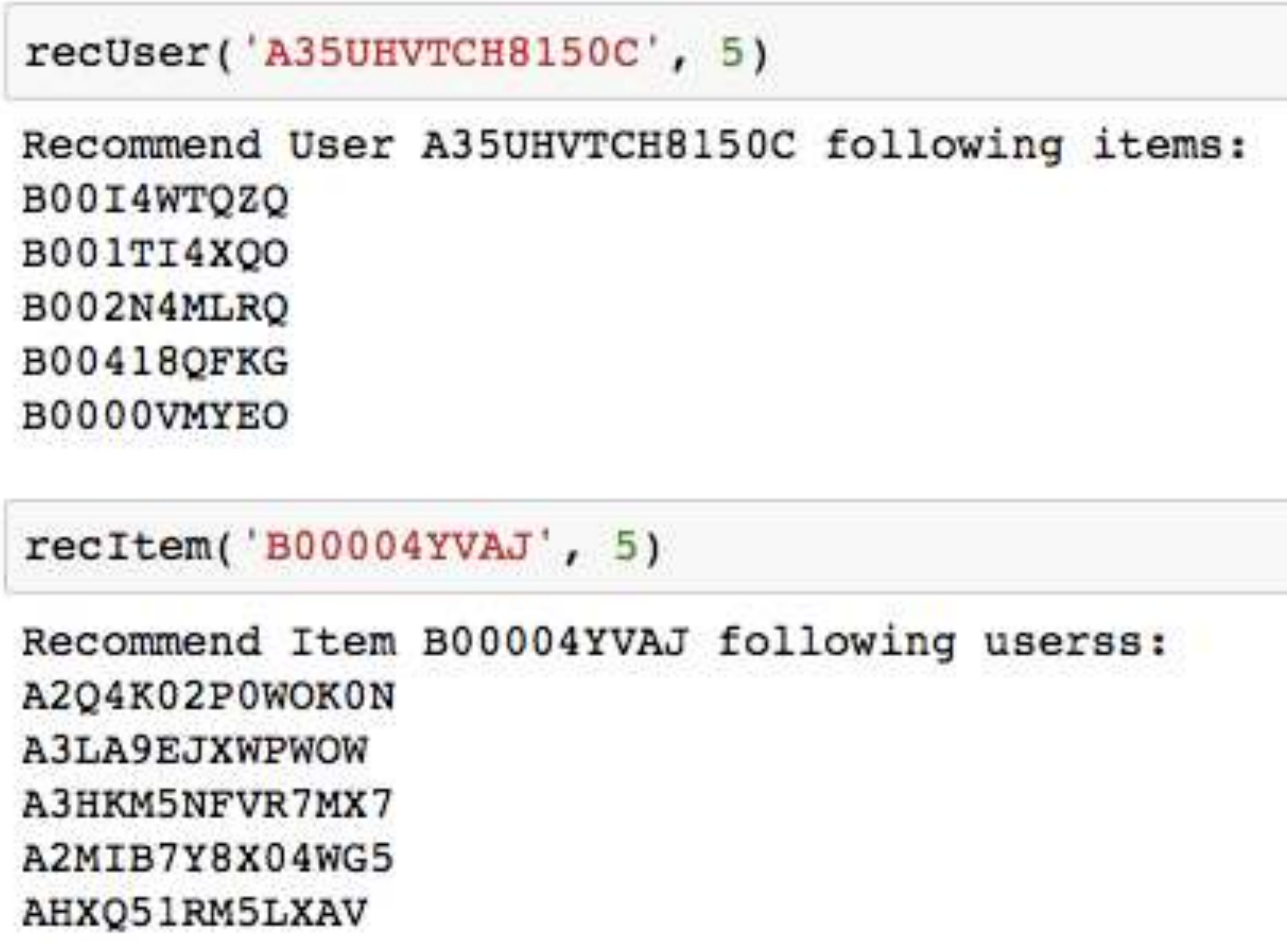






```
RMSE for alpha = 0.002, beta = 0.002 = 0.499148112264
```

Result



```
recUser('A35UHVTCH8150C', 5)
```

```
Recommend User A35UHVTCH8150C following items:
B00I4WTQZQ
B001TI4XQO
B002N4MLRQ
B00418QFKG
B0000VMYEO
```

```
recItem('B00004YVAJ', 5)
```

```
Recommend Item B00004YVAJ following userss:
A2Q4K02P0WOK0N
A3LA9EJXWPWOW
A3HKM5NFVR7MX7
A2MIB7Y8X04WG5
AHXQ51RM5LXAV
```