Amazon Reviews Analysis

Group7:Miko Liu, Mengwei Li, George Bi, Maggie Wu

Agenda

Business Objectives

Data Background

Exploratory Data Analysis

Sentiment Analysis Models

Recommendation System

Graph

7 Conclusion & Improvements

Business Objectives

Predict customer reaction to products and give them appropriate recommendation

- 32.4% growth in 2020, 17.9% in 2021, and will grow to 23.6% and reach 1.6 trillion by 2025
- Big Data is changing the E-commerce game by helping company better understand their clients and forecast consumer behavior patterns and increase revenue

amazon

- One of the Big Five companies in the U.S. information technology industry
- The largest and most successful retailer in the western world
- Occupied almost 50% shares of e-commerce sales in the U.S

Methodology and Evaluations

Goal	Methodology	Evaluate
Predict Customer Ratings for better CRM	- NLP on reviews - Sentimental analysis	AUC
Recommend products to boost potential revenue	Model Based Collaborative Filtering: Alternating Least Square Recommendater	Accuracy RMSE

Datasets Overview and Challenges

Datasets	
5_core_Reviews Data	Product Metadata
Dataset size	
90GB(24 Categories)	30GB
Mainly Used for	
EDA/Sentiment Analysis/Graphs	Recommendation System
Main columns	
ReviewText,ReviewerID,ProductID,Ratings,Time	Category,Description,ProductID,title
Dataset Source	
GitHub and Repository of Recommender Systems Datasets from UCSD	

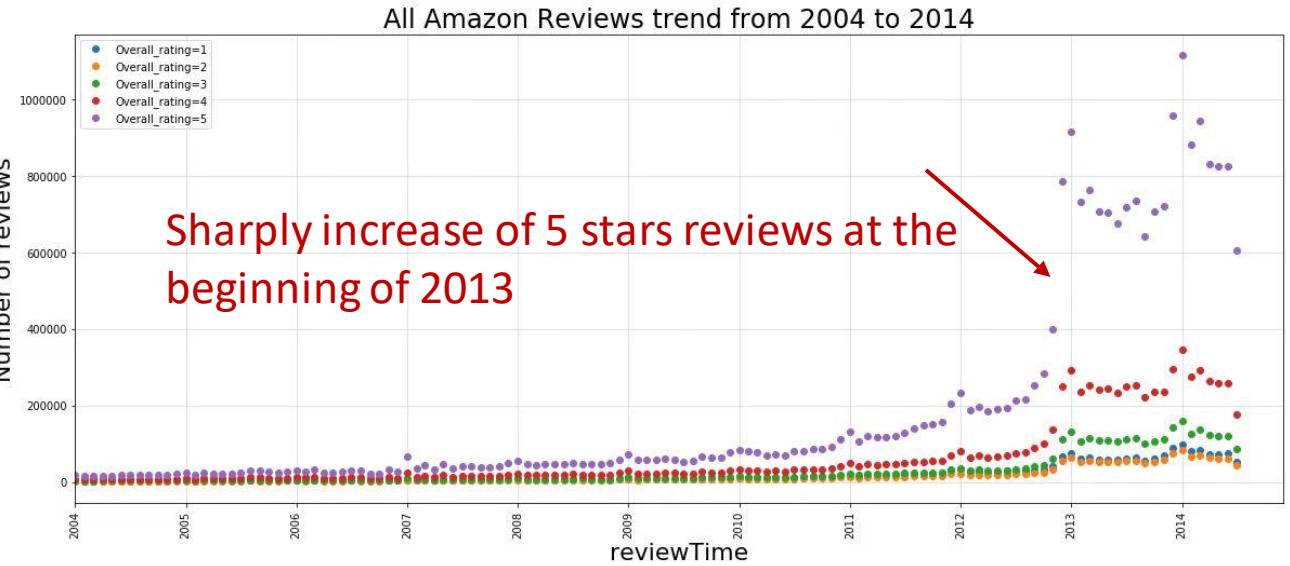
Exploratory Data Analysis

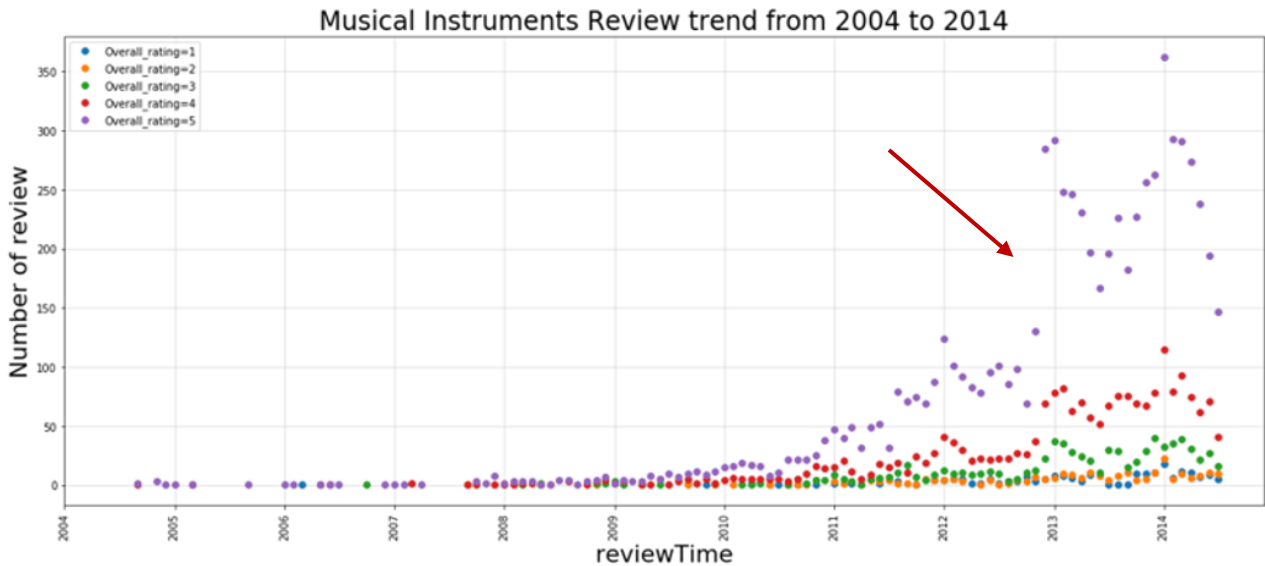
Explored Reviews Trend (Yearly and Monthly)

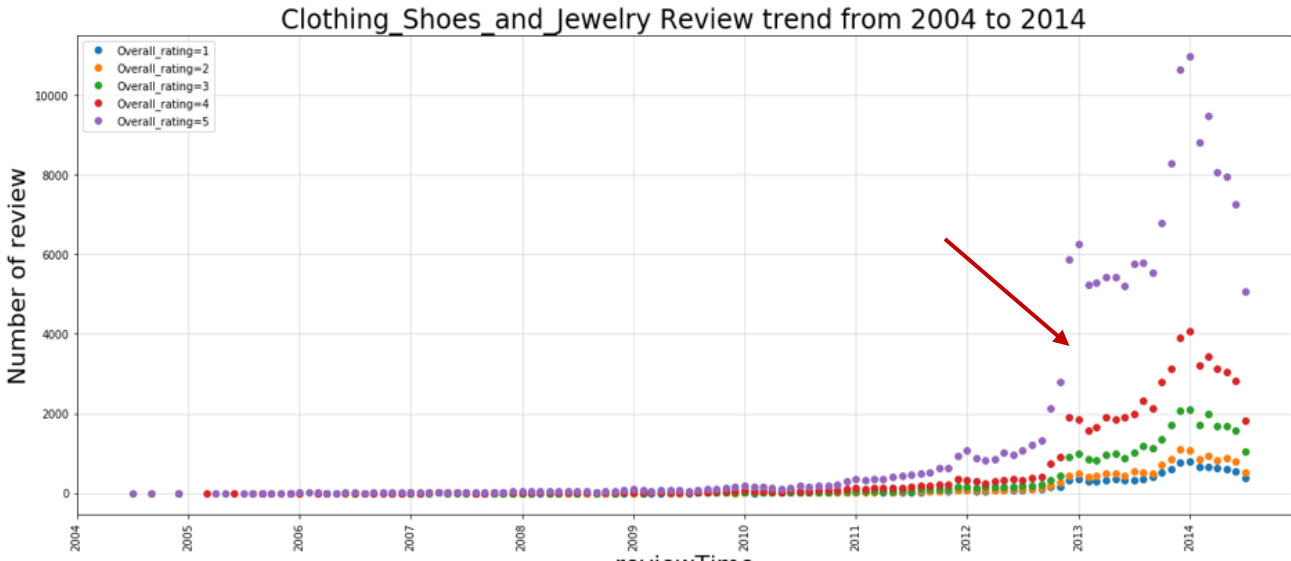
Compared the trends by categories

Relationship between Reviewers and Products

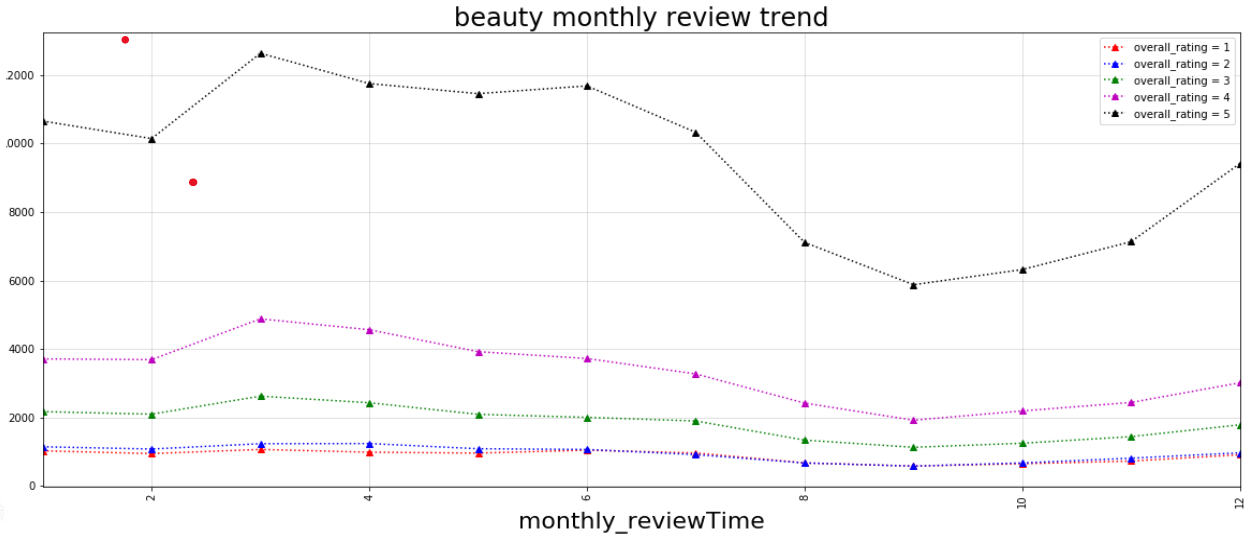
Yearly Reviews Trend



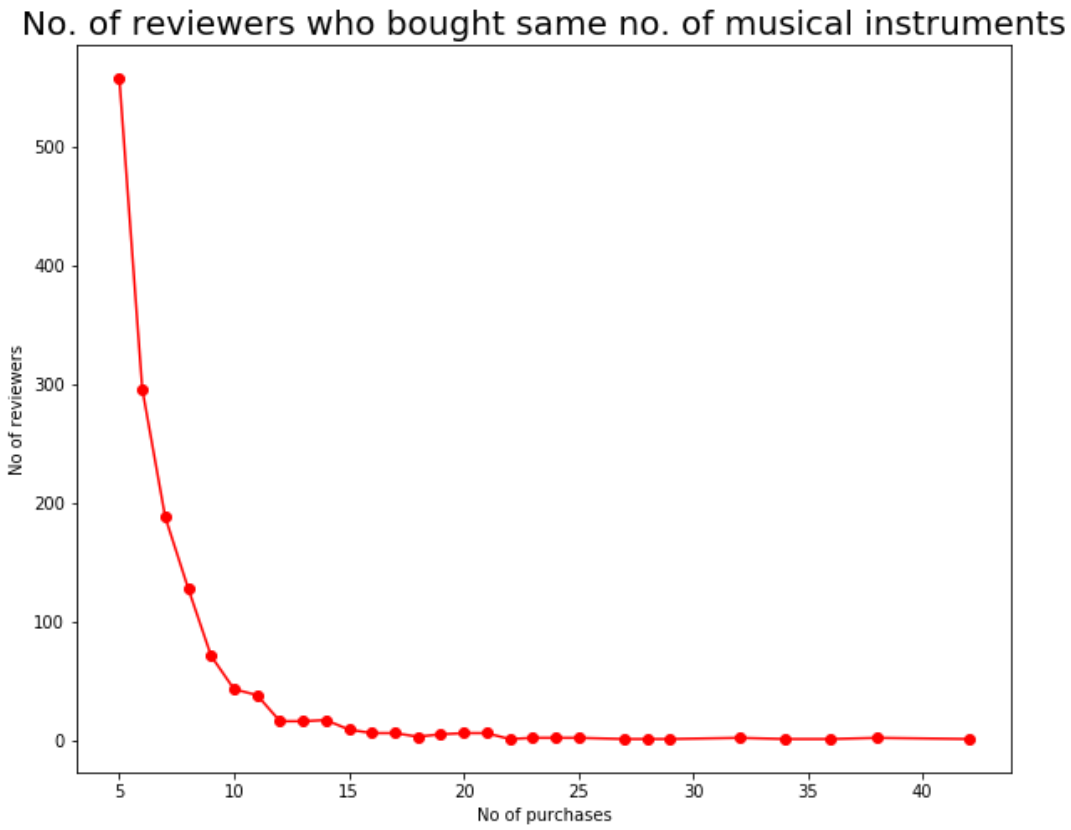


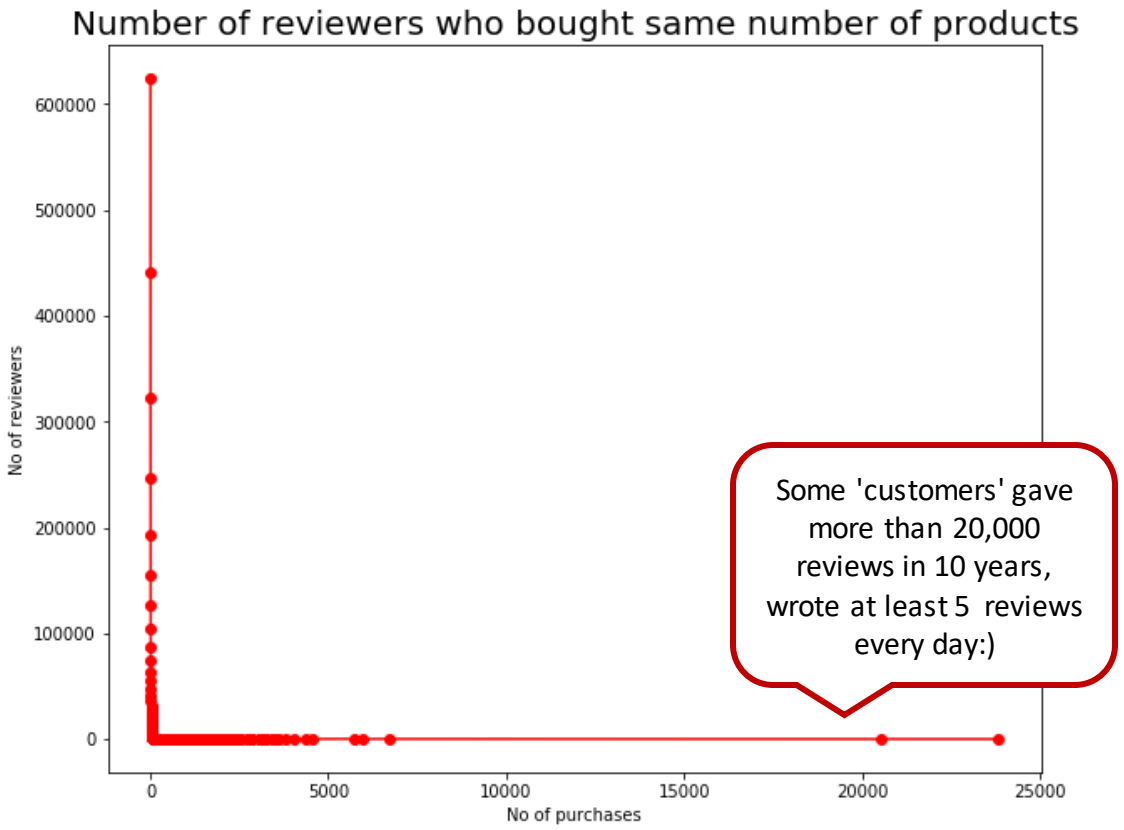


Monthly Reviews Trend

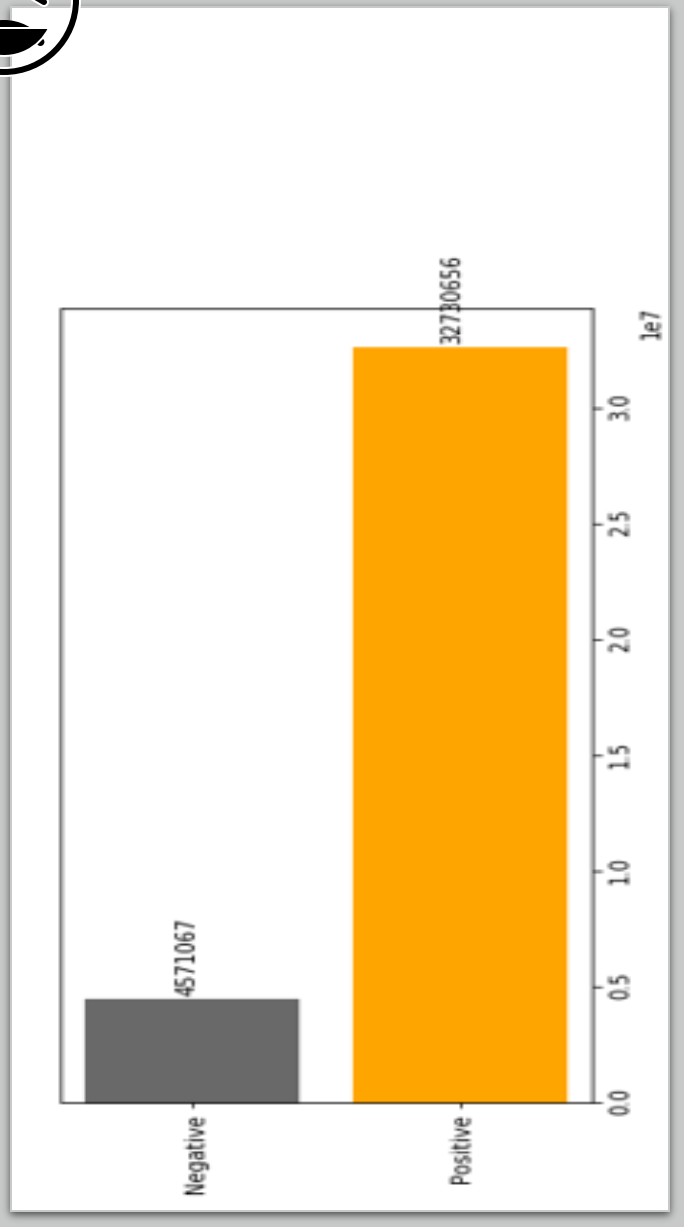


Reviewers and Products





Sentiment Analysis



Data Cleaning

- Remove Null/Nan/Empty/only contains blanks rows
- Remove Duplicated rows

Data Preparation

- Drop overall = 3
- Combine 1 & 2 to Negative, **4 & 5 to Positive**

Pipeline for models

NB/LR

- RegexTokenizer(**It considers a string (a stream of text) without space but with.**)
- Stopwords Remover(with added stop words('http'))
- CountVectorizer(minDF=5)

NB/LR/RF

- Stringindexer

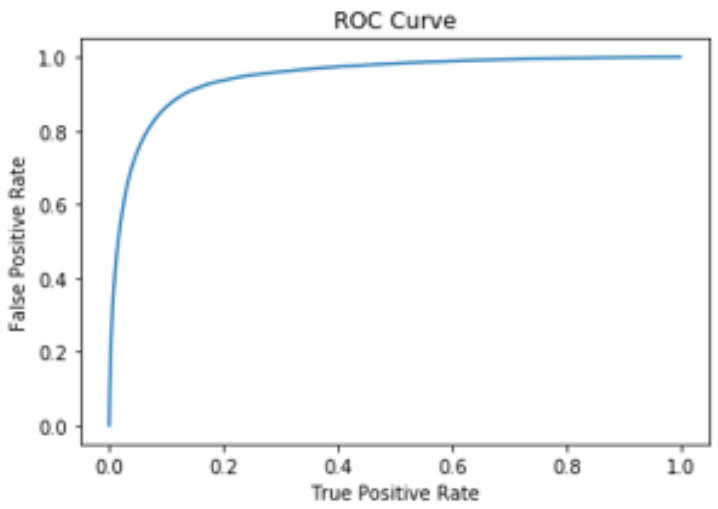
RF

- LabelIndexer
- FeatureIndexer
- Random forest model
- labelConverter

```
+-------+--------------------+--------------------+--------------------+--------------------+-----+
|overall|          reviewText|               words|            filtered|            features|label|
+-------+--------------------+--------------------+--------------------+--------------------+-----+
|    1.0|We use this type ...|[we, use, this, t...|[we, use, this, t...|(10000,[1,2,3,4,5...|  0.0|
|    1.0|I bought this for...|[i, bought, this,...|[i, bought, this,...|(10000,[1,2,3,6,7...|  0.0|
|    1.0|This is a large s...|[this, is, a, lar...|[this, is, a, lar...|(10000,[2,3,6,7,1...|  0.0|
|    1.0|We use this hymn ...|[we, use, this, h...|[we, use, this, h...|(10000,[0,1,3,4,5...|  0.0|
|    1.0|This work bears d...|[this, work, bear...|[this, work, bear...|(10000,[0,1,2,3,5...|  0.0|
+-------+--------------------+--------------------+--------------------+--------------------+-----+
```

Models Evaluation

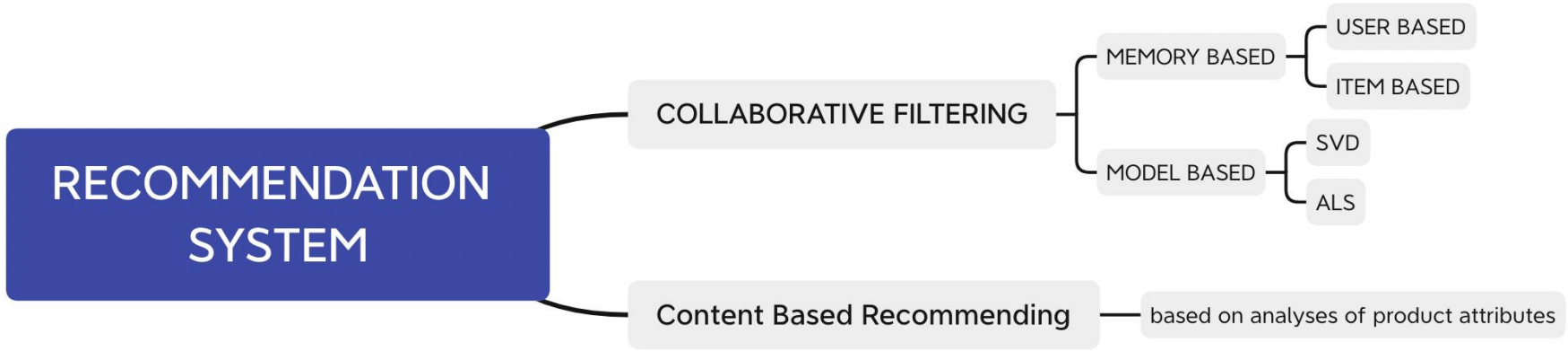
File size	Naïve Bayes	Logistic Regression	Random Forest
90GB	0.44	0.94	Kernel Dead
30GB	0.56	0.92	0.91
5GB	0.52	0.93	0.90




```
+-----------------------------+-------+-----------------------------+-----+----------+
|                   reviewText|overall|                  probability|label|prediction|
+-----------------------------+-------+-----------------------------+-----+----------+
|Junior and narcotics are so...|    0.0|[0.4970094580451036,0.50299...|  1.0|       1.0|
|Sadly, Chris Wallace was ki...|    0.0|[0.4960177256678562,0.50398...|  1.0|       1.0|
|This isn't songwriting, and...|    0.0|[0.49299387149982227,0.5070...|  1.0|       1.0|
|Seeing so many 5 star revie...|    0.0|[0.49217888854854647,0.5078...|  1.0|       1.0|
|Rihanna is an artist who's ...|    0.0|[0.4906702979701311,0.50932...|  1.0|       1.0|
|I bought this album because...|    0.0|[0.48637401843090056,0.5136...|  1.0|       1.0|
|As my title says, Jeremy is...|    0.0|[0.4848517177578043,0.51514...|  1.0|       1.0|
|Good Charlotte are not a pu...|    0.0|[0.48320785177266773,0.5167...|  1.0|       1.0|
|For some weird reason, reco...|    0.0|[0.48232846453805983,0.5176...|  1.0|       1.0|
|This album is so bad that i...|    0.0|[0.48222289571715127,0.5177...|  1.0|       1.0|
+-----------------------------+-------+-----------------------------+-----+----------+
```

RECOMMENDATION SYSTEM

types of recommendation system



hybrid based recommending approach: Grey Sheep Problem

Gray sheep is related to the users whose opinions do not consistently agree or disagree with any group of people. black sheep have so specific tastes that recommending to them or using their opinions for recommendations to others are nearly impossible. Hybrid approach combines content-based and CF recommendations by basing a prediction on a weighted average of the content-based prediction and the CF prediction. Weights of the content-based and CF predictions can be determined on a per-user basis, allowing the system to determine the optimal mix of content-based and CF recommendation for each user, helping to solve the gray sheep problem

Alternating Least Square

A = X * Y

GIVEN FIX

EXPRESSED IN TERMS OF A AND Y

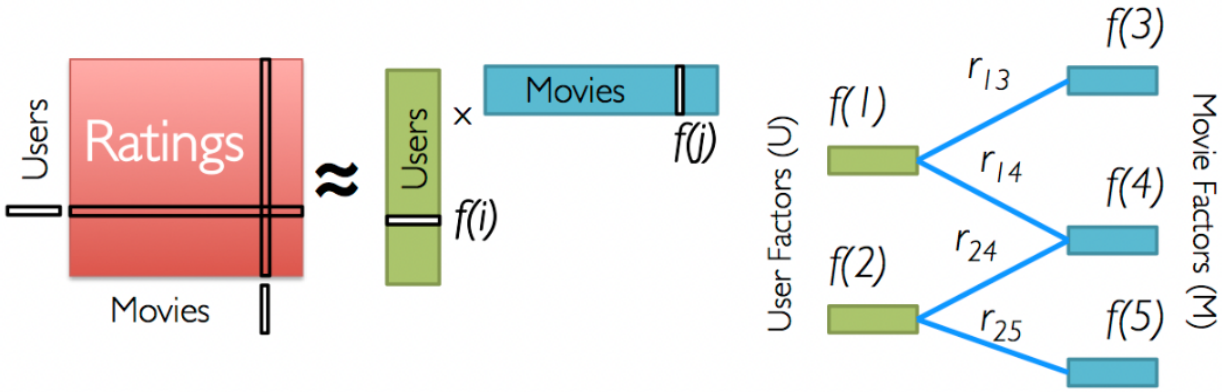
$$A_i Y \left(Y^T Y\right)^{-1} = X_i$$

CANNOT ACHIEVE ABSOLUTE EQUALITY, MINIMIZATION !

$$\sum \left(A_i Y \left(Y^T Y\right)^{-1} - X_i\right)^2 \longrightarrow \hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

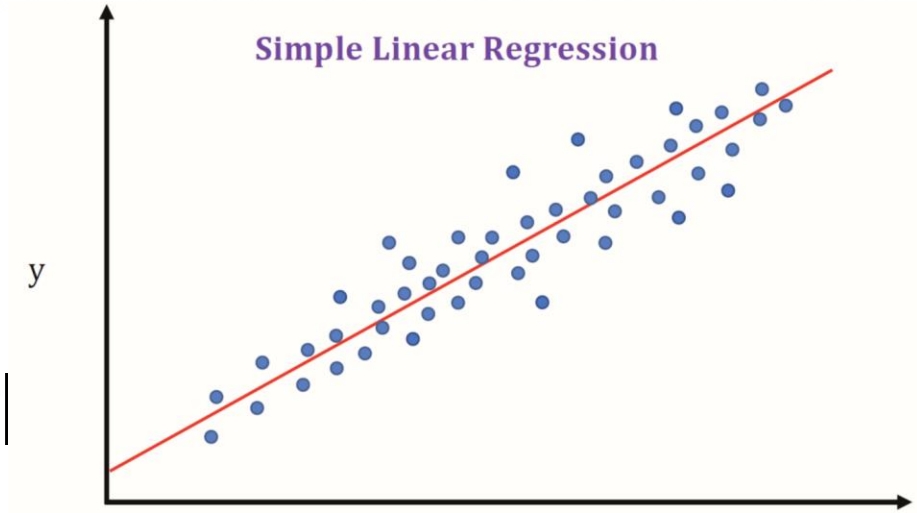
CONNECTION TO LR

Low-Rank Matrix Factorization:



Iterate:

$$f[i] = \arg\min_{w \in \mathbb{R}^d} \sum_{j \in \mathrm{Nbrs}(i)} \left(r_{ij} - w^T f[j]\right)^2 + \lambda ||w||_2^2$$



Models, Parameter tuning and Result

```
# Build the recommendation model using ALS on the training data
# Note we set cold start strategy to 'drop' to ensure we don't get NaN evaluation metrics
als = ALS(maxIter=10, regParam=0.01, userCol="reviewerIDIndex", itemCol="asinIndex", ratingCol="overall",
          coldStartStrategy="drop", nonnegative = True)
```

```
# Build cross validation using CrossValidator
cv3 = CrossValidator(estimator=als, estimatorParamMaps=param_grid3, evaluator=evaluator, numFolds=3)
```

```
# Add hyperparameters and their respective values to param_grid
param_grid3 = ParamGridBuilder() \
            .addGrid(als.rank, [50,100,150]) \
            .addGrid(als.regParam, [0.01,0.05,0.1,0.15]) \
            .build()
```

```
**Best Model**
  Rank: 150
  MaxIter: 10
  RegParam: 0.15
```

```
# View the predictions
test_predictions3 = best_model3.transform(test)
RMSE3 = evaluator.evaluate(test_predictions3)
print(RMSE3)
```

```
0.9824782504983577
```

Books database

```
books1_index.dtypes
```

```
[('asin', 'string'),
 ('helpful', 'array<bigint>'),
 ('overall', 'double'),
 ('reviewText', 'string'),
 ('reviewTime', 'string'),
 ('reviewerID', 'string'),
 ('reviewerName', 'string'),
 ('summary', 'string'),
 ('unixReviewTime', 'bigint'),
 ('asinIndex', 'double'),
 ('reviewerIDIndex', 'double')]
```

Meta books database

```
Meta_Books.select("asin","category",'price','title').show(10)
```

asin	category	price	title
0000092878	[]	$39.94	Biology Gods Livi...
000047715X	[Books, New, Used...		Mksap 16 Audio Co...
0000004545	[Books, Arts & Ph...	$199.99	Flex! Discography...
0000013765	[Books, Arts & Ph...		Heavenly Highway ...
0000000116	[]	$164.10	Georgina Goodman ...
0000555010	[Books, New, Used...		Principles of Ana...
0000477141	[Books, Medical B...		MKSAP 15 Audio Co...
0000230022	[Books, New, Used...		The Simple Truths...
0000038504	[Books, Education...	$198.70	Double-Speak: Fro...
0000001589	[]		LJ Classique Inte...

only showing top 10 rows

10 books Recommendation for all users

```
nrecommendations = df_recommendation.w
nrecommendations.limit(20).show()
```

reviewerIDIndex	asinIndex	overall
134	343863	11.708134
134	274360	11.634775
134	245530	11.588929
134	188082	11.539589
134	153367	11.535437
134	227642	11.518301
134	227242	11.516125
134	188687	11.504252
134	240691	11.491187
134	257191	11.477867
584	351360	10.753606
584	292159	10.599555
584	362007	10.584541
584	188082	10.328151
584	260682	10.3169365
584	257191	10.299189
584	197354	10.245995
584	188687	10.210522
584	341812	10.187984
584	281154	10.186992

USER 134 ACTUAL PREFERENCE

```
df134 = books_index_interpret.join(Meta_Books_df_interpret, on='asin').filter("reviewe
```

```
df134['category'].astype('str').value_counts()
```

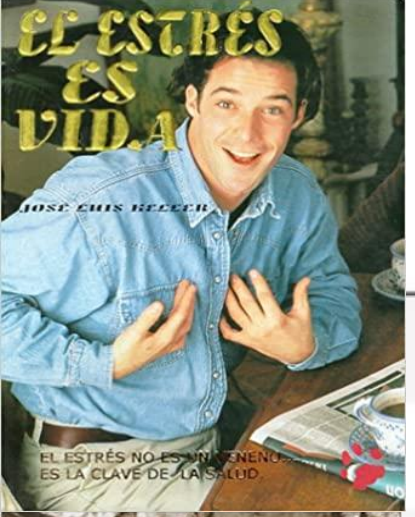
```
['Books', 'Literature &amp; Fiction', 'Genre Fiction']                          14
['Books', 'Mystery, Thriller &amp; Suspense', 'Thrillers &amp; Suspense']       11
['Books', 'Mystery, Thriller &amp; Suspense', 'Mystery']                         7
['Books', 'Science &amp; Math', 'Biological Sciences']                           7
['Books', 'Cookbooks, Food &amp; Wine', 'Regional &amp; International']          7
['Books', 'Literature & Fiction', 'Genre Fiction']                               6
['Books', 'Mystery, Thriller & Suspense', 'Mystery']                             4
['Books', 'Literature &amp; Fiction', 'United States']                           3
['Books', 'Cookbooks, Food &amp; Wine', 'Cooking Education &amp; Reference']     3
['Books', 'Literature &amp; Fiction', 'Contemporary']                            3
['Books', 'Mystery, Thriller & Suspense', 'Thrillers & Suspense']                3
['Books', 'Humor &amp; Entertainment', 'Humor']                                  2
['Books', 'Biographies &amp; Memoirs', 'Arts &amp; Literature']                  2
['Books', 'Medical Books', 'Medicine']                                           2
['Books', 'Humor & Entertainment', 'Humor']                                      2
['Books', 'Literature & Fiction', 'United States']                               1
['Books', 'Literature &amp; Fiction', 'World Literature']                        1
['Books', 'Cookbooks, Food & Wine', 'Cooking by Ingredient']                     1
['Books', 'Cookbooks, Food &amp; Wine', 'Baking']                                1
['Books', 'Cookbooks, Food & Wine', 'Cooking Education & Reference']             1
```

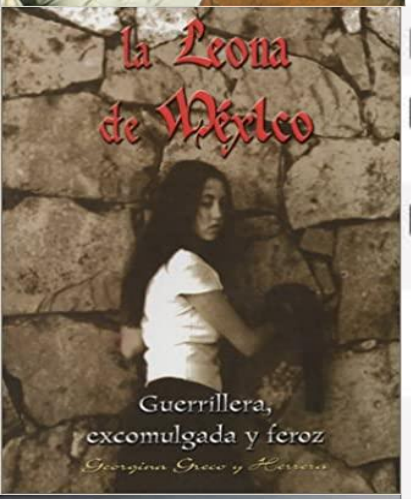
USER 134 RECOMMENDATIONS

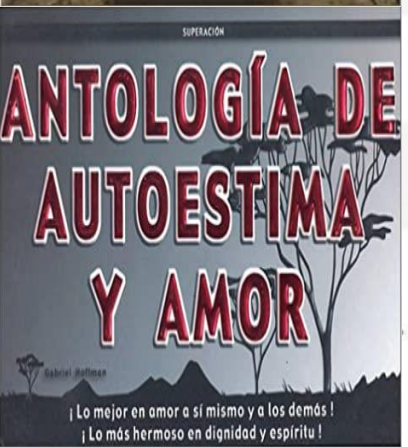
```
nrecommendations.join(books_meta_index_recommend, on='asinIndex').filter("reviewerIDIndex = '134'").toPandas()
```

	asinIndex	reviewerIDIndex	overall	asin	category	price	title
0	227642	134	11.518301	9687968257	[Books, Politics & Social Sciences]	$9.66	El Derecho Prohibido...(A Baby:Forbiden Right)...
1	188687	134	11.504252	9687968281	[]	$45.98	Antologa de Autoestima y Amor (The Best of Sel...
2	343863	134	11.708134	9686801693	[Books, Health, Fitness & Dieting]		El estres es Vida (Spanish Edition)
3	257191	134	11.477867	9686801707	[Books, Health, Fitness & Dieting, Women's...	$70.33	Libro de Oro del Embarazo (Spanish Edition)
4	240691	134	11.491187	9706061789	[Books, Health, Fitness & Dieting, Exercis...		Ejercicios Isomtricos (Isometric Exercises) (S...
5	245530	134	11.588929	9706061576	[Books, Cookbooks, Food & Wine, Cooking Educat...	.a-section.a-spacing-mini{margin-bottom:6px!im...	Cocina Mexicana de los siglos XVI al XIX (Mex...
6	227242	134	11.516125	1857910478	[Books, Reference, Dictionaries & Thesauruses]	$16.60	Focloir Poca: English-Irish Irish-English Dict...
7	274360	134	11.634775	9706061681	[]	$10.00	Dios Mio ! ¡Hazme Delgada!(Oh,Lord, Mak...
8	153367	134	11.535437	9706061908	[Books, Literature & Fiction, History &amp...	$29.95	La Leona de Mxico (Mexico`s Lioness) (Spanish ...
9	188082	134	11.539589	9686801278	[Books, Cookbooks, Food & Wine, Regional &...		Cocina para celebrar (Spanish Edition)

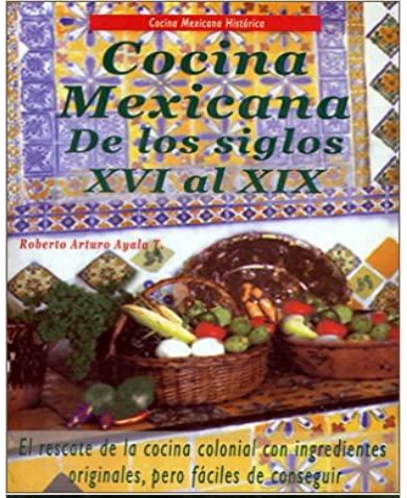
OVERLAPPING BETWEEN ACTUAL PREFERENCE & RECOMMENDATIONS?

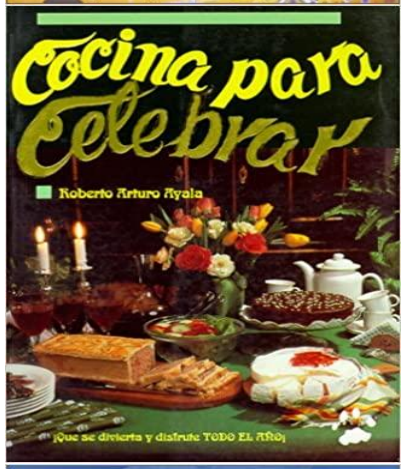


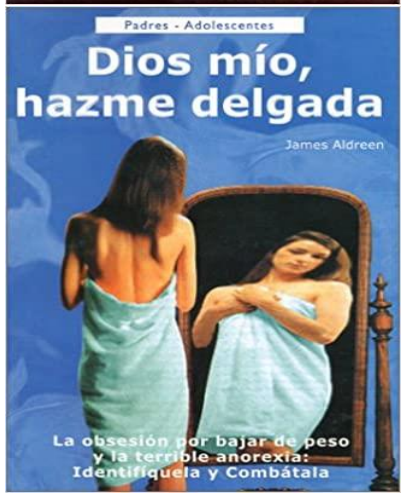




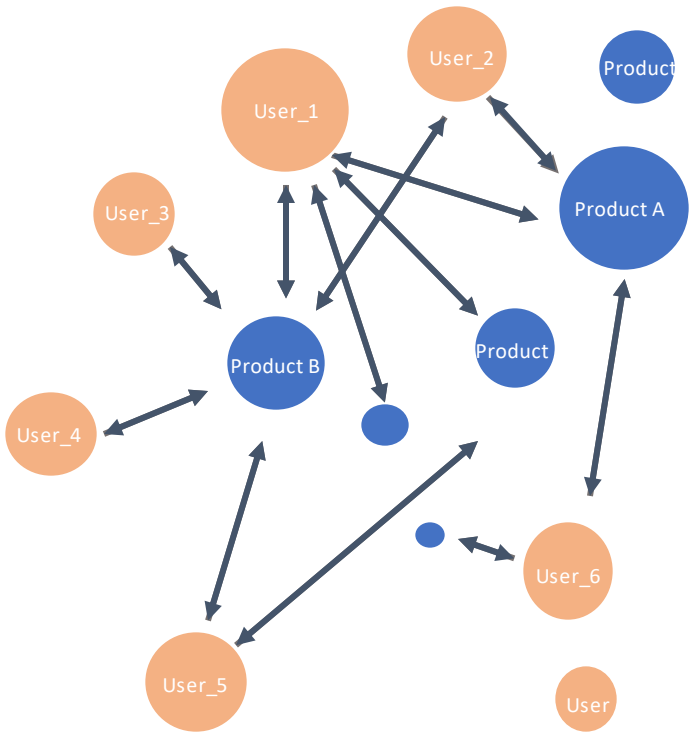
category	
[Books, Politics & Social Sciences]	'Literature & Fiction', 'Genre Fiction']
[]	'Mystery, Thriller & Suspense', 'Thrillers & S
	'Mystery, Thriller & Suspense', 'Mystery']
	'Science & Math', 'Biological Sciences']
[Books, Health, Fitness & Dieting]	'Cookbooks, Food & Wine', 'Regional & Internat
	'Literature & Fiction', 'Genre Fiction']
[Books, Health, Fitness & Dieting, Women's...	'Mystery, Thriller & Suspense', 'Mystery']
	'Literature & Fiction', 'United States']
[Books, Health, Fitness & Dieting, Exercis...	'Cookbooks, Food & Wine', 'Cooking Education &
	'Literature & Fiction', 'Contemporary']
[Books, Cookbooks, Food & Wine, Cooking Educat...	'Mystery, Thriller & Suspense', 'Thrillers & Suspense'
	'Humor & Entertainment', 'Humor']
[Books, Reference, Dictionaries & Thesauruses]	'Biographies & Memoirs', 'Arts & Literature']
	'Medical Books', 'Medicine']
	'Humor & Entertainment', 'Humor']
	'Literature & Fiction', 'United States']
	'Literature & Fiction', 'World Literature']
[]	'Cookbooks, Food & Wine', 'Cooking by Ingredient']
	'Cookbooks, Food & Wine', 'Baking']
[Books, Literature & Fiction, History &amp...	'Cookbooks, Food & Wine', 'Cooking Education & Referen
[Books, Cookbooks, Food & Wine, Regional &...	







Graph



Building the graph – GraphFrame

- Vertices
  - Products (ProductID…)
  - Users (ReviewerID, ReviewerName…)
- Edges
  - Review (ReviewTime, ReviewText…)

Directions

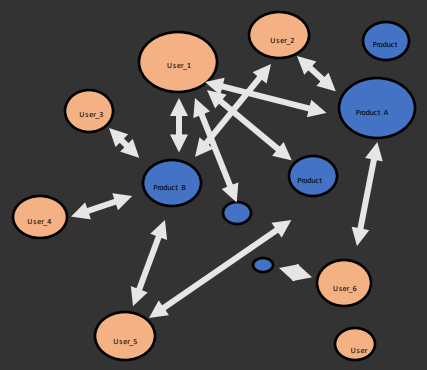
- Single direction

  (User_1)-[reviewed]->(Product A);
  (User_2)-[reviewed]->(Product A);

- Bi-direction

  (User_1)-[reviewed]->(Product A);
  (Product A)-[reviewed]->(User_2);

Graph



g.vertices .filtering('value')	g.edges.groupBy()	g.bfs(id1, id2, maxPath)
g.edges .filtering('value')	g.find(route)	g.connectedComponents
g.degrees •inDegrees •outDegrees	g.triangleCount()	g.pageRank

```
graph.bfs("id = 'B003VWJ2K8'" , "id = 'B003VWKPHC'", maxPathLength = 10).show(truncate = False)
```

```
|[B003VWJ2K8,]|[AE9C0UNXBV8CB, B003VWJ2K8, 5.0, [0, 0], 03 11, 2014, I own several of the Snark tuners.  The SN-1 s
ettles quickly, and works with a wide array of instruments.  Tuning couldn't be simpler.  Just clip it to your inst
rument and strum or play a note and, in an instant, there it is.  The display is bright and easy to read, and the g
raph has a lot of resolution.  The SN-1 senses frequency through vibration.  It is transposable, and it offers a me
tronome feature.  The SN-1 is finished with a rubbery texture and is a beautiful &#34;electric blue&#34; color., B0
03VWJ2K8, AE9C0UNXBV8CB]
|[AE9C0UNXBV8CB, Mike Lovelace]      |[AE9C0UNXBV8CB, B003VWKPHC, 5.0, [0, 0], 03 11, 2014, I own several of the Sn
ark tuners, but the Snark SN-2 is my personal favorite.  It settles extremely fast, and works with a wide array of
instruments and couldn't be simpler.  Just clip it to your instrument and strum or play a note and, in an instant,
there it is.  The display is bright and easy to read, and the graph has a lot of resolution.  The SN-2 senses frequ
ency through vibration, but it also has a selectable microphone pickup.  It is transposable, and it offers a metron
ome feature.  The SN-2 is finished with a rubbery texture and is a beautiful &#34;electric red&#34; color., AE9C0UN
XBV8CB, B003VWKPHC]
|[B003VWKPHC,]|
|[B003VWJ2K8,]|[A2Y7BSQG9V3LNG, B003VWJ2K8, 5.0, [0, 0], 03 10, 2014, I used these on all my instruments, guitar, m
andolin, and violin.  It works fine and I would recommend them at this good price., B003VWJ2K8, A2Y7BSQG9V3LNG]
|[A2Y7BSQG9V3LNG, James M. Bailey]   |[A2Y7BSQG9V3LNG, B003VWKPHC, 5.0, [0, 0], 05 31, 2011, I highly recommend the
se tuners.  They are the best I have found for Fiddle and Mandolin.  They also work well with acoustical guitar., A
2Y7BSQG9V3LNG, B003VWKPHC]
```

```
graph.triangleCount().show()
```

count	id	reviewerName
0	A17A1KTVI3DG6U	Nathan A. Edwards
0	A2DG65AWX5RJ4J	Carlos
0	A2IZ3ST24HSO4H	David McCarthy
0	A36C867ZDP30NQ	John D
0	A5MC7LP0ZBO4Q	Lets.Be.Reasonable.
0	B000MWWT6E	null
0	B00	
0	B00	

EW	8
1G	8
HK	20
6U	5
NQ	6

```
# Products has the most reviews
(graph.edges
    .groupBy("src")
    .count()
    .orderBy("count", ascending = False)
    .withColumnRenamed('src', 'productID')
    .show())
```

productID	count
B003VWJ2K8	163
B0002E1G5C	143
B0002F7K7Y	116
B003VWKPHC	114
B0002H0A3S	93
B0002CZVXM	74

Graph - takeaways

Why

Unstructured data

Complex structure

Space saving by index

	Structured	Graph
# of rows	2431	210

Challenges

Data attributes limitation

Different node structures

Platform limitation

How

Prepare 2 datasets

'GraphFrame' – generating graph

Querying

Visualization: Neo4j on Databricks

Conclusion and Future Work

Achievements

- ✓ Dealing with Extremely large data ~90GB
- ✓ SparkML on NLP
  - MNB
  - LR
  - RF
- ✓ Recommendation System
- ✓ Graph

Experience

- System crash
- Switching between platforms (GCP, RCC, Databricks...)
- Building pipelines/ experiments

Limitations

- ❑ Dataset attributes
- ❑ Computation Power
- ❑ Spark supported packages
  - Lib factorization machine, stochastic SVD

Improvements

- ➢ ML models
  - Cross validation
- ➢ Recommendation System
  - Explore different methods other than ALS or more regularization parameters
  - Evaluation methods
- ➢ Graph – API of neo4j
- ➢ Scale up!

Thank you!