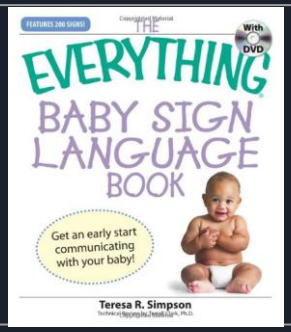
Hybrid Scalable Online Recommendations

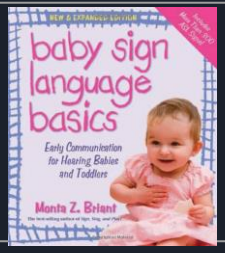
Diego Fanesi
Farooq Qaiser
Sai Chaitanya

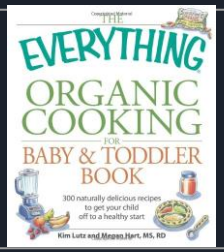
Amazon

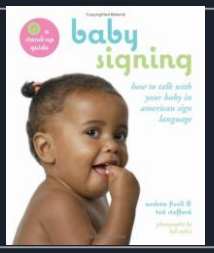


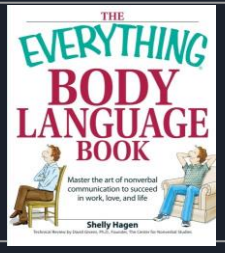
Title: The Everything Baby Sign Language Book
Price: $9.99

Similar items to what you're viewing

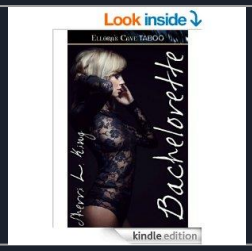


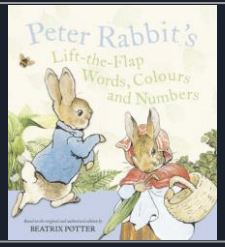


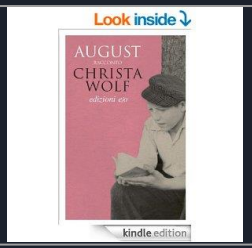


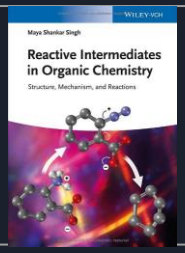


Other items you might like









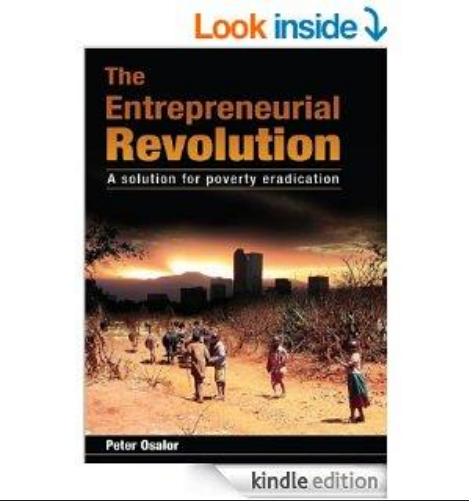
Problem	Experiments	Final Solution
Holistic recommendations	1. Content Based Filtering 2. Collaborative Filtering	Hybrid model
Scalable implementation	1. JSON vs Parquet 2. Local vs Cluster	1. Amazon S3 + Parquet 2. Amazon EC2 + Spark
Online capabilities	1. N^2 approach 2. LSH approach 3. ALS approach 4. SSGD approach	1. Content Based Filtering using LSH 2. Collaborative Filtering using SSGD

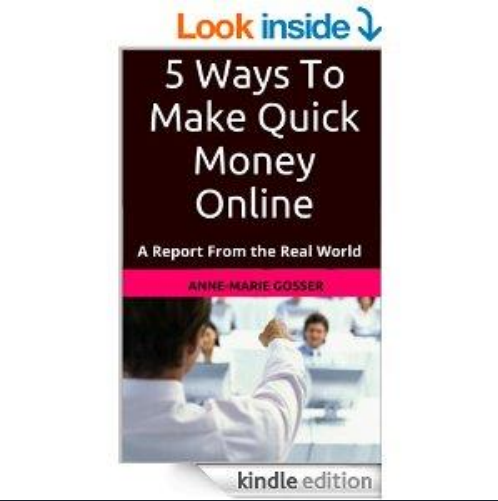
Content Based Recommender

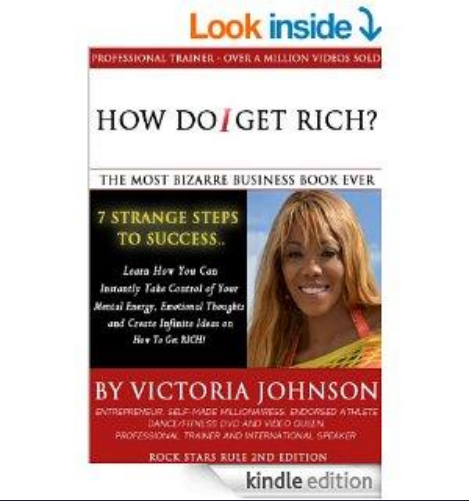
How it works

Selected Recommendations

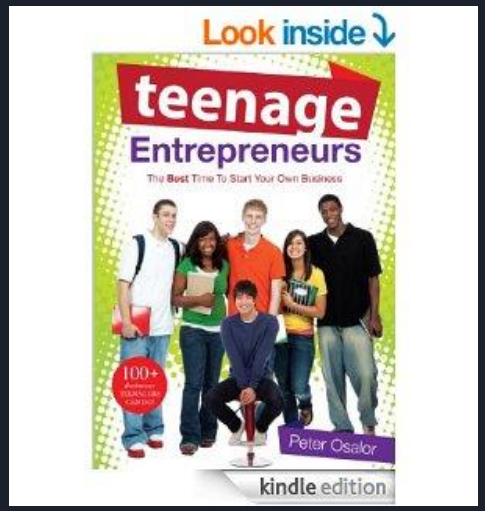
Item

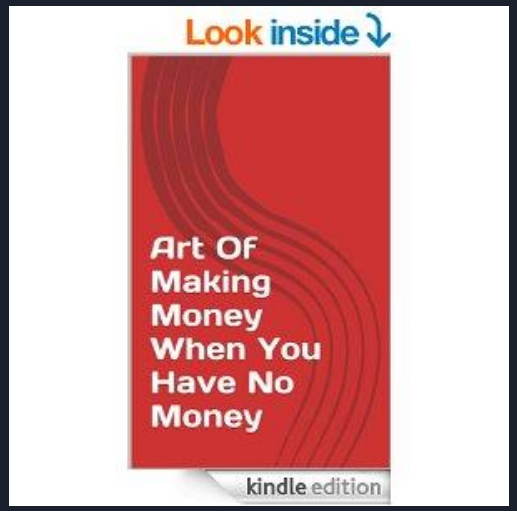


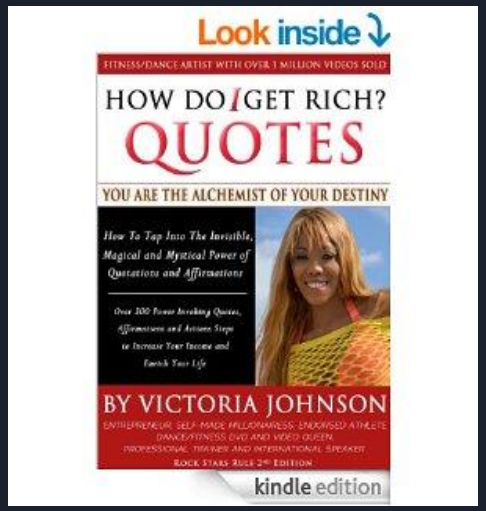


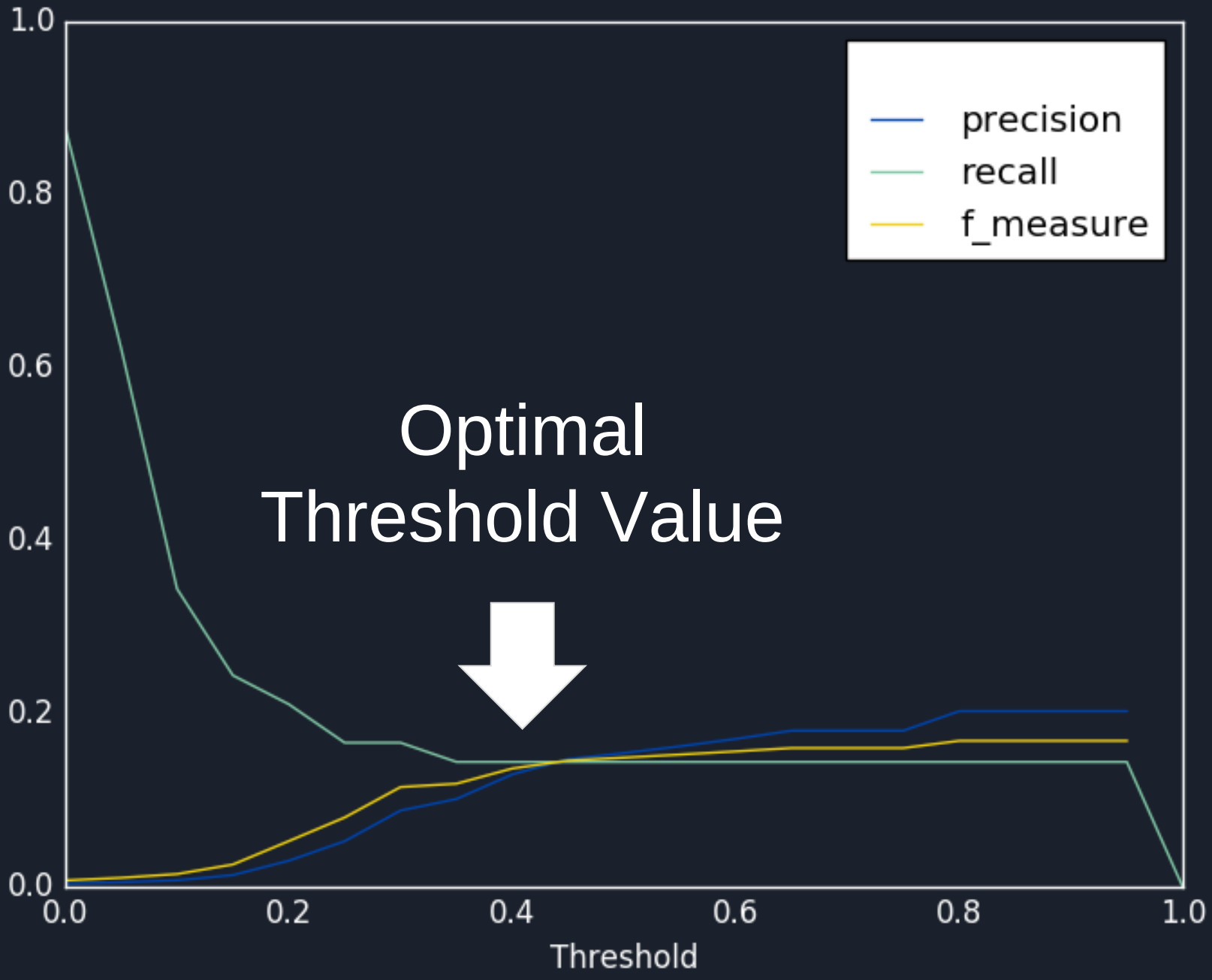


Similar Item

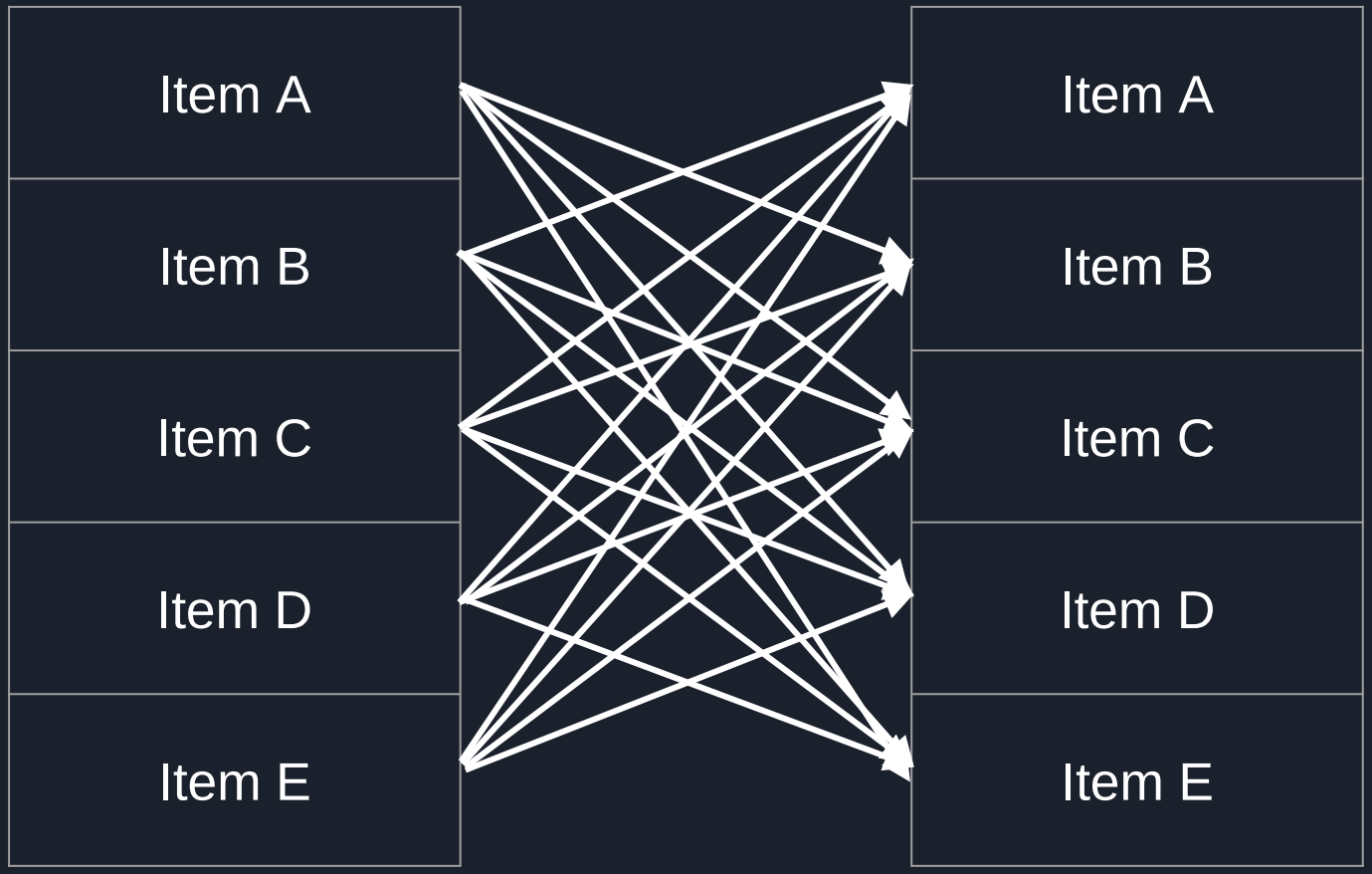




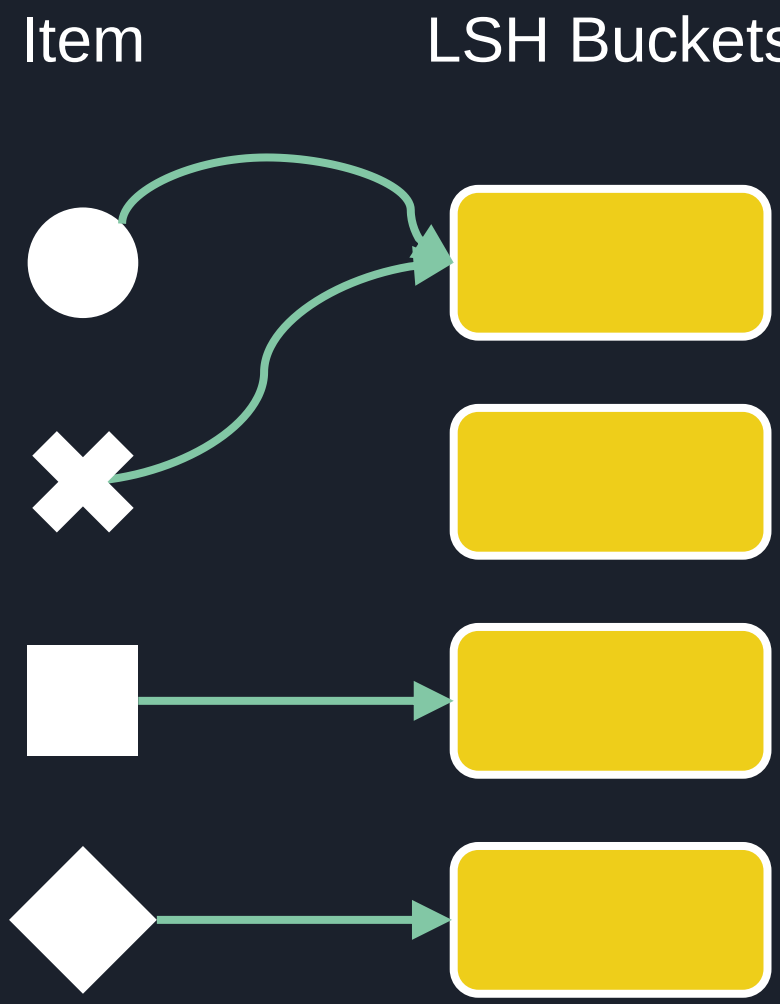


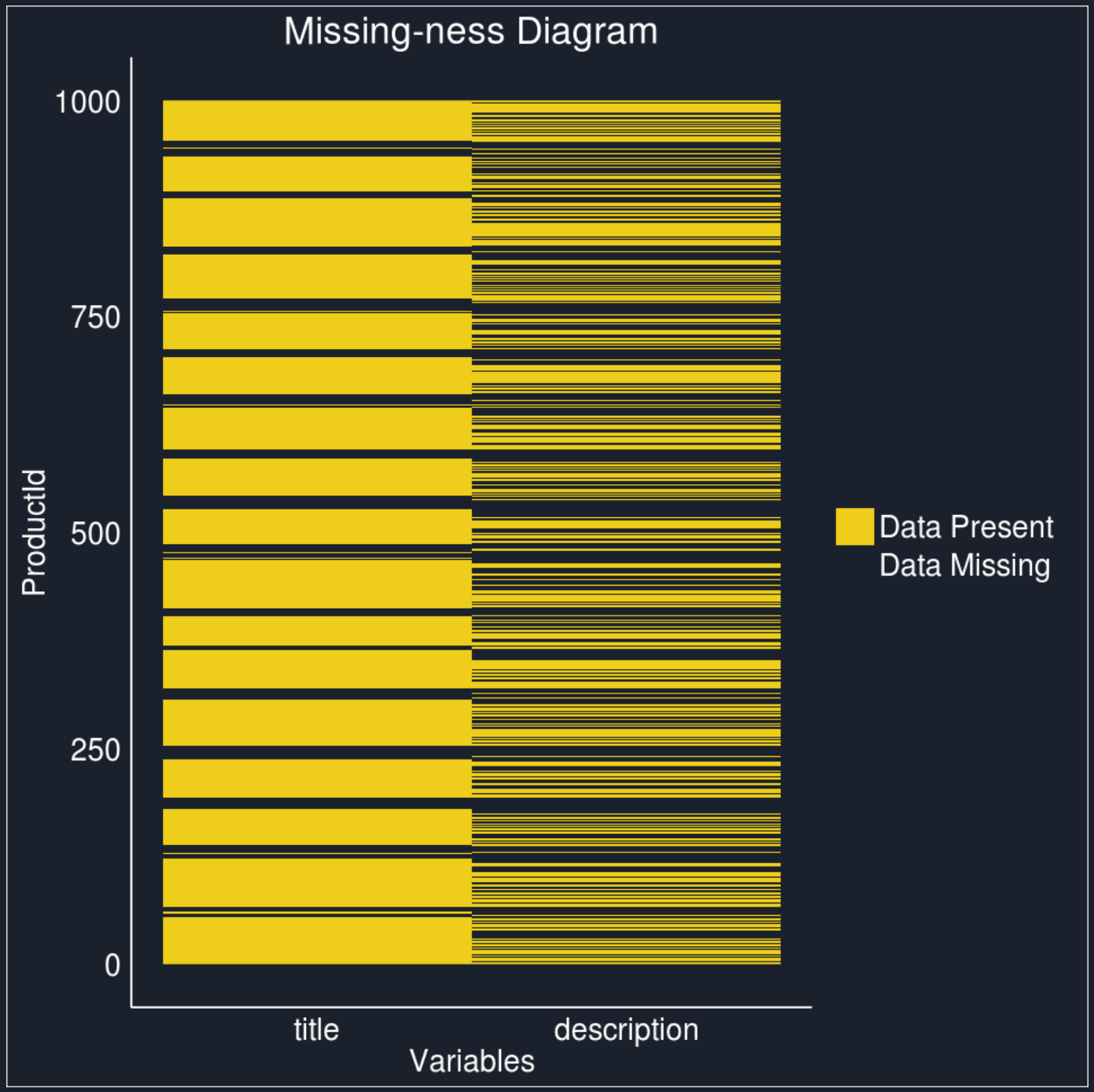

1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Threshold
precision
recall
f_measure
Optimal
Threshold Value

N^2 Approach
How it works



Locality Sensitive Hashing Approach
How it works



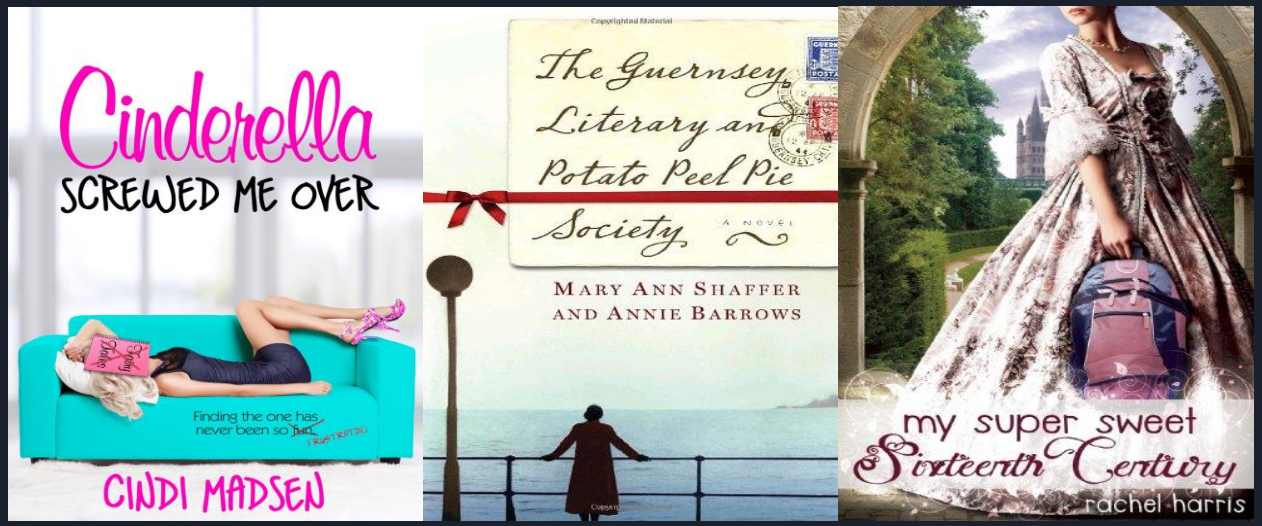

Missing-ness Diagram
1000
750
500
250
0
ProductId
title
description
Variables
Data Present
Data Missing

Collaborative Filtering Recommender

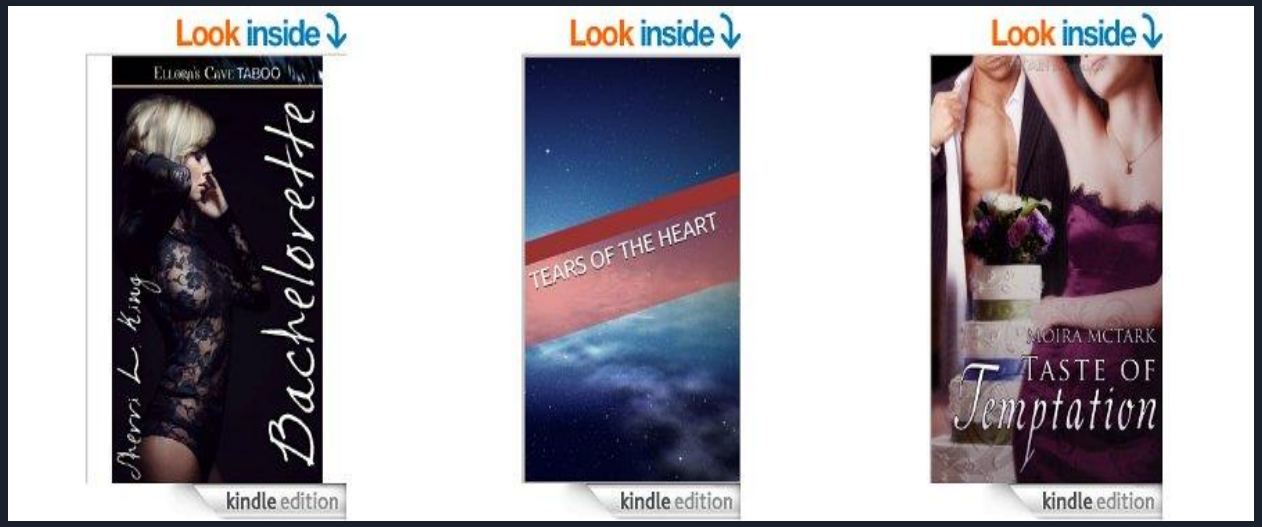
Collaborative Filtering
How it works

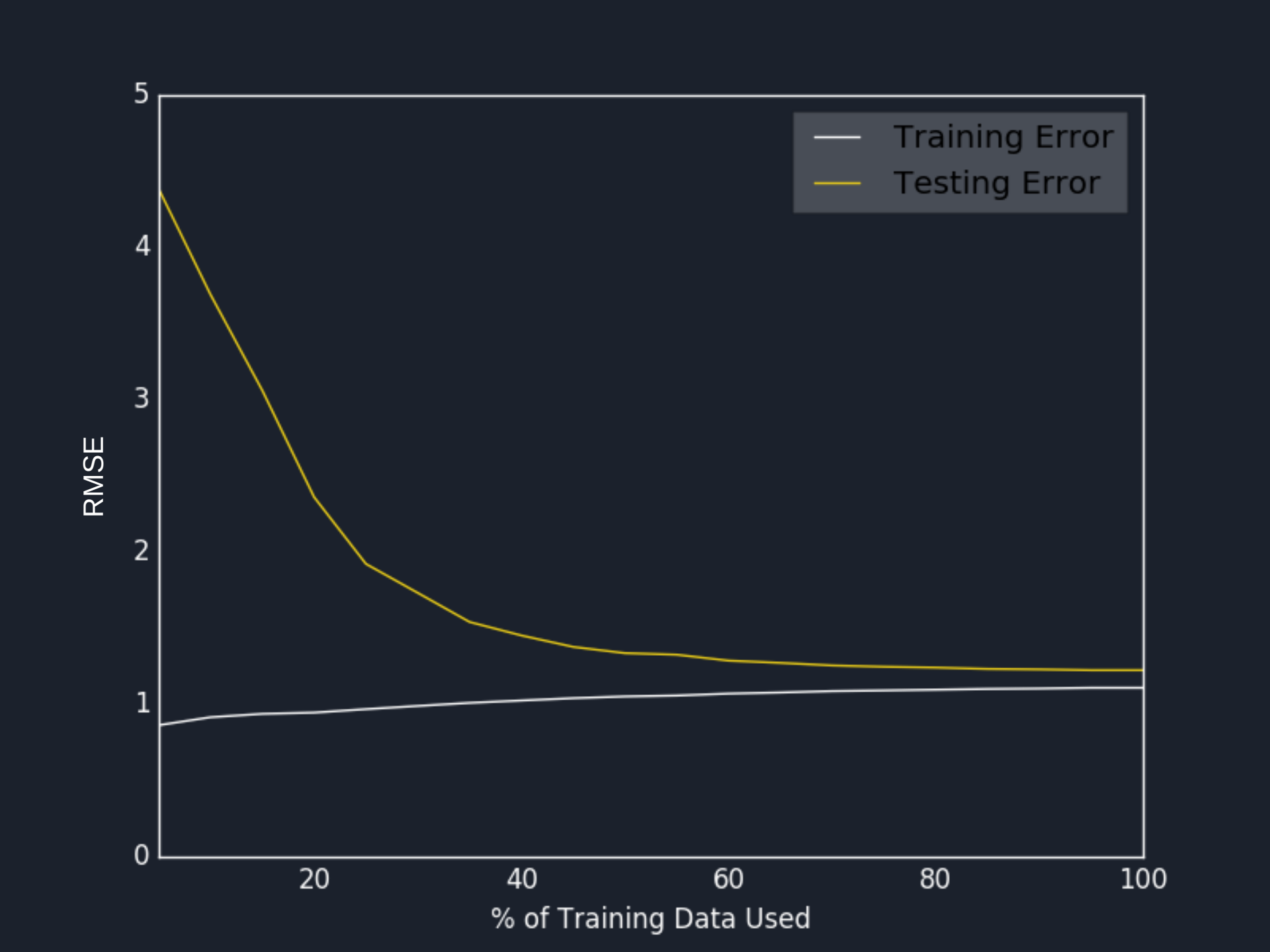
Selected Results

Items rated highly by the user

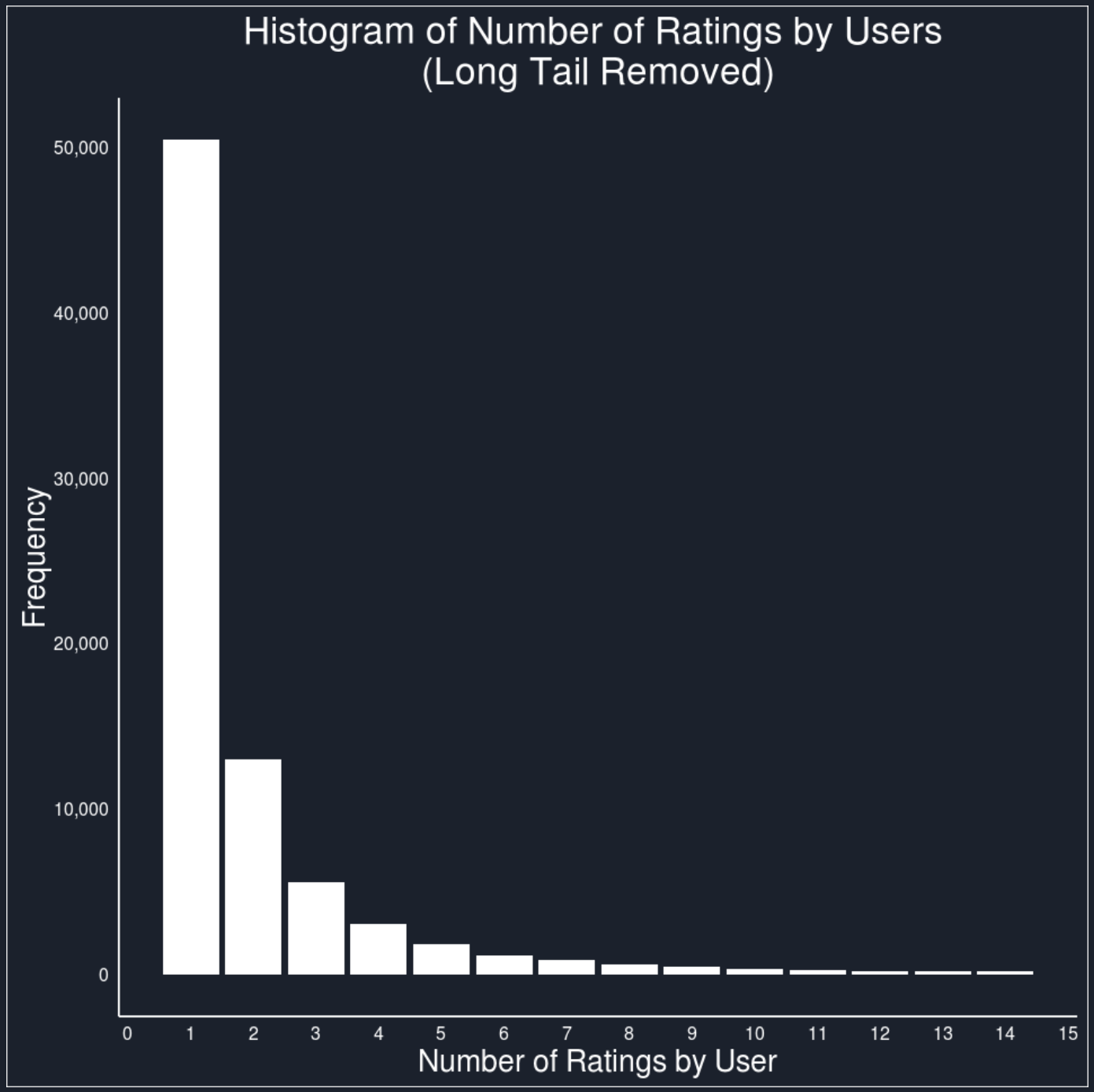


Predictions from the Collaborative Filtering model




Training Error
Testing Error
RMSE
% of Training Data Used
0
1
2
3
4
5
20
40
60
80
100

ALS	SSGD
Alternating Least Squares	Stratified and Distributed Stochastic Gradient Descent
Spark ML and MLLib	Custom Implementation


Histogram of Number of Ratings by Users
(Long Tail Removed)
Frequency
50,000
40,000
30,000
20,000
10,000
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Number of Ratings by User

Conclusions And Productionizing

Amazon

Title
Description
Price

Similar Items to what you're viewing

Content Based Filtering Recommender

Other items you might like

Collaborative Filtering Recommender

Conclusions

1. LSH > N2 for content based approach
2. SGD > ALS for collaborative filtering approach
3. Reviews > Ratings
4. CB is great as a content similarity tool
5. CF great as a content discovery tool

From PoC to Production

- Rewrite in scala to improve performances
- Improve the implementation of algorithms
- Implement Spark Streaming for online prediction and training
- Add autotests to make it more robust
- Tune up the models
- Semantic understanding for Content Based
- ngrams for Content Based

We would like to thank Julian McAuley for graciously sharing the full Amazon Books datasets that were used for this project.

Appendix

Supplemental Material

Storage

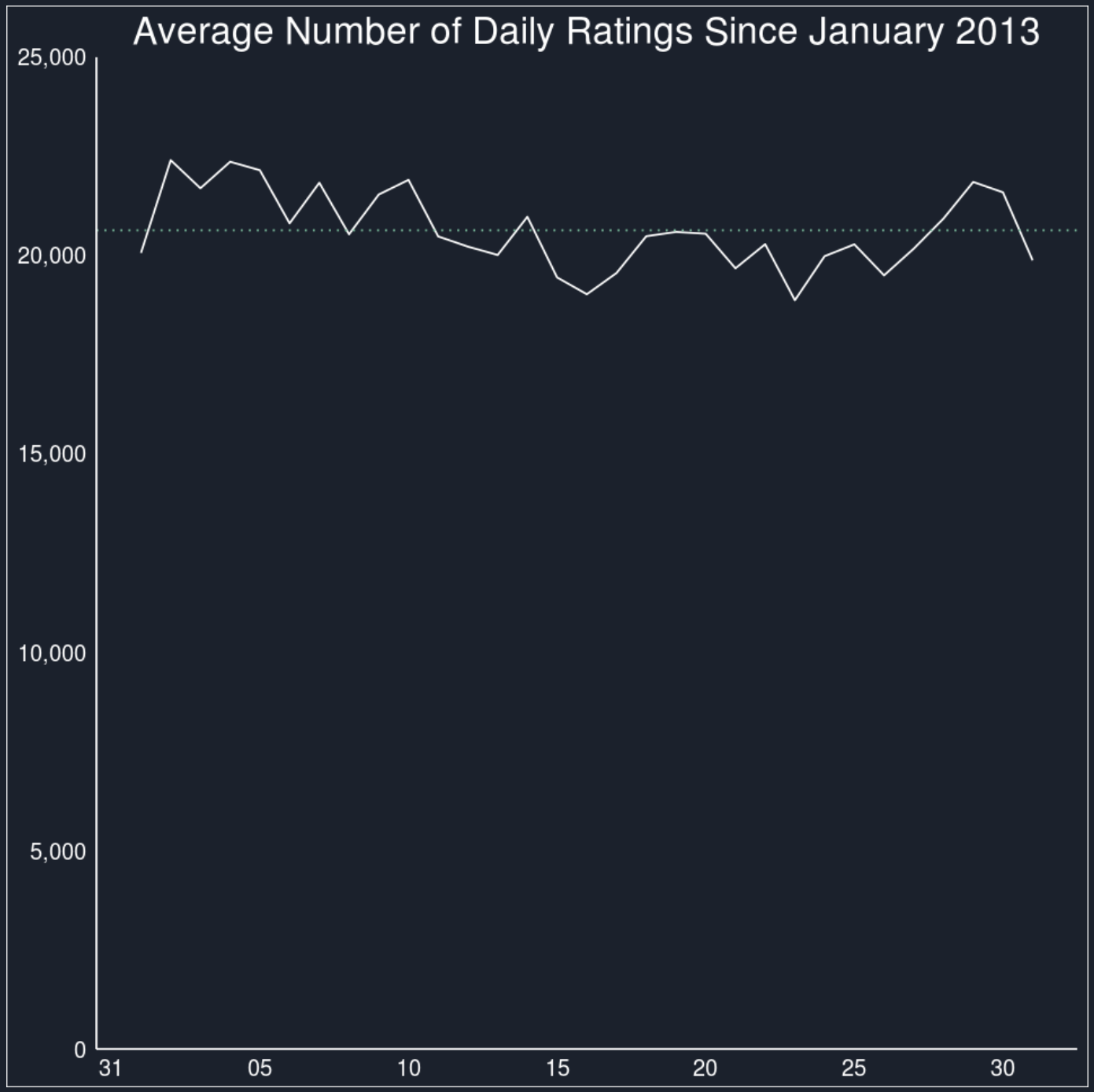
Computation

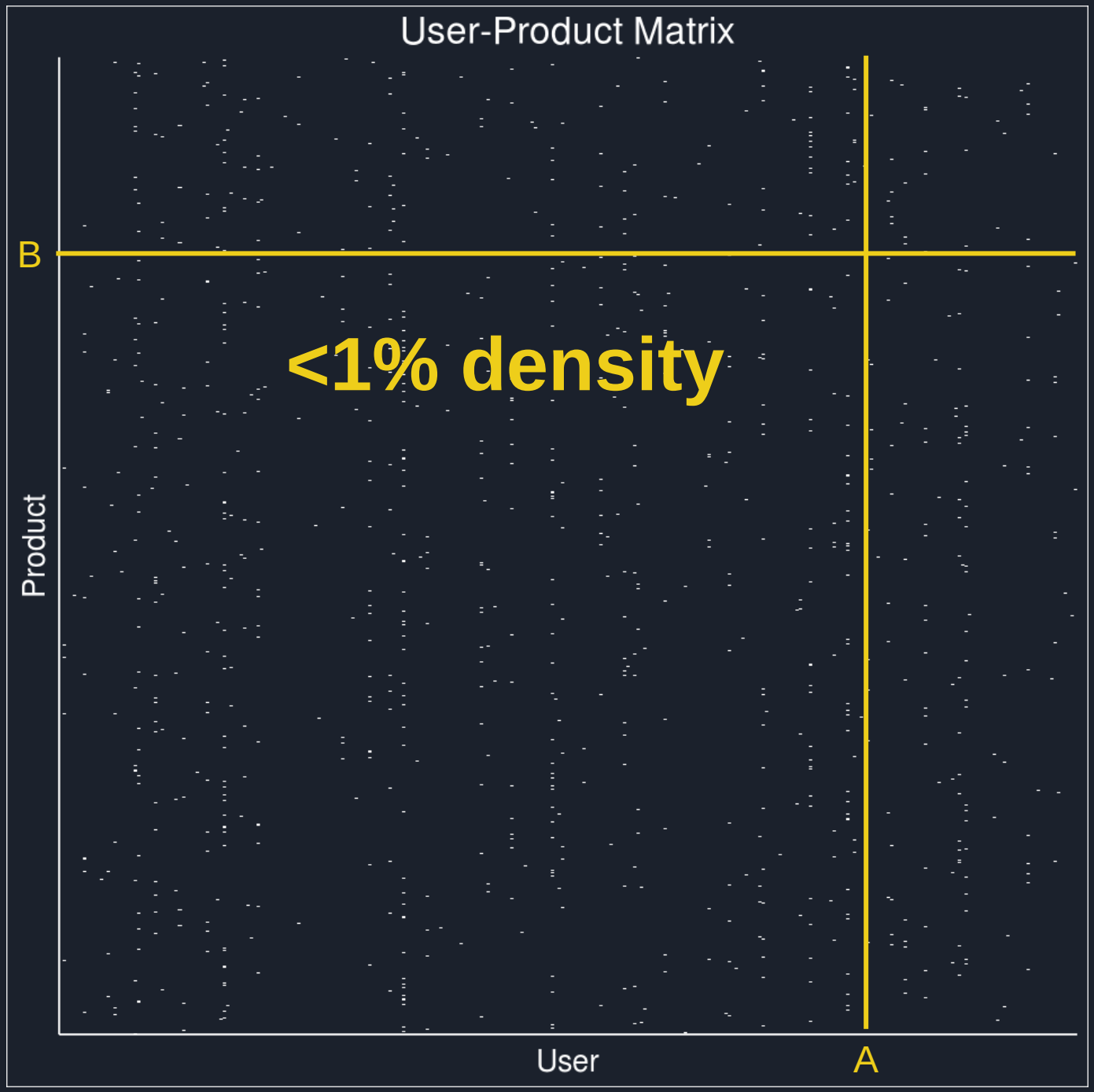
Exploratory Data Analysis

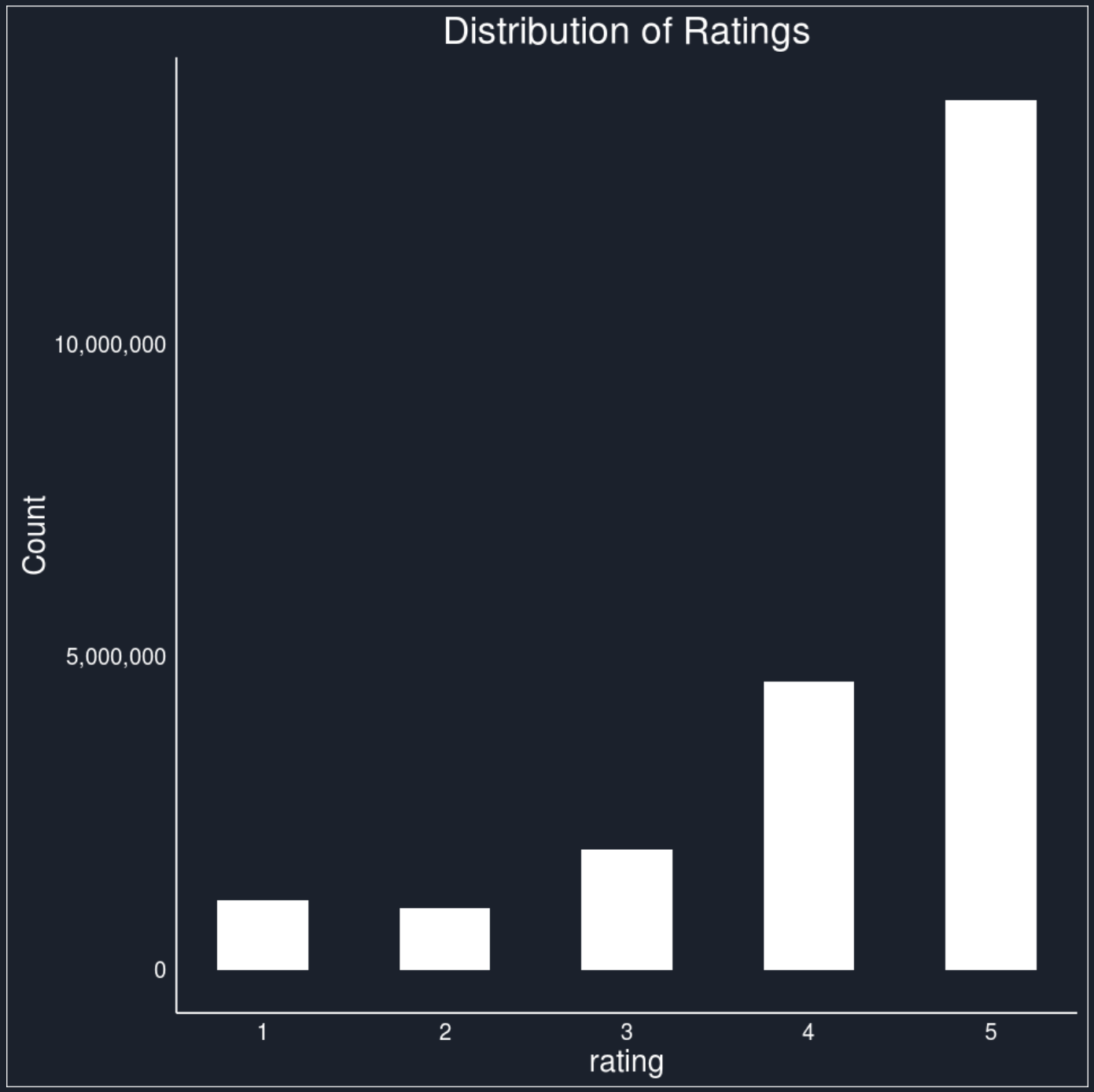
Supplemental Material

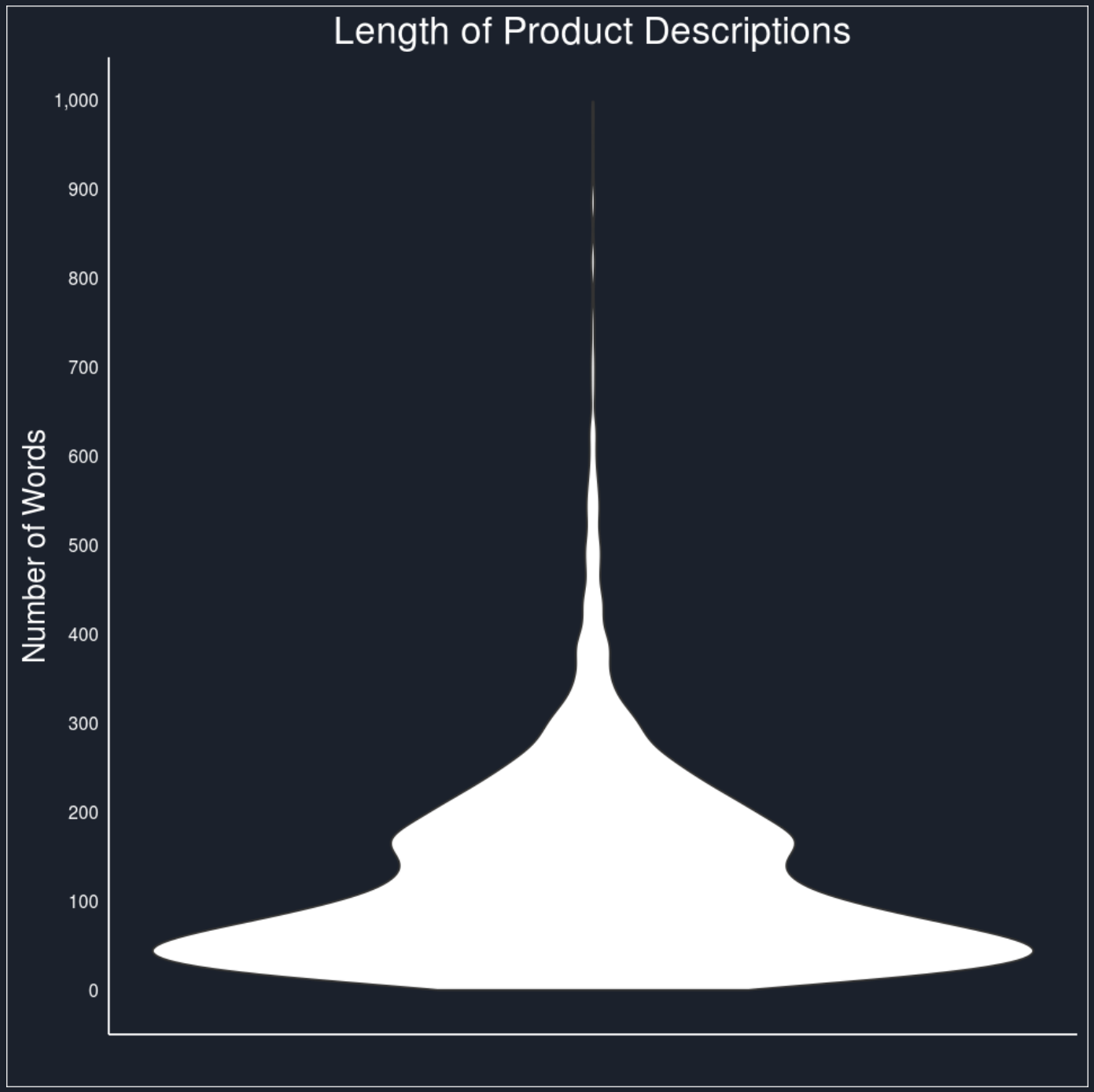
Books dataset

Users	8,026,459
Books	3,941,625
Ratings	22,507,206
Reviews	8,899,474


Average Number of Daily Ratings Since January 2013
25,000
20,000
15,000
10,000
5,000
0
31
05
10
15
20
25
30


User-Product Matrix
B
<1% density
Product
User
A


Distribution of Ratings
Count
10,000,000
5,000,000
0
1
2
3
4
5
rating

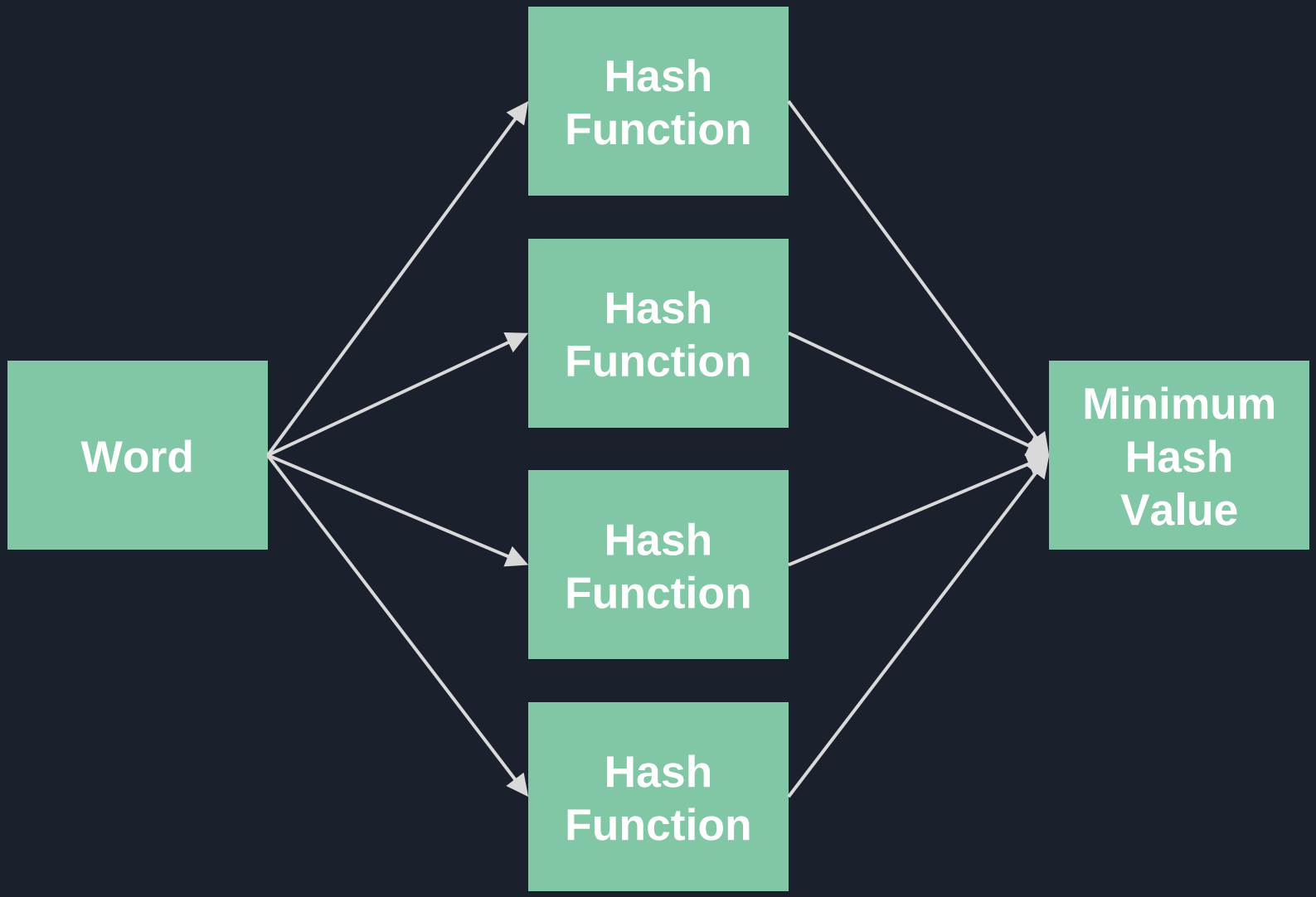

Length of Product Descriptions
Number of Words
1,000
900
800
700
600
500
400
300
200
100
0

Content Based Recommender

Supplemental Material


MinHash
LSH

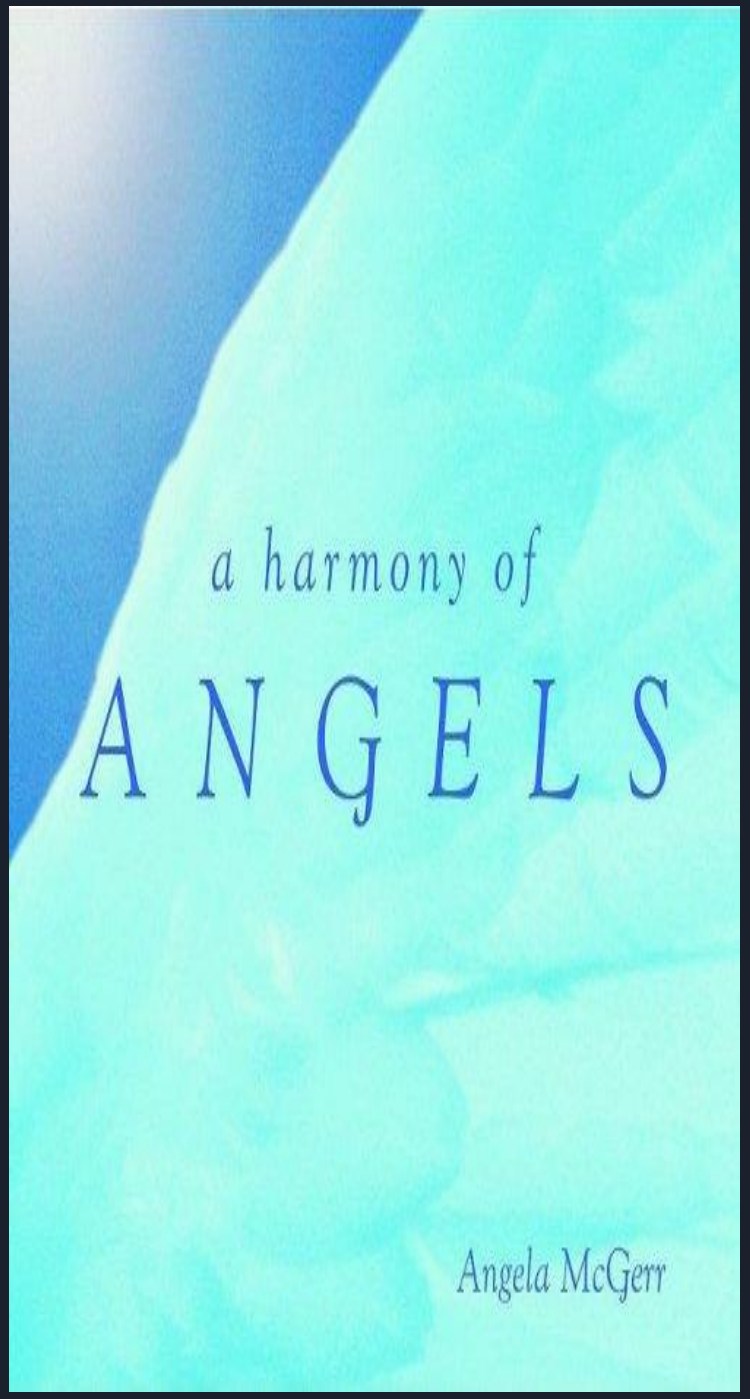
MinHash Function



Locality Sensitive Hashing

Item	Bucket 1	Bucket 2	Bucket 3	Bucket 4	Bucket 5
A	0	1	2	3	3
B	5	3	2	1	1
C	6	9	11	13	13
D	5	3	2	1	1
E	30	0	2	1	1

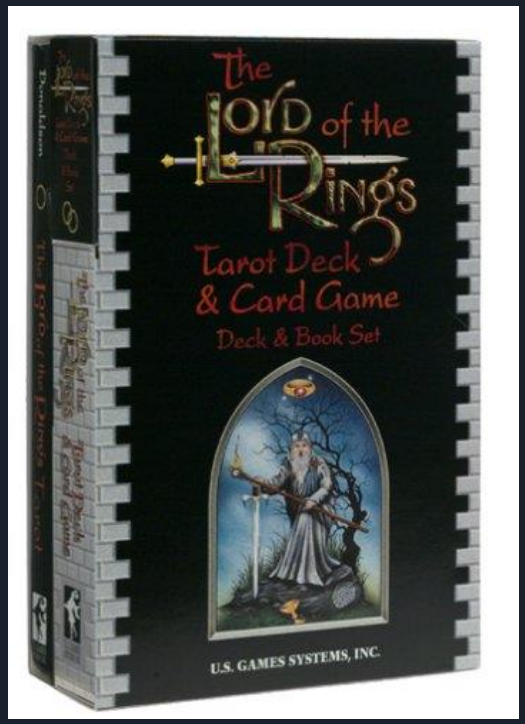
Product ID	682117
Title	A Harmony of Angels
Product Description	"This is the loveliest book on angel work that I have ever seen." (Reader comment) "This breathtakingly beautiful book and card set brings you so many uses." (Reader comment)" I recommend it for any beginner on angels." (Reader comment)
Price	None

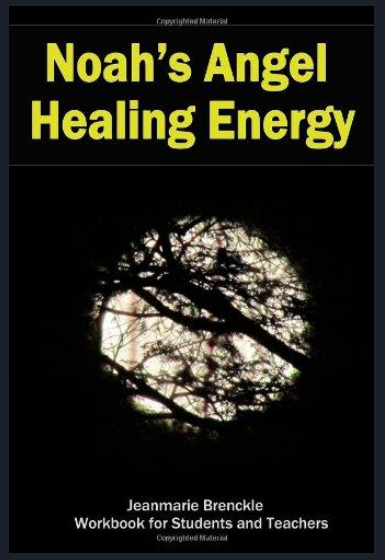


Official NFL Pro Set Card Book

National Football League

Note: This is not the actual book cover



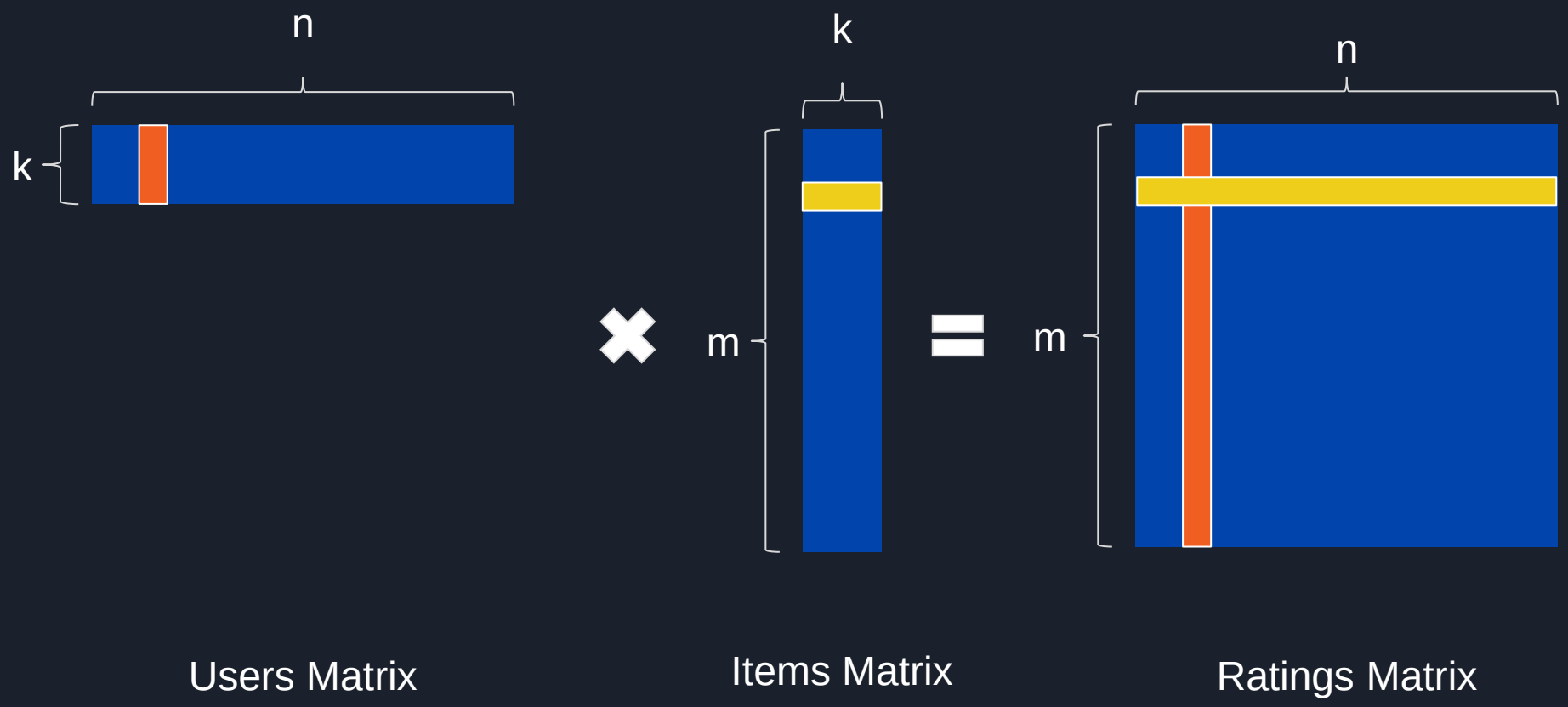


N^2	LSH
Slow	Fast
Accurate	Approximate
Scales very poorly	Scales extremely well

Collaborative Filtering Recommender

Supplemental Material

Low Rank Matrix Factorization
How it works

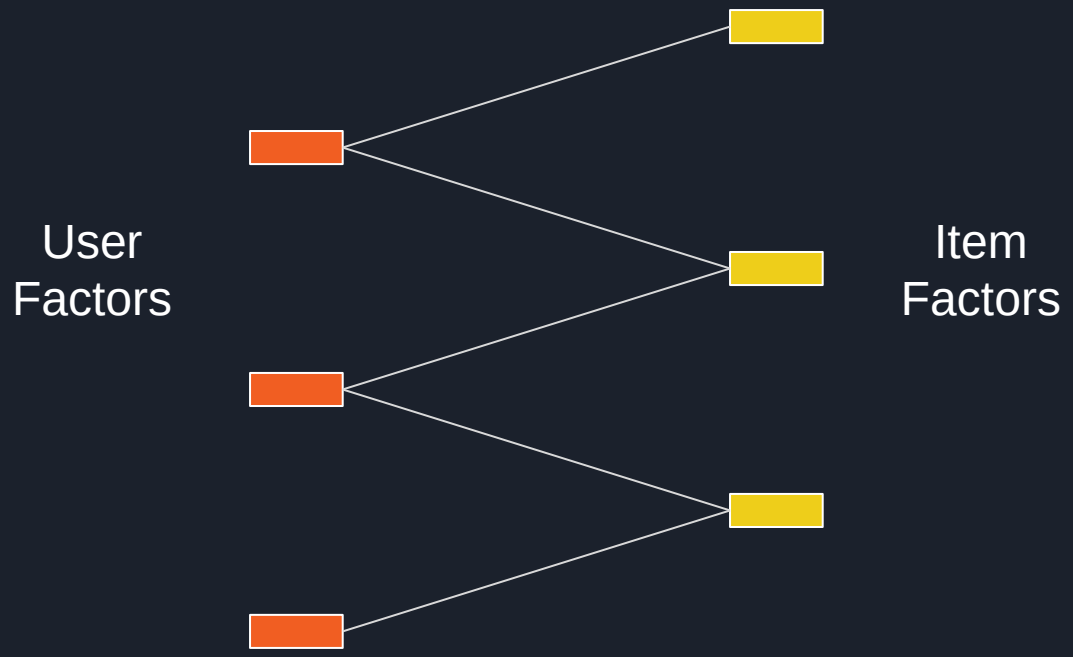


Users Matrix

Items Matrix

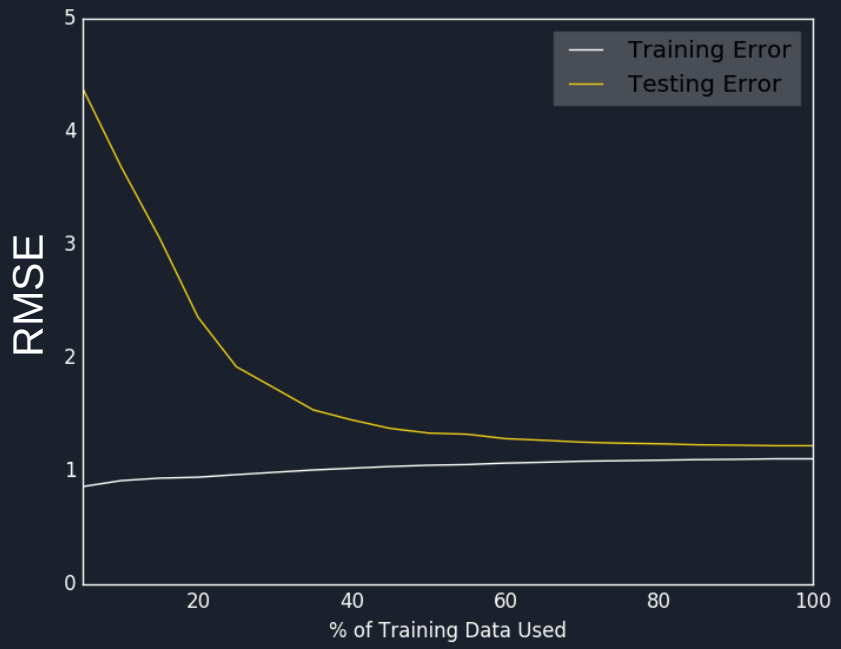
Ratings Matrix

Alternating Least Squares
How it works

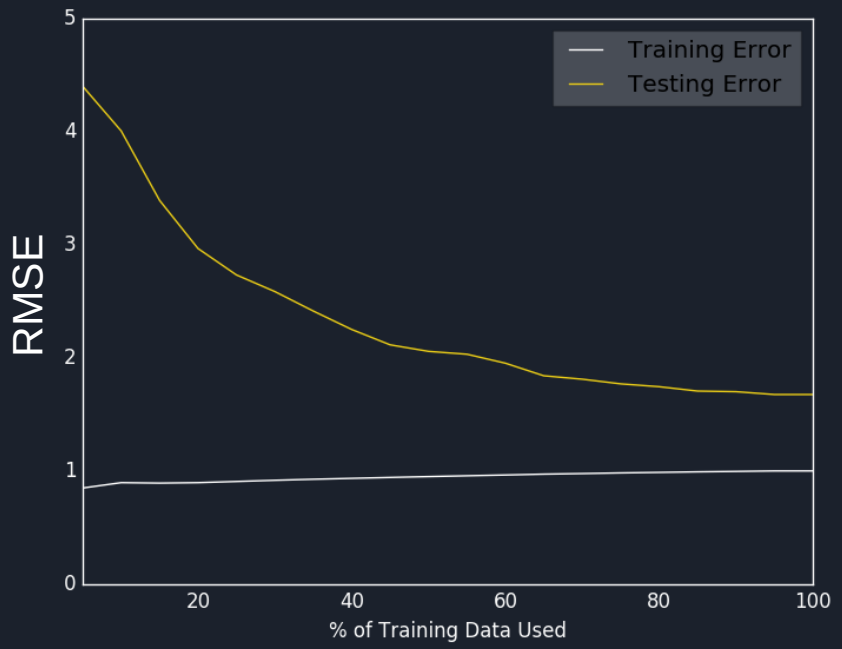


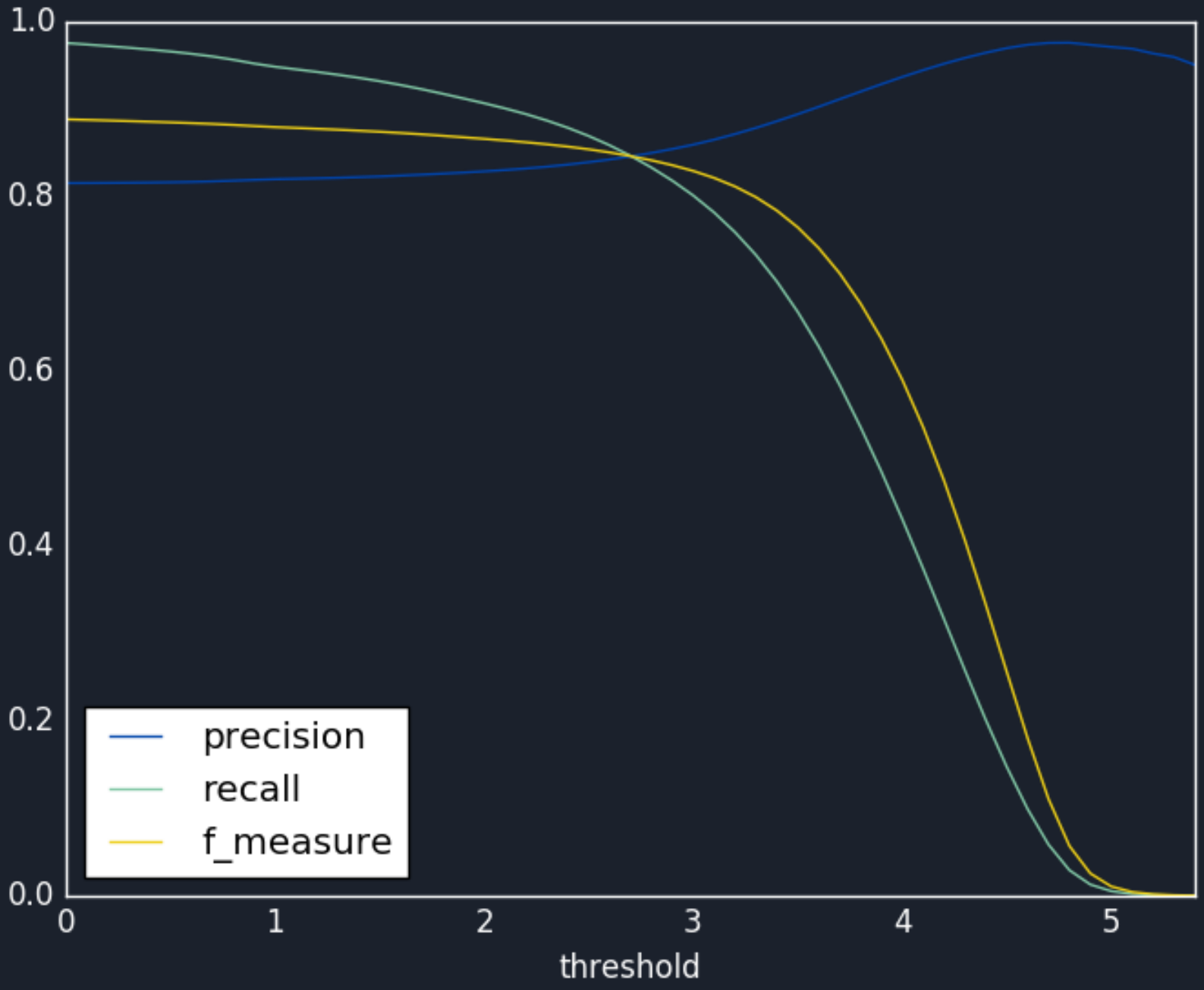
Collaborative Filtering Learning Curves

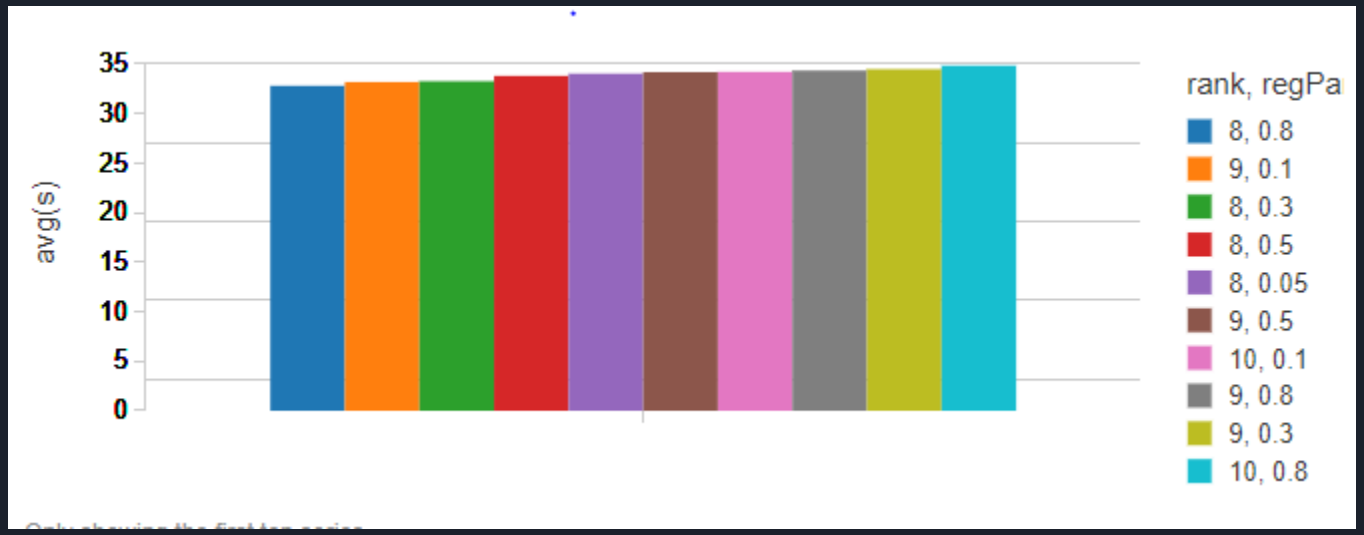
Reviews Only

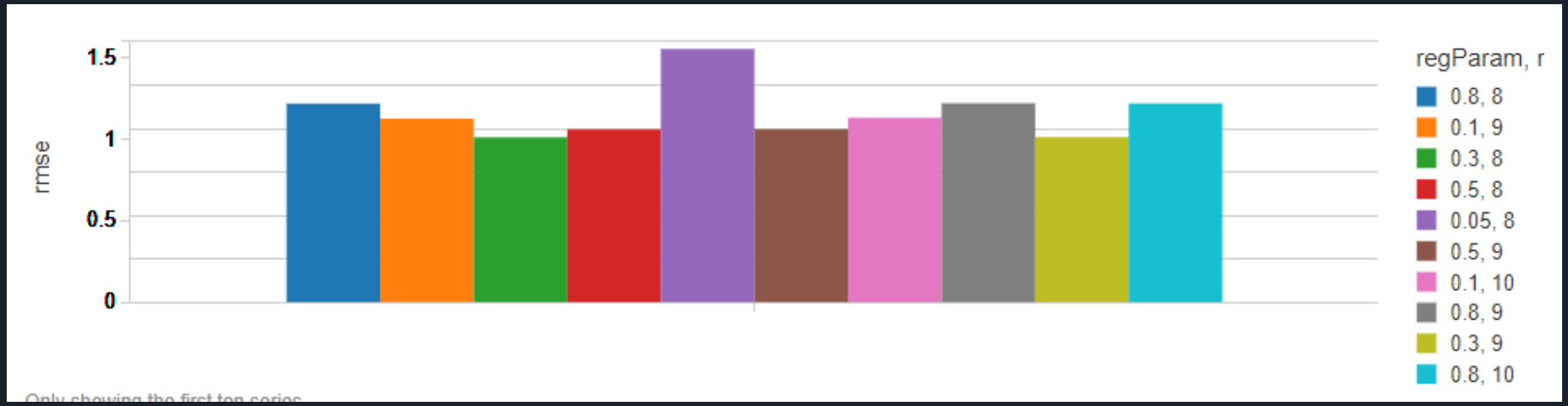


Ratings Only




1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
threshold
precision
recall
f_measure


35
30
25
20
15
10
5
0
avg(s)
rank, regPa
8, 0.8
9, 0.1
8, 0.3
8, 0.5
8, 0.05
9, 0.5
10, 0.1
9, 0.8
9, 0.3
10, 0.8


1.5
1
0.5
0
rmse
regParam, r
0.8, 8
0.1, 9
0.3, 8
0.5, 8
0.05, 8
0.5, 9
0.1, 10
0.8, 9
0.3, 9
0.8, 10

Stochastic Gradient Descent

Arberger (2009) shows that SGD is generally faster and more accurate than ALS except in situations of extremely sparse data where ALS tends to performs better.

Stratified Stochastic Gradient Descent
How it works