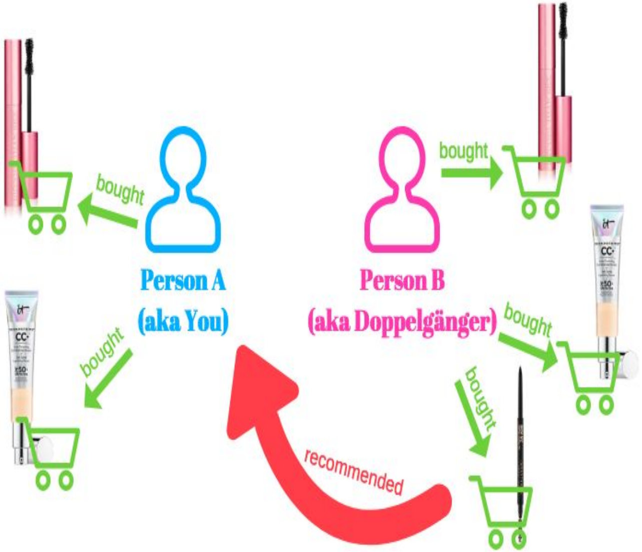
Capstone Project
Product Recommendation Engine

Content

Problem Statement

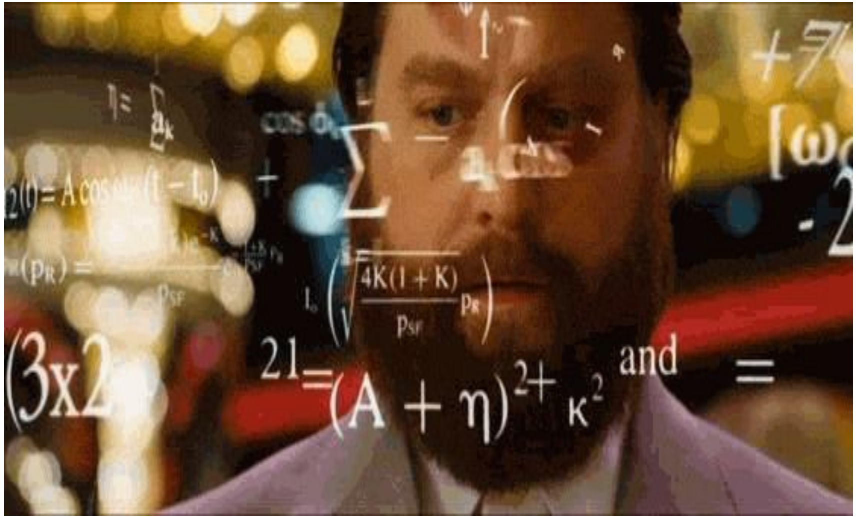
- Many online businesses rely on customer reviews and ratings. Explicit feedback is especially important in the ecommerce industry where all customer engagements are impacted by these ratings. Amazon relies on such rating data to power its recommendation engine to provide the best beauty product recommendations that are personalized and most relevant to the user.
- Build a recommender engine that reviews the customer ratings and purchase history to recommend items and improve sales for beauty products.



Data Summary

- reviewerID - UserId
- asin - ProductId
- reviewer Name - User Name
- helpful
- review Text
- overall - Rating
- summary
- unixReviewTime
- review Time

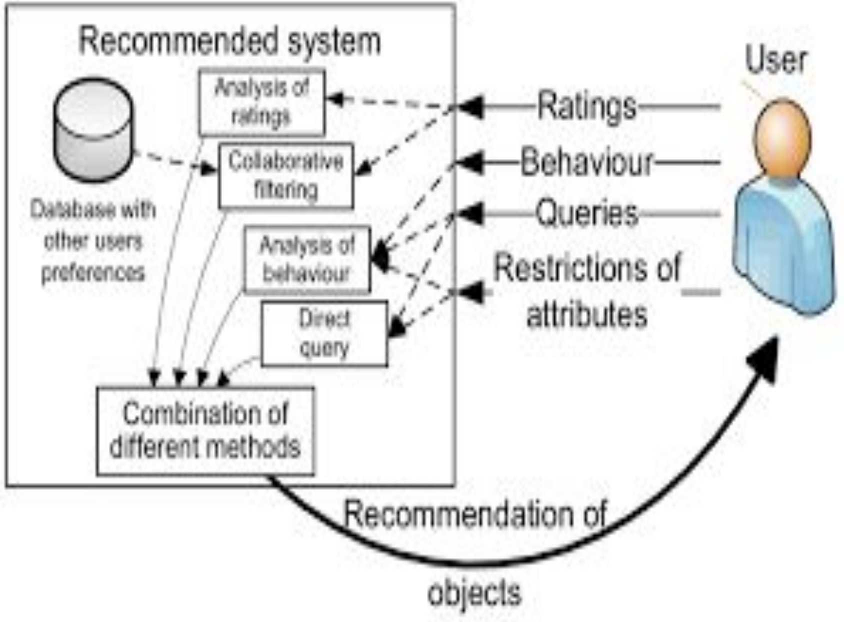
Total Row - 198502
Total columns - 9



Recommended Systems

- Sharp system that provides idea about item to users that might interest them.

- Recommendation system is subclass of information filtering to predict preferences to the items used by or for users. It personalize recommendation and deals with information overload. These demands throws some challenges so different approaches like memory based, model based are used.



Product Recommendation

Need of recommendation systems:

Why there is a need?

"Getting Information off the internet is like taking a drink from a fire hydrant" - Mitchell Kapor

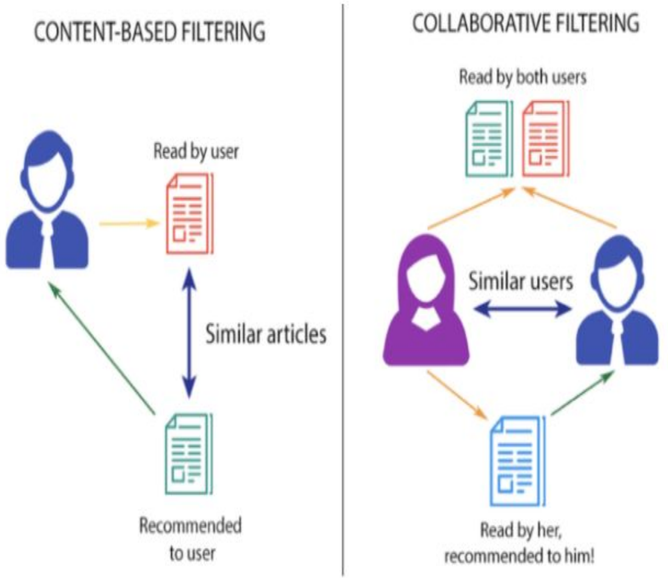
- Information Overload
- User Experience
- Revenues

Recommender systems help in addressing the information overload problem by retrieving the information desired by the user based on his or similar users' preferences and interests.

Approaches of Recommendation System

Recommendation system is usually classified on rating estimation:

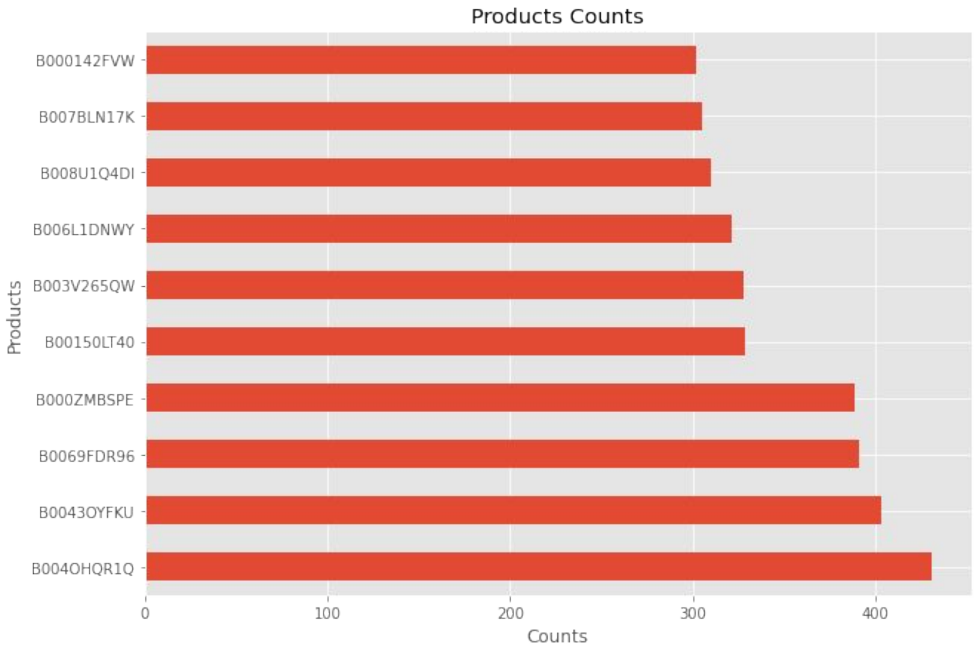
- ➢ **Collaborative Filtering system**
- ➢ **Content based system**
- ➢ **Hybrid based system**



Exploratory Data Analysis

Feature Name - ProductId

- ➢ **Plotted graph has only top 10 products.**
- ➢ **The graph is showing how many times a particular product has been sold.**

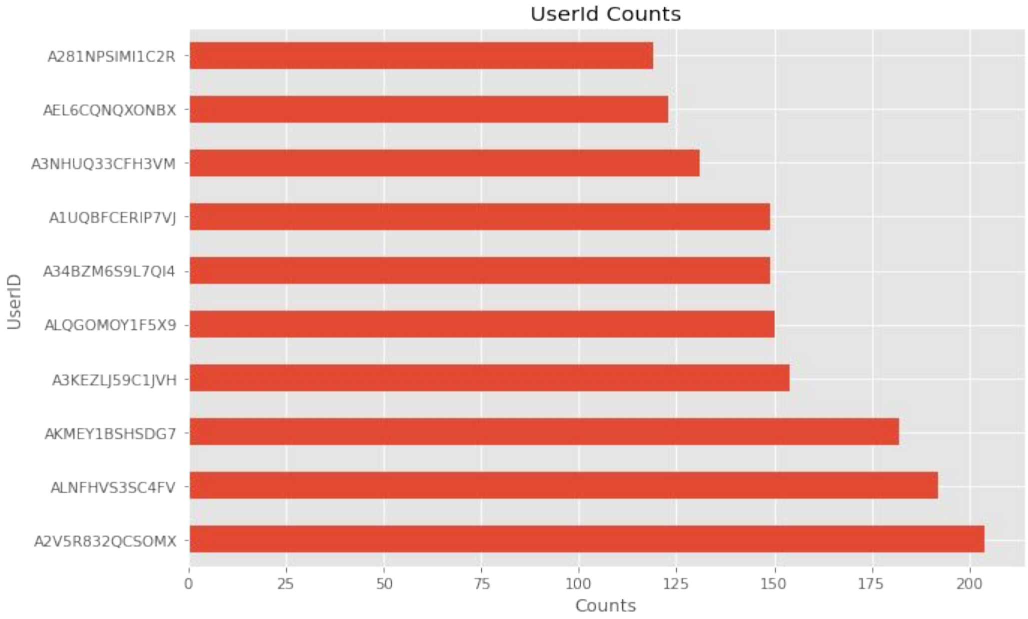


EDA Contd.

Feature Name - UserId

- ➢ Plotted graph has only top 10 Users.

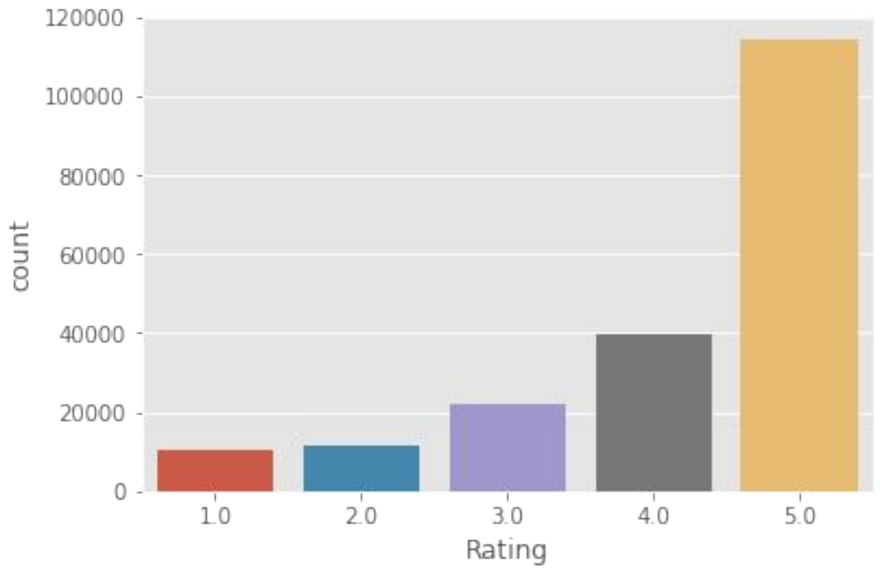
- ➢ The graph is showing how many times a particular user has purchased a products.



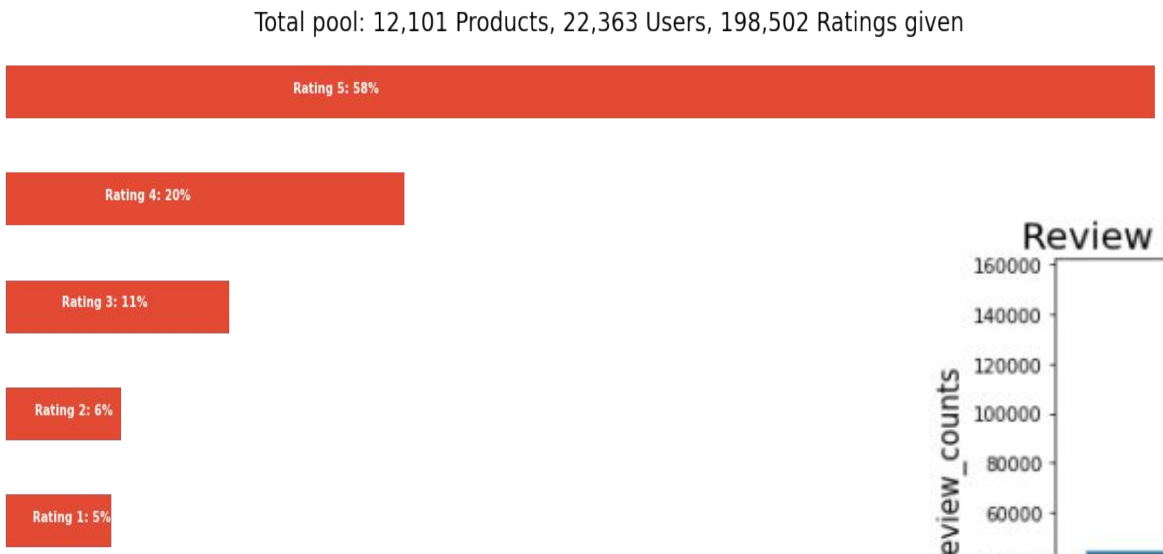
EDA Contd.

From chart it's clear that -

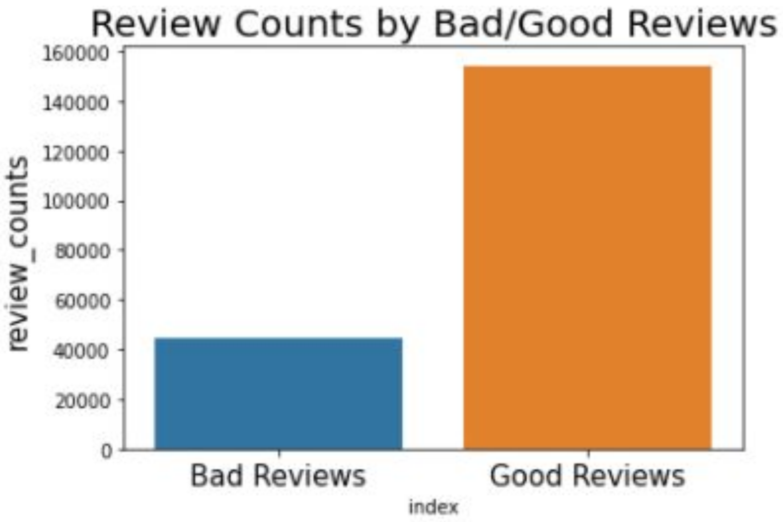
- ➢ **Most of the product has given as highest rating.**
- ➢ **Very less number of product has low rating.**



EDA Contd.



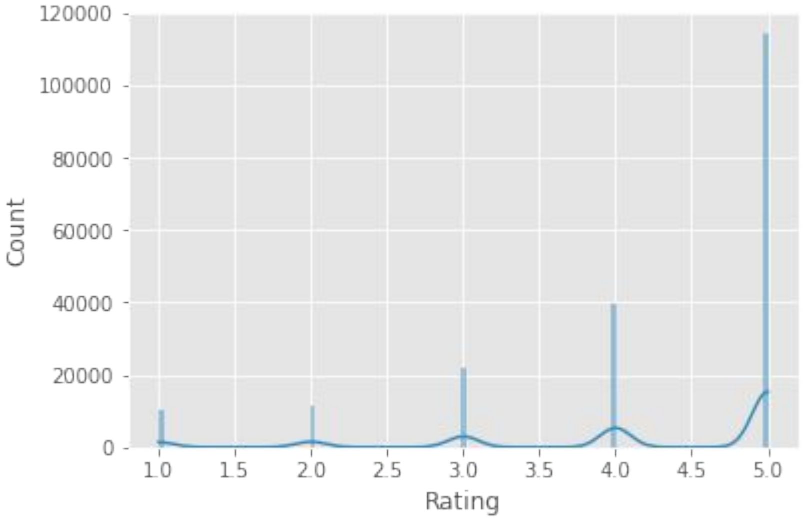
Total pool: 12,101 Products, 22,363 Users, 198,502 Ratings given
Rating 5: 58%
Rating 4: 20%
Rating 3: 11%
Rating 2: 6%
Rating 1: 5%



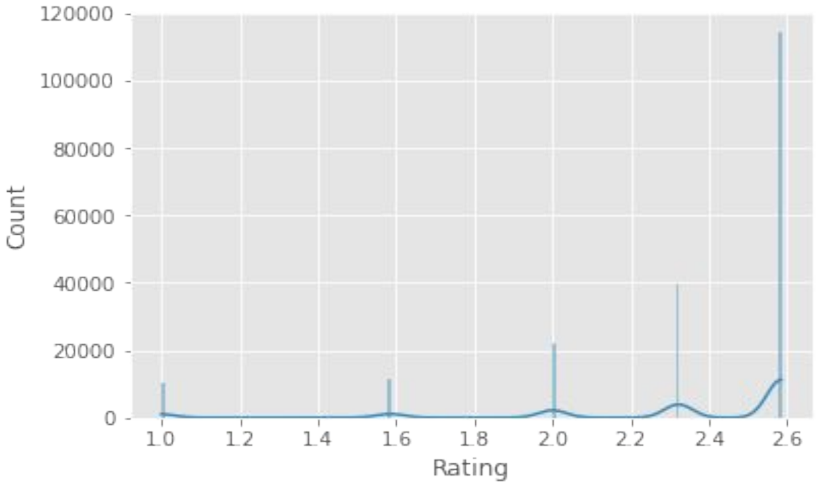
Review Counts by Bad/Good Reviews
review_counts
0
20000
40000
60000
80000
100000
120000
140000
160000
Bad Reviews
Good Reviews
index

EDA Contd.

Before Log Transformation

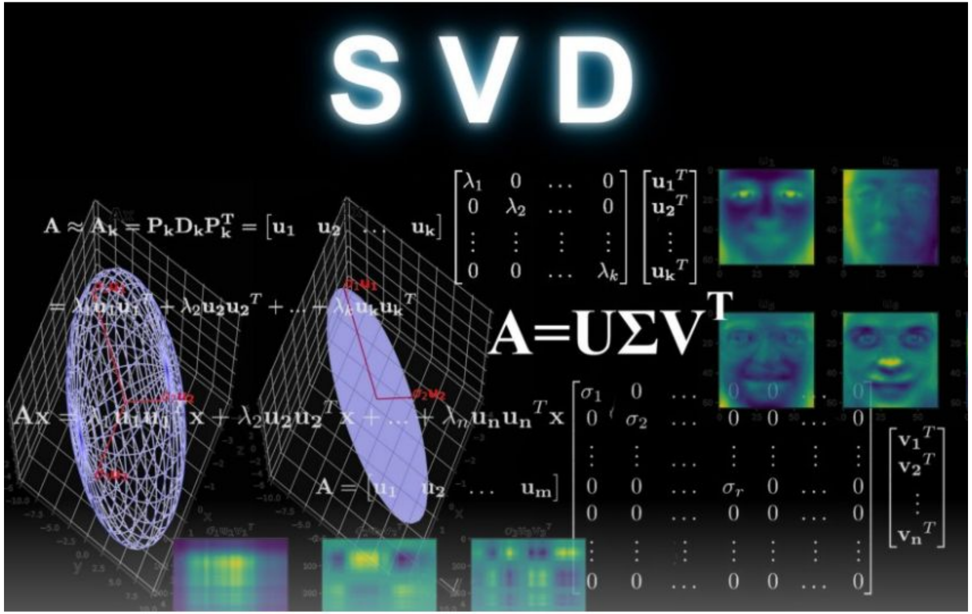


After Log Transformation



SVD - Singular Value Decomposition

➢ The Singular-Value Decomposition, is a matrix decomposition method for reducing a matrix to its constituent parts in order to make certain subsequent matrix calculations simpler. It provides another way to factorize a matrix, into singular vectors and singular values.



Machine Learning Algorithm

Collaborative Recommendation System

No of Components in SVD = 35

Pivot Matrix: Shape (22363,12101)

	A00414041RD0BXM6WK0GX	A00473363TJ8YSZ3YAGG9	A00700212KB3K0MVESPIY	A0078719IR14X3NNUG0F
ProductId				
7806397051	0.000056	0.000018	0.000076	-0.000218
9759091062	0.000055	0.000092	-0.000034	-0.000336
9788072216	0.000008	0.000121	0.000187	0.000046
9790790961	0.000222	0.000351	-0.000056	0.012612
9790794231	0.000005	0.000701	0.000289	0.000169

Evaluation for Collaborative Filtering

AI

Random 10 User Id Details

Overall Accuracy

recall@5: 0.3847
recall@10: 0.4759
recall@15: 0.5358

	hits@5_count	hits@10_count	hits@15_count	recall@5	recall@10	recall@15	interacted_count	UserId
1263	7	10	14	0.170732	0.243902	0.341463	41	A2V5R832QCSOMX
1094	34	36	36	0.871795	0.923077	0.923077	39	ALNFHVS3SC4FV
210	15	21	21	0.405405	0.567568	0.567568	37	AKMEY1BSHSDG7
1989	5	9	11	0.161290	0.290323	0.354839	31	A3KEZLJ59C1JVH
1711	17	18	19	0.566667	0.600000	0.633333	30	A34BZM6S9L7QI4
197	1	5	8	0.033333	0.166667	0.266667	30	ALQGOMOY1F5X9
992	27	28	30	0.900000	0.933333	1.000000	30	A1UQBFCERIP7VJ
137	23	23	24	0.884615	0.884615	0.923077	26	A3NHUQ33CFH3VM
545	22	22	22	0.880000	0.880000	0.880000	25	AEL6CQNQXONBX
1761	4	6	8	0.166667	0.250000	0.333333	24	A281NPSIMI1C2R

Content-Based Recommendation System

Dataset

	UserId	ProductId	reviewerName	helpful	reviewText	overall	summary	unixReviewTime	reviewTime	Rating	Timestamp
0	A1YJEY40YUW4SE	7806397051	Andrea	[3, 4]	Very oily and creamy. Not at all what I expect...	1	Don't waste your money	1391040000	01 30, 2014	1.0	1391040000
1	A60XNB876KYML	7806397051	Jessica H.	[1, 1]	This palette was a decent price and I was look...	3	OK Palette!	1397779200	04 18, 2014	3.0	1397779200
2	A3G6XNM240RMWA	7806397051	Karen	[0, 1]	The texture of this concealer pallet is fantas...	4	great quality	1378425600	09 6, 2013	4.0	1378425600
3	A1PQFP6SAJ6D80	7806397051	Norah	[2, 2]	I really can't tell what exactly this thing is...	2	Do not work on my face	1386460800	12 8, 2013	2.0	1386460800

TF-IDF Vectorizer Technique

```
vectorizer = TfidfVectorizer(analyzer='word',
                             ngram_range=(1, 2),
                             min_df=0.003,
                             max_df=0.5,
                             max_features=3000,
                             stop_words=stopwords_list)

item_ids = review_rating_df['ProductId'].tolist()
tfidf_matrix = vectorizer.fit_transform(review_rating_df['reviewText'] + "" + review_rating_df['summary'])
tfidf_feature_names = vectorizer.get_feature_names()
```

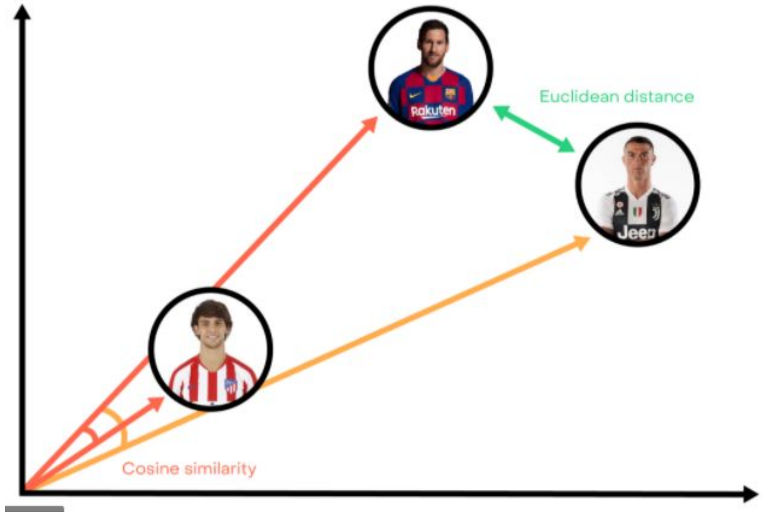
Contt..

Low Score Words

High Score Words

Cosine Similarity

- Cosine similarity measures the similarity between two vectors of an inner product space.
- It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction.
- It is often used to measure document similarity in text analysis.



$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}},$$

Token Relevance to a Particular User

AI

UserId: A00414041RD0BXM6WK0GX

	token	relevance
0	wig	0.715549
1	head	0.286604
2	fit	0.170562
3	average	0.140115
4	cap	0.139048
5	entire	0.123835
6	come	0.121345
7	hair	0.114995
8	fabulous	0.114200
9	mention	0.113886
10	paid	0.111911
11	super	0.110223
12	quality	0.106566
13	forever	0.105580
14	well	0.105234
15	totally	0.099761
16	size	0.098901
17	believe	0.093585
18	small	0.091248
19	blonde	0.087882

Evaluation for Content - Based Recomm...

AI

recall@5:0.83814

recall@10:0.8630

recall@15:0.8680

	hits@5_count	hits@10_count	hits@15_count	recall@5	recall@10	recall@15	interacted_count	UserId
1263	7	10	10	0.170732	0.243902	0.243902	41	A2V5R832QCSOMX
1094	12	14	15	0.307692	0.358974	0.384615	39	ALNFHVS3SC4FV
210	5	6	7	0.135135	0.162162	0.189189	37	AKMEY1BSHSDG7
1989	4	6	7	0.129032	0.193548	0.225806	31	A3KEZLJ59C1JVH
1711	7	9	9	0.233333	0.300000	0.300000	30	A34BZM6S9L7QI4

	hits@5_count	hits@10_count	hits@15_count	recall@5	recall@10	recall@15	interacted_count	UserId
10717	1	1	1	1.0	1.0	1.0	1	A2CMHND1J2REXO
10718	1	1	1	1.0	1.0	1.0	1	AT9IIRZG9EA
10719	1	1	1	1.0	1.0	1.0	1	A9V313DO1PZTF
10720	1	1	1	1.0	1.0	1.0	1	A18I3C6E5VKADI
22362	1	1	1	1.0	1.0	1.0	1	A3U46FFN9OP7BL

Challenges

- High Volume of Data.
- Elevating evaluation score for the models.
- Crashing of session due to large pivot matrix.
- Choosing optimal number of Factors in SVD.
- Implementing of Hybrid recommendation Model.

Conclusion

- We got recall@5: 0.3847 and recall@10: 0.4759 for collaborative Model.
- We got recall@5: 0.83814 and recall@10: 0.8630 for content-based Model.
- As we can see We are getting better recall value for content-based model than collaborative model.
- So we can conclude that content-based model is optimal model for product recommendation.

Thank You