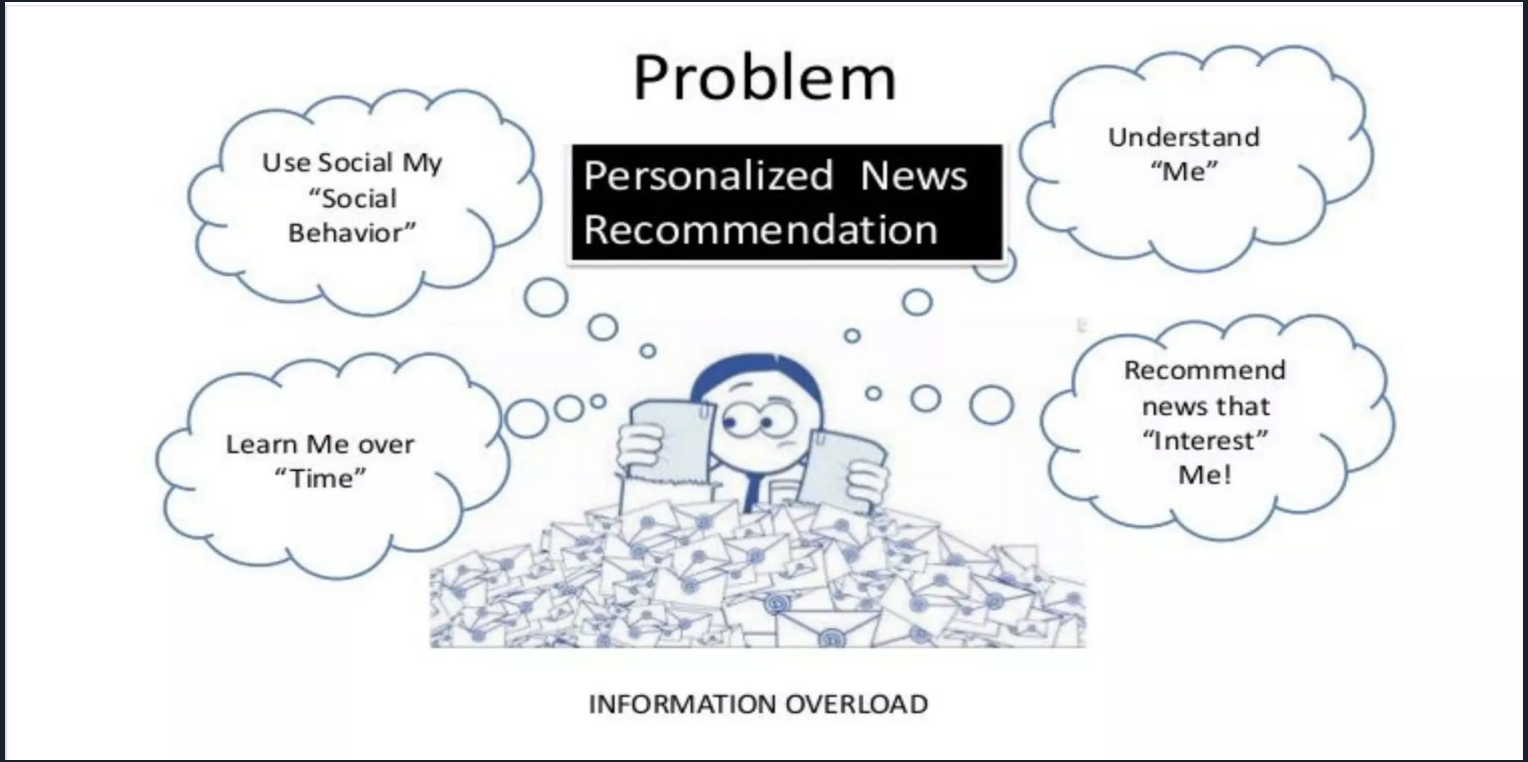
Article Recommendation System

DS5230 Final:

Bharath Gajula
Laasya Anantha Prasad
Nitika Jain

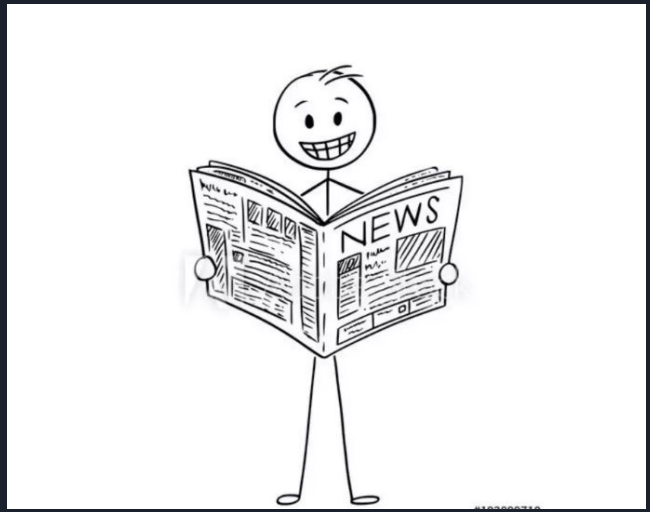
PROBLEM STATEMENT



DATASET

1. Articles Dataset (articles.csv) :
Each row represents a single article, with columns providing various attributes and metadata about it.

2. User Interaction Dataset (user_interaction.csv) :
This dataset captures user interactions with the articles. Each row in this dataset likely represents a single interaction event between a user and an article, with ails.



RECOMMENDATION ENGINE

In technical terms, a recommendation engine problem is to develop a **mathematical model or objective function** which can **predict** how much a user will like an item.

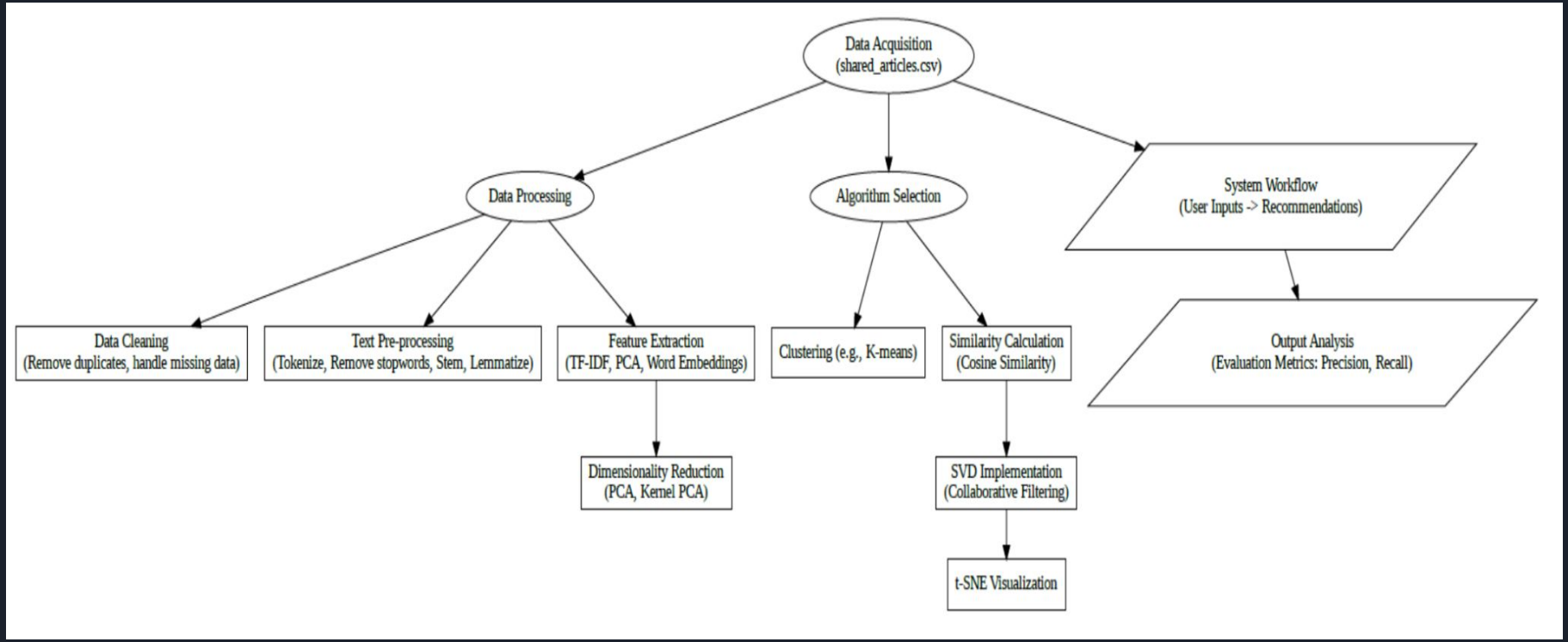
If U = {users}, I = {items} then F = Objective function and measures the usefulness of item I to user U, given by: $F: U \times I \rightarrow R$

Where R = {recommended items}.

For each user u, we want to choose the item i that maximizes the objective function:

$$u \in U, I' = \text{argmax } u(u, i)$$

SYSTEM ARCHITECTURE



INTERACTIONS SCORING EDA

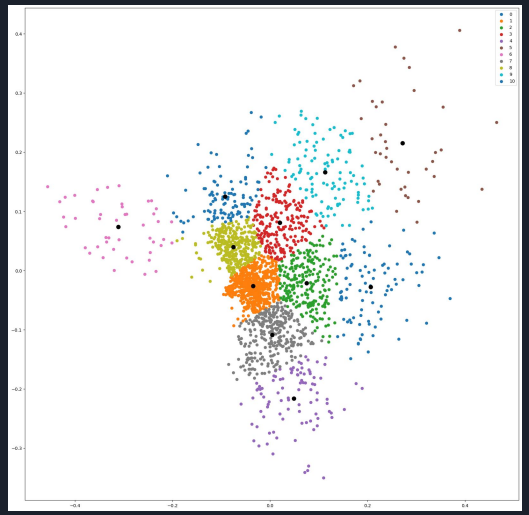
```
event_type_strength = {
   'VIEW': 1.0,
   'LIKE': 2.0,
   'BOOKMARK': 2.5,
   'FOLLOW': 3.0,
   'COMMENT CREATED': 4.0,
}
```

# of unique user/item interactions: 39106

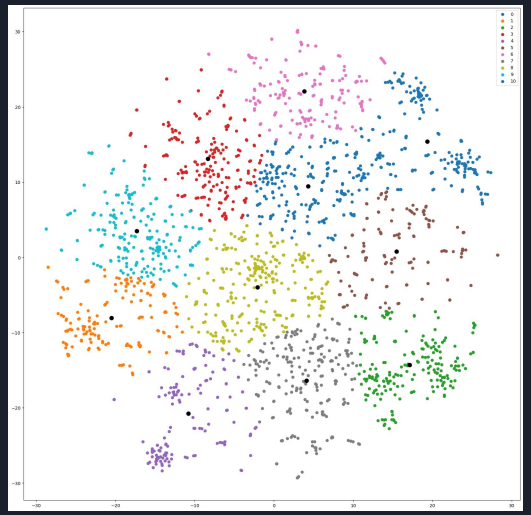
	personId	contentId	eventStrength
0	-9223121837663643404	-8949113594875411859	1.000000
1	-9223121837663643404	-8377626164558006982	1.000000
2	-9223121837663643404	-8208801367848627943	1.000000
3	-9223121837663643404	-8187220755213888616	1.000000
4	-9223121837663643404	-7423191370472335463	3.169925

contentType	url	title	text	lang	content	tags	clusters_text
HTML	http://www.nytimes.com/2016/03/28/business/dea...	Ethereum, a Virtual Currency, Enables Transact...	All of this work is still very early. The firs...	en	Ethereum, a Virtual Currency, Enables Transact...	[Ethereum, Bitcoin, Microsoft, Ether, Buterin]	1
HTML	http://cointelegraph.com/news/bitcoin-future-w...	Bitcoin Future: When GBPcoin of Branson Wins O...	The alarm clock wakes me at 8:00 with stream o...	en	Bitcoin Future: When GBPcoin of Branson Wins O...	[USDcoin, Trump, TeachBot, GBPcoin, TradeBot]	1
HTML	https://cloudplatform.googleblog.com/2016/03/G...	Google Data Center 360° Tour	We're excited to share the Google Data Center ...	en	Google Data Center 360° TourWe're excited to s...	[YouTube, Cardboard, Google Data Center 360, t...	1

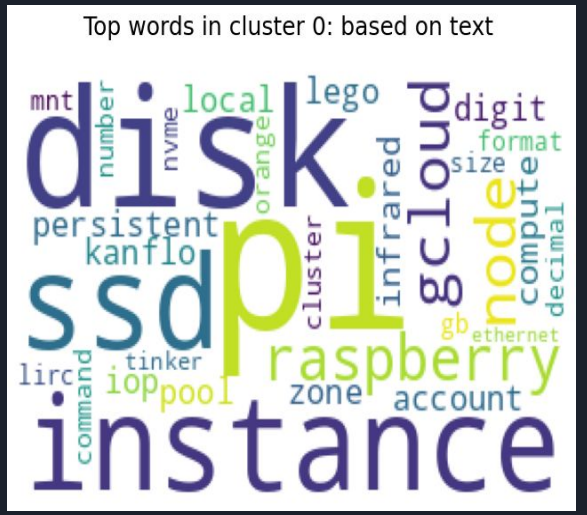
Dimensionality Reduction and visualization:



PCA Clustering



T-sne Visualization



K-means Cluster

Baseline Recommender

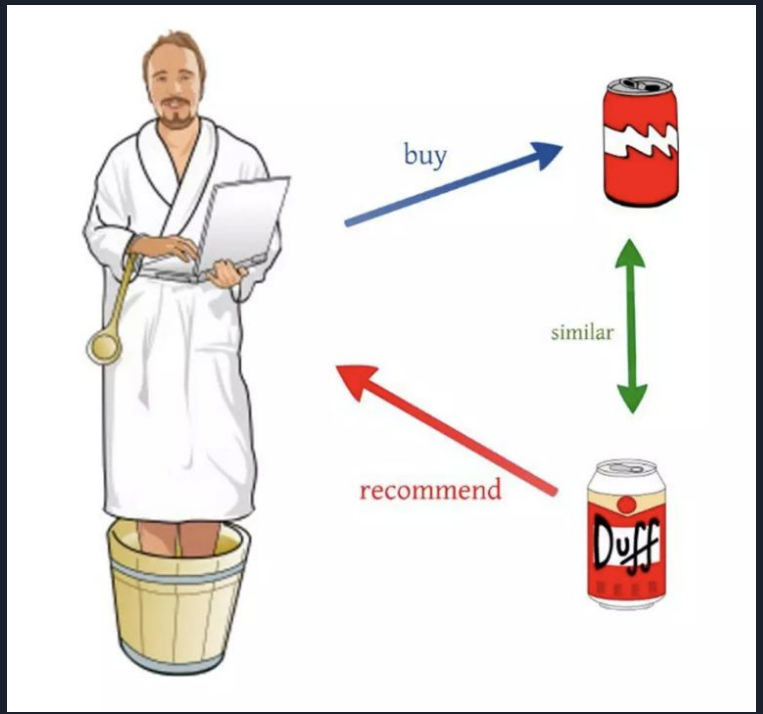
```
user_input = ['Learning to think critically about machine learning']
res = recommend(user_input, False, True)

print('Recommending top 5 articles according to the user input based on title:')
res
```

```
Cluster number is:  8
Recommending top 5 articles according to the user input based on title:
{'Deep Learning for Chatbots, Part 1 – Introduction',
 'Machine Learning for Designers',
 'My Top 9 Favorite Python Deep Learning Libraries – PyImageSearch',
 'Stop Coding Machine Learning Algorithms From Scratch – Machine Learning Mastery',
 'Why Learning Angular 2 Was Excruciating'}
```

Content Based Filtering
(Item Based Collaborative Filtering

- Based on the description of the item and profile of the user's preference.
- Keywords are used to describe the items: TFIDF
- User profile is built to indicate the type of item user may like.



(1, 5000)

	token	relevance
0	learning	0.294228
1	machine learning	0.245745
2	machine	0.236812
3	de	0.201505
4	google	0.194384
5	data	0.166056
6	ai	0.131141
7	que	0.121237

Evaluation Metrics

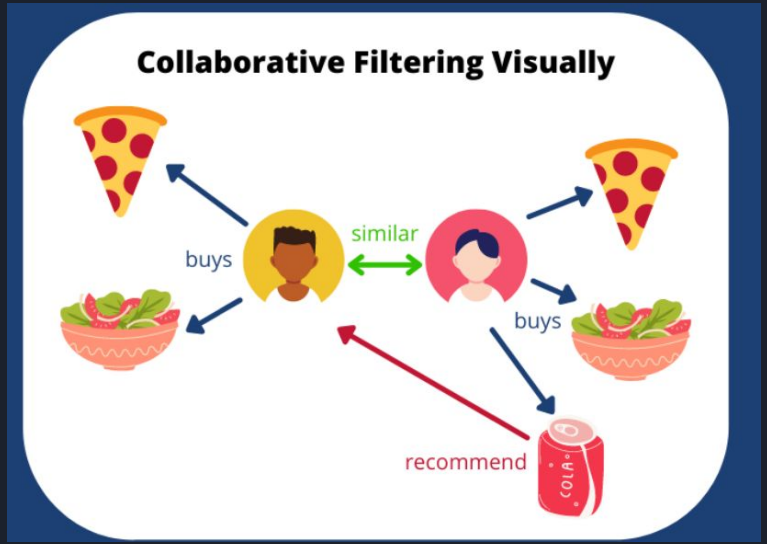
- Retrieve the set of items that a given user has interacted with.
- Generated a sample set of items that the user has not interacted with.
- Verify if an interacted item is within the top N recommended items.
- Evaluate the recommendation model for a single users like hits at 5 and 10, Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain (NDCG) based on these interactions.

```
Global metrics:
{'modelName': 'Content-Based',
'recall@5': 0.18818716440807978,
'recall@10': 0.27192533878803377,
'mrr': 0.2947636507945819,
'ndcg': 0.49570080981643383}
```

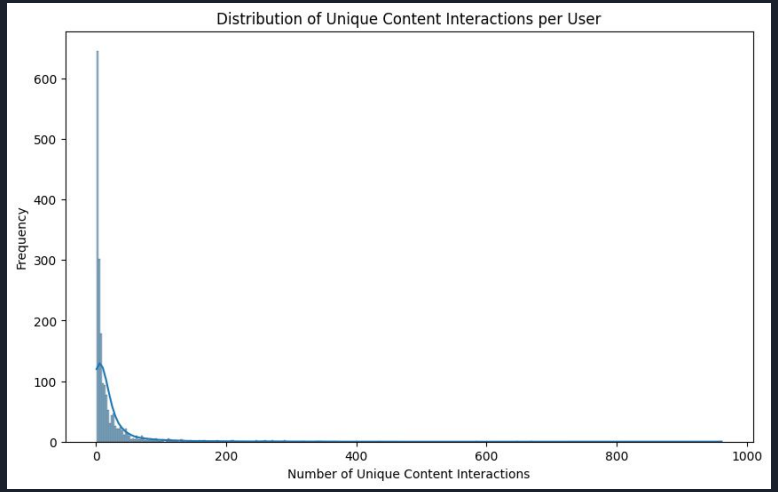
	hits@5_count	hits@10_count	interacted_count	recall@5	recall@10	mrr	ndcg	_person_id
76	10	21	192	0.052083	0.109375	0.124684	0.660217	3609194402293569455
17	8	12	134	0.059701	0.089552	0.133778	0.565099	-2626634673110551643
16	16	27	130	0.123077	0.207692	0.192364	0.867519	-1032019229384696495
10	20	37	117	0.170940	0.316239	0.180677	1.145046	-1443636648652872475
82	0	3	88	0.000000	0.034091	0.059720	0.311067	-2979881261169775358

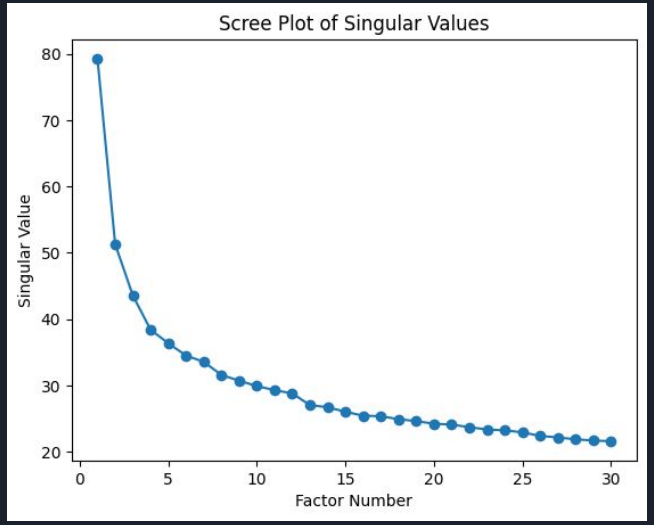
Collaborative Based Filtering (User Based Filtering)

- Based on large amount of information on users behaviour, activities or preferences.
- Predicting what users will like based on their similarity to other users.
- Advantage : Accurately recommends complex items such as keyworks and articles without understanding of the item itself.



COLLABORATIVE RECOMMENDATION EVALUATION

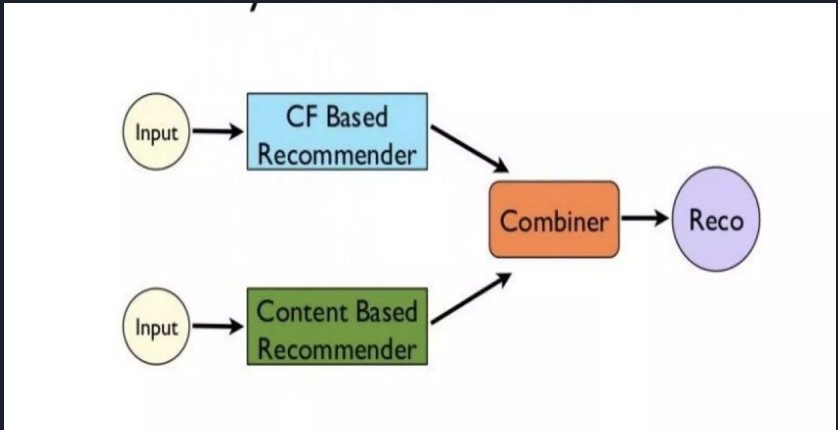




	recall@5	recall@10	mrr	ndcg
Collaborative Filtering	0.268346	0.401048	0.264244	0.548831

	hits@5_count	hits@10_count	interacted_count	recall@5	recall@10	ndcg	mrr	_person_id
112	31	50	192	0.161458	0.260417	0.007580	0.071429	3609194402293569455
47	26	44	134	0.194030	0.328358	0.016642	0.250000	-2626634673110551643
62	16	36	130	0.123077	0.276923	0.000000	0.000000	-1032019229384696495
51	35	46	117	0.299145	0.393162	0.006746	0.012658	-1443636648652872475
8	37	49	88	0.420455	0.556818	0.052318	1.000000	-2979881261169775358

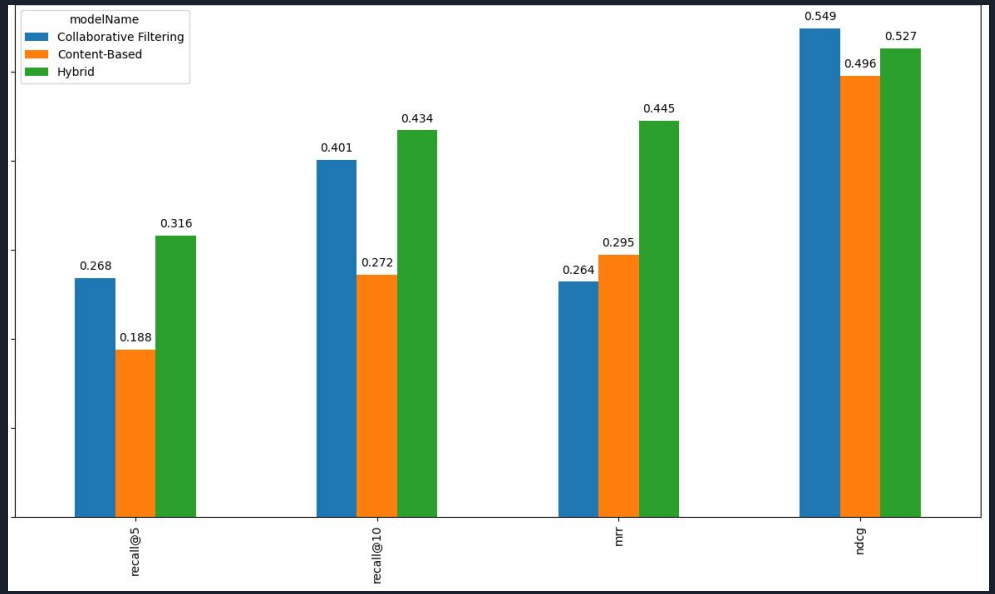
HYBRID RECOMMENDER



- Integrates content-based and collaborative filtering models.
- Calculates hybrid scores by multiplying CB and CF scores.

	recStrengthHybrid	contentId	title	url	lang
0	0.404984	5258604889412591249	Machine Learning Is No Longer Just for Experts	https://hbr.org/2016/10/machine-learning-is-no...	en
1	0.403726	-9033211547111606164	Google's Cloud Machine Learning service is now...	https://techcrunch.com/2016/09/29/googles-clou...	en
2	0.369363	-7126520323752764957	How Google is Remaking Itself as a "Machine Le...	https://backchannel.com/how-google-is-remaking...	en
3	0.350236	-4571929941432664145	Machine Learning as a Service: How Data Scienc...	http://www.huffingtonpost.com/laura-dambrosio/...	en
4	0.304937	2589533162305407436	6 reasons why I like KeystoneML	http://radar.oreilly.com/2015/07/6-reasons-why...	en
5	0.289812	3269302169678465882	The barbell effect of machine learning.	http://techcrunch.com/2016/06/02/the-barbell-e...	en
6	0.270451	524776334673868069	Graph-powered Machine Learning at Google	https://research.googleblog.com/2016/10/graph-...	en
7	0.270093	5092635400707338872	Power to the People: How One Unknown Group of ...	https://medium.com/@atduskgreg/power-to-the-pe...	en
8	0.267138	3320201327008235211	How Mark Zuckerberg Led Facebook's War to Crus...	http://www.vanityfair.com/news/2016/06/how-mar...	en
9	0.256739	-1901742495252324928	Designing smart notifications	https://medium.com/@intercom/designing-smart-n...	en
10	0.252827	-4541461982704074404	Exclusive: Why Microsoft is betting its future...	http://www.theverge.com/2016/7/7/12111028/micr...	en
11	0.249146	7395435905985567130	The AI business landscape	https://www.oreilly.com/ideas/the-ai-business-...	en

Post EDA Analysis and Important Insights



modelName	recall@5	recall@10	mrr	ndcg
Hybrid	0.316287	0.434416	0.445120	0.526664
Collaborative Filtering	0.268346	0.401048	0.264244	0.548831
Content-Based	0.188187	0.271925	0.294764	0.495701