Recommendation Systems - End Term Project

News Recommendation System

•••

April 28, 2023

Team Members

Pavan Thanay	(IMT2020024)
Krushikar Reddy	(IMT2020043)
Chaithanya Reddy	(IMT2020054)
Samarth Gattu	(IMT2020062)
Sougandh Krishna	(IMT2020120)

Why News Recommendation System? Social utilities?

- Combatting biases: News recommendation systems can help break users out of filter bubbles by exposing them to news articles that challenge their existing beliefs and opinions. This can promote critical thinking and reduce polarization in society.
- Encouraging engagement with news: News recommendation systems can suggest articles that are relevant to users' interests, making it more likely that they will engage with the news and stay informed about current events.
- Supporting the news industry: News recommendation systems can help news organizations reach new audiences and increase engagement with their content, which can ultimately support the financial sustainability of the news industry.

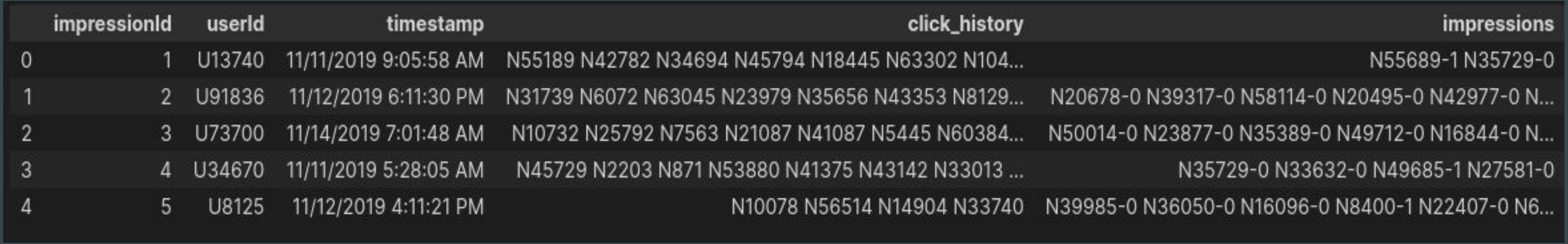
How is News Recommendation System challenging compared to others?

- Severe Cold-start problem
  - User cold-start problem - Login and new user problem
  - Product cold-start problem - Large number of new news articles
- In traditional recommendation systems we have user ratings to each item, but we do not have explicit news article ratings.
- Item representation for news articles must be based on it’s content.
- Real time Updates to the dataset.

Hence a Challenging Problem!!

MIND dataset

- behaviour.tsv - This file contains the click history of the user, all the impressions shown to the user and the impression time, and the user clicks.
- An impression log records the news articles displayed to a user when the user visits the news page at some specific time.

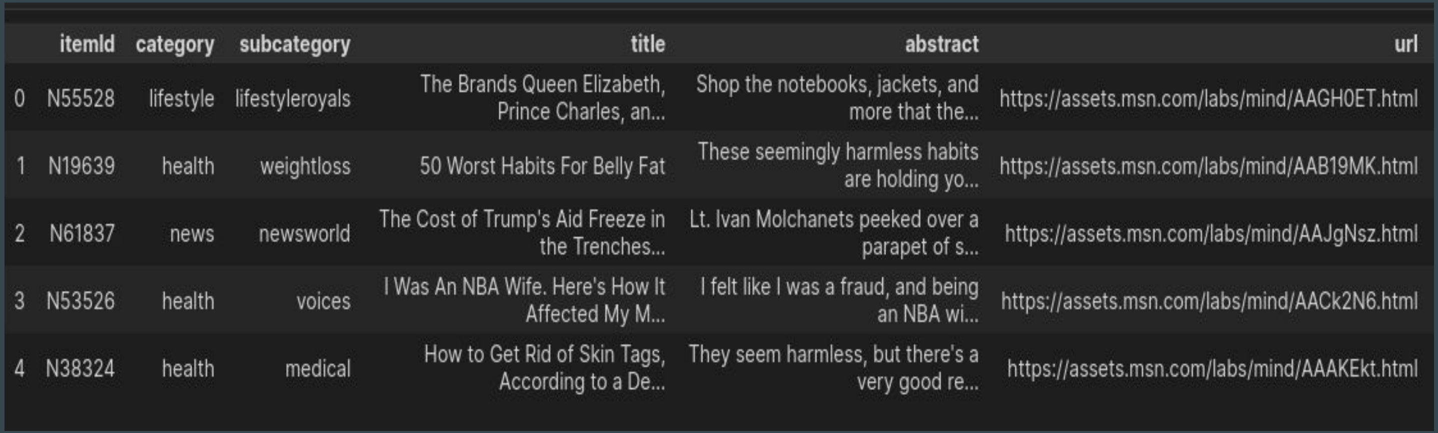


	impressionId	userId	timestamp	click_history	impressions
0	1	U13740	11/11/2019 9:05:58 AM	N55189 N42782 N34694 N45794 N18445 N63302 N104...	N55689-1 N35729-0
1	2	U91836	11/12/2019 6:11:30 PM	N31739 N6072 N63045 N23979 N35656 N43353 N8129...	N20678-0 N39317-0 N58114-0 N20495-0 N42977-0 N...
2	3	U73700	11/14/2019 7:01:48 AM	N10732 N25792 N7563 N21087 N41087 N5445 N60384...	N50014-0 N23877-0 N35389-0 N49712-0 N16844-0 N...
3	4	U34670	11/11/2019 5:28:05 AM	N45729 N2203 N871 N53880 N41375 N43142 N33013 ...	N35729-0 N33632-0 N49685-1 N27581-0
4	5	U8125	11/12/2019 4:11:21 PM	N10078 N56514 N14904 N33740	N39985-0 N36050-0 N16096-0 N8400-1 N22407-0 N6...

MIND dataset

- news.tsv - This file contains the description of each news article.
- The survival time of more than 84.5% news articles is less than two days.

- News-id
- URL
- Category
- Sub-category
- Title
- Abstract

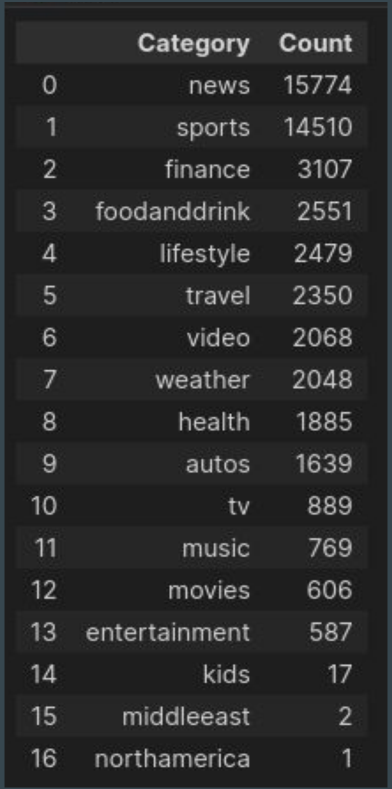


	itemId	category	subcategory	title	abstract	url
0	N55528	lifestyle	lifestyleroyals	The Brands Queen Elizabeth, Prince Charles, an...	Shop the notebooks, jackets, and more that the...	https://assets.msn.com/labs/mind/AAGH0ET.html
1	N19639	health	weightloss	50 Worst Habits For Belly Fat	These seemingly harmless habits are holding yo...	https://assets.msn.com/labs/mind/AAB19MK.html
2	N61837	news	newsworld	The Cost of Trump's Aid Freeze in the Trenches...	Lt. Ivan Molchanets peeked over a parapet of s...	https://assets.msn.com/labs/mind/AAJgNsz.html
3	N53526	health	voices	I Was An NBA Wife. Here's How It Affected My M...	I felt like I was a fraud, and being an NBA wi...	https://assets.msn.com/labs/mind/AACk2N6.html
4	N38324	health	medical	How to Get Rid of Skin Tags, According to a De...	They seem harmless, but there's a very good re...	https://assets.msn.com/labs/mind/AAAKEkt.html

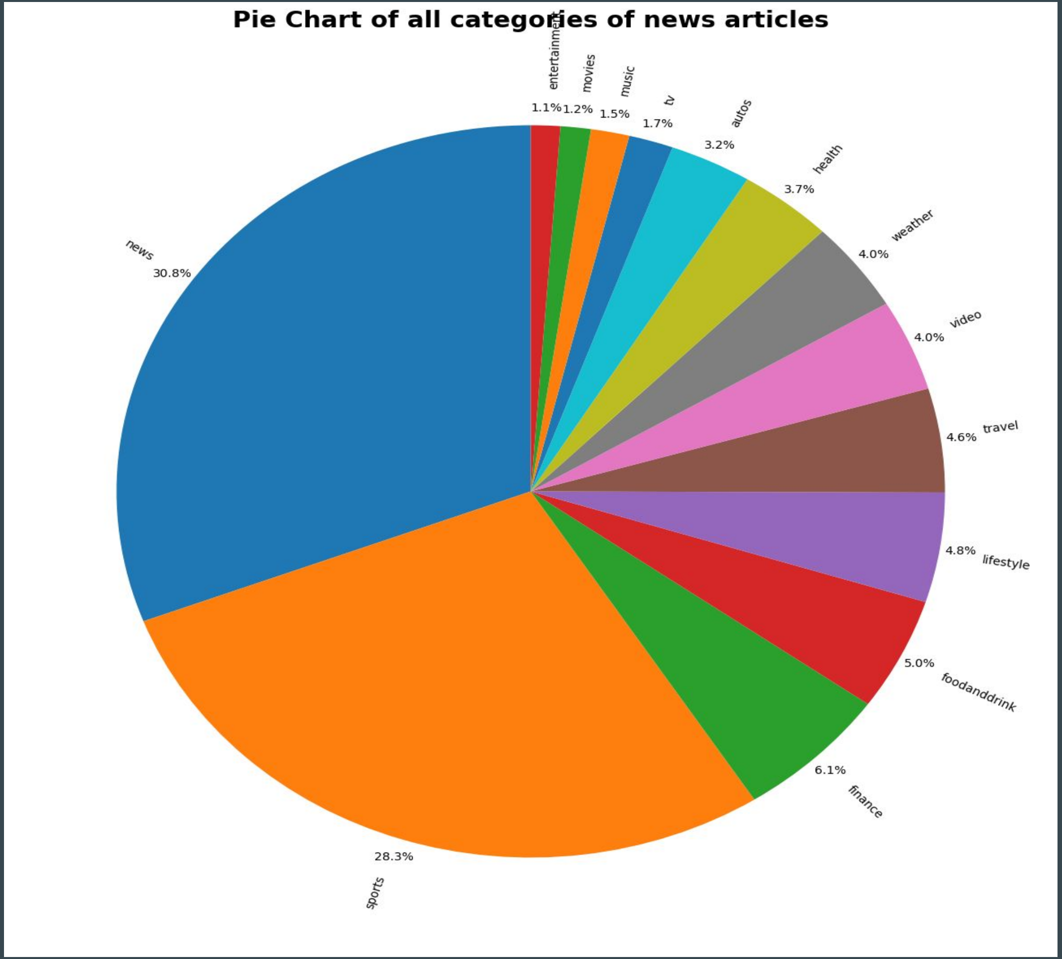
EDA

- The news dataset contains 17 categories and 264 sub categories.
- 60% of the news articles belong to either news or sports category.
- The dataset consists of the user behaviour records between the period November 9, 2019 to November 14, 2019.
- Users click only approximately 4% of the news articles shown to him.
- Approximately more than 30,000 news articles (60%) are not even displayed to the user in this time period.
- Each user in the dataset have at least 17 news article clicks in this period.

EDA

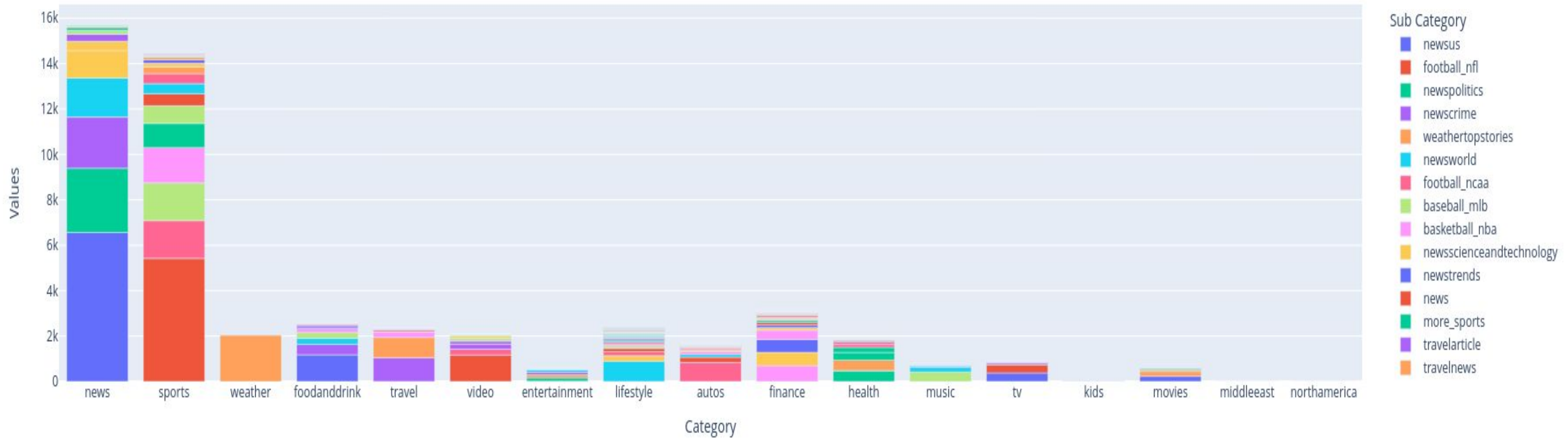


	Category	Count
0	news	15774
1	sports	14510
2	finance	3107
3	foodanddrink	2551
4	lifestyle	2479
5	travel	2350
6	video	2068
7	weather	2048
8	health	1885
9	autos	1639
10	tv	889
11	music	769
12	movies	606
13	entertainment	587
14	kids	17
15	middleeast	2
16	northamerica	1



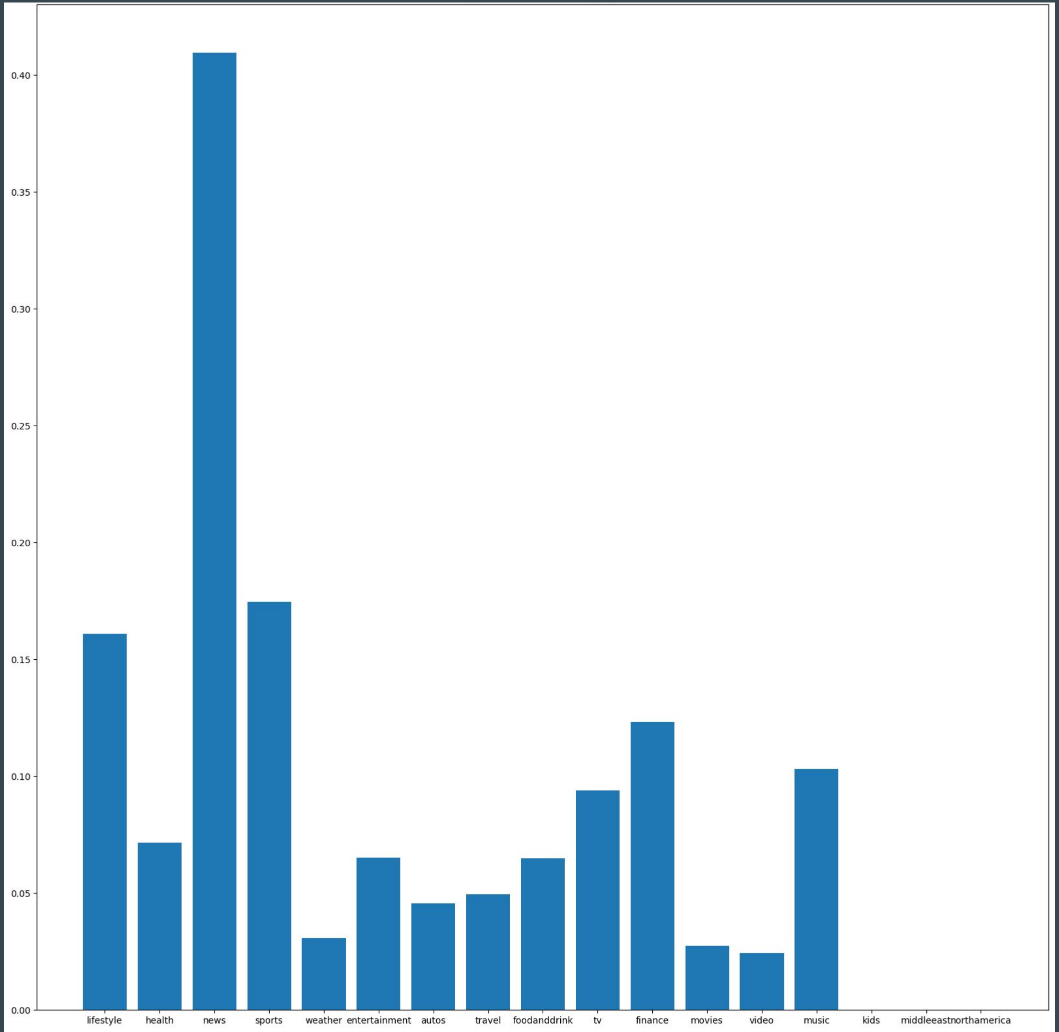
Pie Chart of all categories of news articles

EDA


Values
16k
14k
12k
10k
8k
6k
4k
2k
0
news
sports
weather
foodanddrink
travel
video
entertainment
lifestyle
autos
finance
health
music
tv
kids
movies
middleeast
northamerica
Category
Sub Category
newsus
football_nfl
newspolitics
newscrime
weathertopstories
newsworld
football_ncaa
baseball_mlb
basketball_nba
newsscienceandtechnology
newstrends
news
more_sports
travelarticle
travelnews

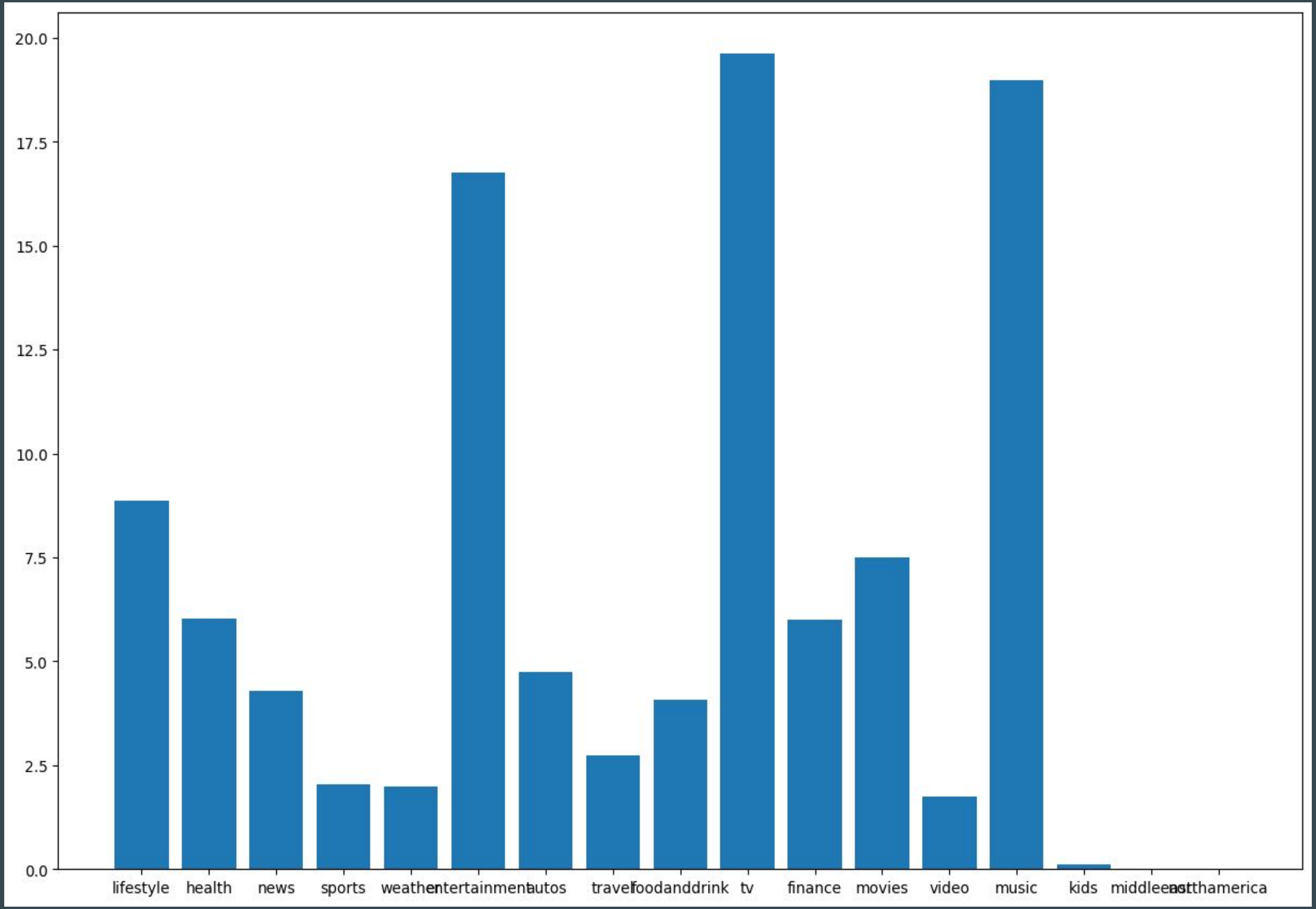
EDA

Distribution of Clicks of users



EDA

Total Clicks
―――
No of articles



Pre-processing

- Drop URL column - The given URL for news article is not accessible.
- Replacing with empty string (" ") for NaN values.
- Pre-processed the concatenated strings in the news dataset (to lower case, remove symbols, remove stop words, applied tokenization & lemmatization)
- News-embeddings
- User-embeddings
- Creating click and non-click list for all the impressions.
- Changing the data types as per the actual data type of the column.

News Embeddings

- Used sentence transformers for embedding the news article representation.

  - **Representation 1** - Concatenate category, subcategory, title, abstract
  - **Representation 2** - Treat each of the above as different entity and then take the weighted average.
  - **Representation 3** - Individually find embeddings of category, subcategory, title, abstract and then concatenate the embeddings.

  Using the sentence transformer based on BERT to map sentences to a 384 dimensional vector space.

User Embedding

- **Representation - 1**:
  - We defined the user embedding to be the mean of news article embeddings (Representation 1) in the user's click history.

- **Representation 2:**
  - 'K' is defined as the average vectors of all the news article in the user's click history. The weights are the values of the dot product of 'K' with each of the item embedding of the click history. Then we take the weighted average of these vectors with these weights for the final user embedding.

  - Final user representation $= \Sigma_{j=1}^{n} (K.v_j)\ v_j$

User Embedding

- **Representation 3 (Attention):**
  - In this method we construct ‘n’ new vectors V_i' as below and then take the mean as the final user representation.
  - $V_i' = \Sigma_{j=1}^{n} (v_i.v_j)v_j$

User Embedding

- **Representation 4:**
    - In this method we take the ‘n’ vectors V_i obtained in previous method, and then apply the first method for these ‘n’ vectors, to find the final user embedding.

  - The above mentioned methods are content based, collaborative based embeddings will be discussed later.

News Recommendation System

- A Layman Recommendation System
- K-Means Clustering
  - Clustering Users and News articles
  - Clustering Users
  - Clustering Articles
- BPR
- MAB
- Trending
- A different approach

A simple Recommendation System

- Ask the user to choose some articles from randomly selected ‘25’ articles.
- **Method -1** : Represent this new user as the average of all these news article embeddings, now take the top 5 closest news articles to this representation and recommend them to the user.
- **Method -2** : For all the articles chosen by the user find the nearest articles to each of the news article and recommend to the user.
- Both types of news article embeddings (category, sub-category, title, abstract as single entity and multiple entities) are used to design these 2 methods.

K-Means Clustering

Fixed number of clusters as 2000 for both news articles and users.

- Clustering Users and News articles

  After assigning the user to the cluster randomly pick 25 articles from the articles cluster which is closest to the user cluster representation.

- Clustering Users

  After assigning the user to the cluster randomly pick 25 articles from the articles viewed by the users in the cluster and is not seen by the user.

- Clustering Articles

  Assign the user to the closest Article cluster and randomly pick 25 articles from the cluster

Trending news articles

- News article publication time assumption:
  The publication time of the news article is the time of the impression it first appeared in.
- Trending Article prediction is a function of the publication time and the number of clicks for the article.
- Top ‘k’ articles which have the highest number of clicks in the past 2 days are considered as trending articles.

BPR - Bayesian Personalized Ranking

- **Dataset creation:**
  - Each impression is converted into a table (liked, not liked) where each combination of (click, non-click) of this impression is placed in the table.
  - A user’s dataset for this model is the union of all these tables for all the user’s impressions.
  - Now the news article in the dataset is replaced with the Hadamard product of user embedding and the news article embedding.

BPR - Bayesian Personalized Ranking

- Example:

  1st Impression:
  clicks = 1, non-clicks = 2,3

  2nd Impression:
  clicks = 4,5 non-clicks = 6

Liked (click)	**Not liked**	True/False
1	2	1
1	3	1
4	6	1
5	6	1
2	1	0
3	1	0
6	4	0
6	5	0

BPR - Bayesian Personalized Ranking

Prediction

- Now we trained this dataset using Logistic Regression.
- Suppose we have ‘k’ articles from which we need to provide the impression for the user then we need to compute k x k matrix where (i,j) cell indicates whether the user likes the news article ‘i’ compared to the news article ‘j’.
- Now for each article ‘i’ maintain the vector of count of articles ‘j’ such that user likes the news article ‘i’ compared to the article ‘j’ according to the model prediction.
- The articles with top ‘k’ values are the final impressions given to the user.

BPR - Bayesian Personalized Ranking

Prediction

	1	2	3	4
1		0	1	0
2	1		1	0
3	0	0		0
4	1	1	1	

⇨

	1	2	3	4
count	1	2	0	3

Learning User and Item Embeddings “collaboratively”

- Consider a user with embedding u and let the embeddings that they liked and disliked be l and d respectively.
- We build a model that expects the dataset containing (u, l, d) tuples.
- The model “transforms” u to $W_u u$; l and d to $W_a l$ and $W_a d$, respectively.
- Let the new embeddings be u’, l’ and d’.
- The objective function $L = \Sigma (\langle u', l' \rangle - \langle u', d' \rangle)$
- The weight matrices are learnt in such a way that L is maximized.
- Some inspiration has been taken from Linear Metric Learning.
- The model achieves about 66 % accuracy.
- The model can easily be improved by using a neural network to learn better transformations.

Multi Arm Bandits - Thompson Sampling Version

- This is demonstrated as a purely **“Reels”** version
- All articles are categorised into 4 main categories. There are 4 arms and each arm represents these main categories.

Arm	**Categories**
Life	Lifestyle, health, weather, food and drink, travel, kids
Entertainment	Entertainment, TV, movies, sports
World	news , finance, middle-east, north america, autos
Video	video

Multi Arm Bandits - Thompson Sampling Version

- Each arm is picked with probability p which is drawn from distribution beta(α,β).
- These (α,β) are updated based on whether user liked the article or not.
- Note that there are only 4 arms so that during demonstration, we can see user’s interest converge to specific category / categories quickly.

Future Scope

- Improved user and news article embeddings.
- Applying RNN to train and predict user’s preferences based on view history.
- News article distribution is not uniform, so better sampling can be done.
- Develop it into a web based application.
- BPR set representations.
- Granular user type.

Thank You!