Jhakass NewsVala

News Recommender System

-Midway Report

Group -16 CARDS

Chhavi Chahar MS18136

Amlan Nayak MS18197

Rimjhim Goel MS18133

Dheer Mankad MS18140

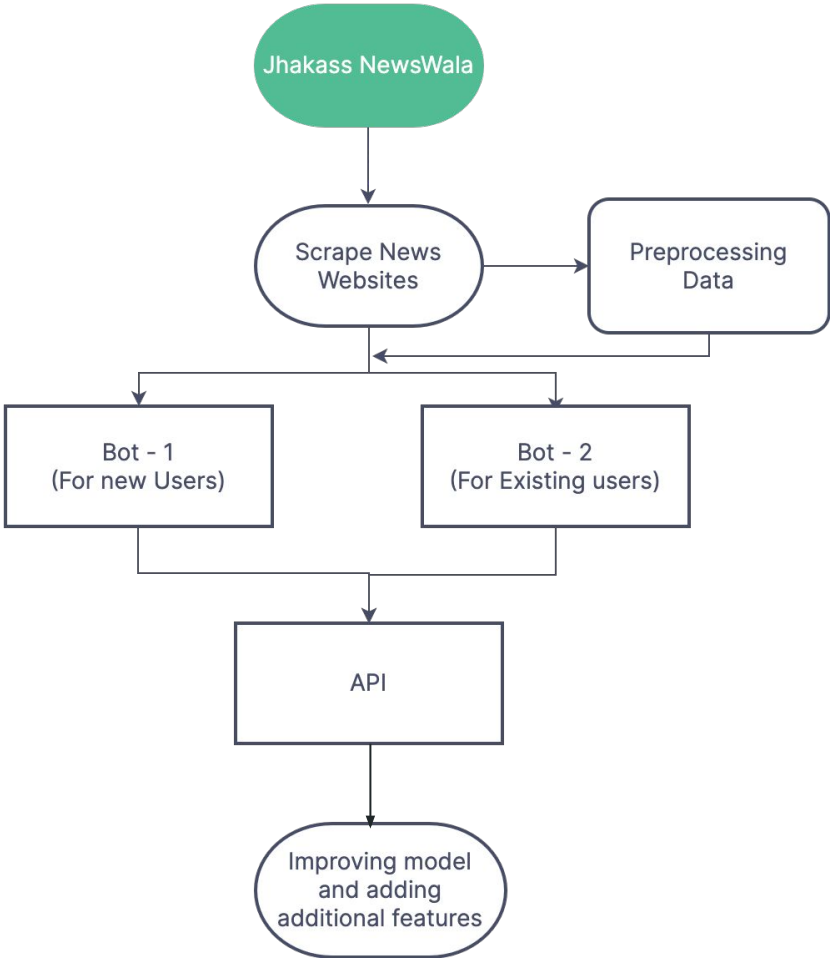
Smriti Chhibber MS18142

Introduction

- The aim of this project is to build a recommender system for the start-up Jhakaas NewsVala.
- The recommender system targets the working professionals in the age group 21-40 and serves them with 10 articles per scroll as per their interests and preferences.
- The goal is to retain maximum amount of new users and serving the existing ones with exciting news while keeping their interests in mind.
- As discussed in the strategy plan, we have done web scrapping of news articles and performed exploratory data analysis. Bot1 (for the new users) has been developed. Bot2 is still being worked on.

Further plan is under process.

Workflow


Jhakass NewsWala
Scrape News Websites
Preprocessing Data
Bot - 1 (For new Users)
Bot - 2 (For Existing users)
API
Improving model and adding additional features

Web Scraping Data

- News sources: The Republic, The Hindu, International Business Times(IBT), India Today
- Total no. of articles scraped: 2956
- Categories: Business, Entertainment, General, Sci-tech and Sports
- Avg. reading time: Approximately 1 minute 40 seconds
- Date of publication between 8th Dec,2021 and 9th April,2022

Output

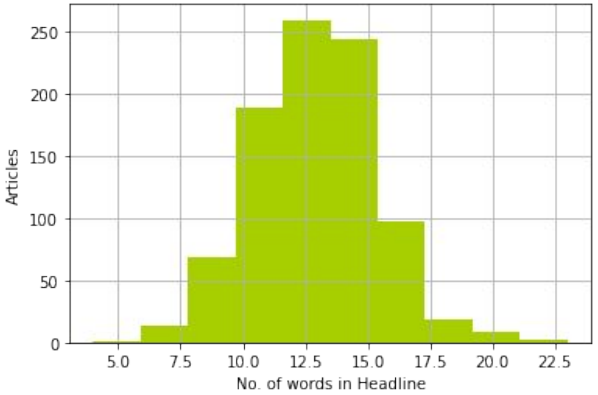
	Category	Sub-Category	Headline	Entire_News	Author	News_Link	Mean_Time	DateTime
0	general	politics	Bill To Ban Financing WMDs Introduced In LS	New Delhi, Apr 5 (PTI) The government on Tuesd...	Press Trust Of India	https://www.republicworld.com/india-news/polit...	112	2022-04-05 16:58:33
1	general	politics	No One Can Erase My Name From Nagpur-Mumbai Ex...	Nagpur, Apr 5 (PTI) Maharashtra BJP leader De...	Press Trust Of India	https://www.republicworld.com/india-news/polit...	49	2022-04-05 16:51:03
2	general	politics	Congress Faces Big Blow As Ahmed Patel's Son F...	In a massive development, Republic has now lea...	Vishnu V V	https://www.republicworld.com/india-news/polit...	132	2022-04-05 16:47:40
3	general	politics	Congress To Fight On Its Own In Telangana; Rah...	Congress will not ally with the Telangana Rash...	Abhishek Raval	https://www.republicworld.com/india-news/polit...	86	2022-04-05 16:32:39
4	general	politics	Parliament Approves Accountancy Bill	New Delhi, Apr 5 (PTI) Parliament on Tuesday a...	Press Trust Of India	https://www.republicworld.com/india-news/polit...	88	2022-04-05 16:29:34

Exploratory Data Analysis (EDA)

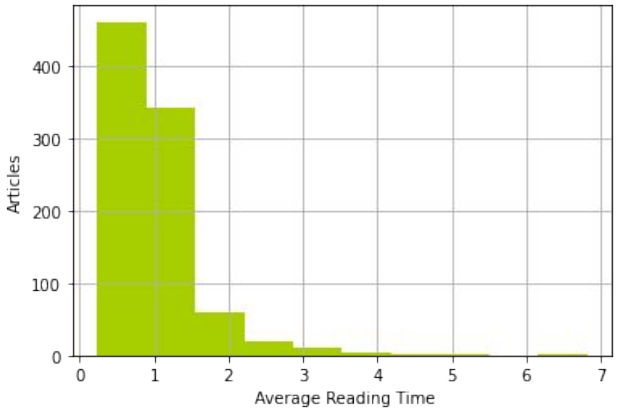
The following is the EDA presented for International Business Times

count	900.000000
mean	449.500000
std	259.951919
min	0.000000
25%	224.750000
50%	449.500000
75%	674.250000
max	899.000000

Articles vs Number of words in headlines



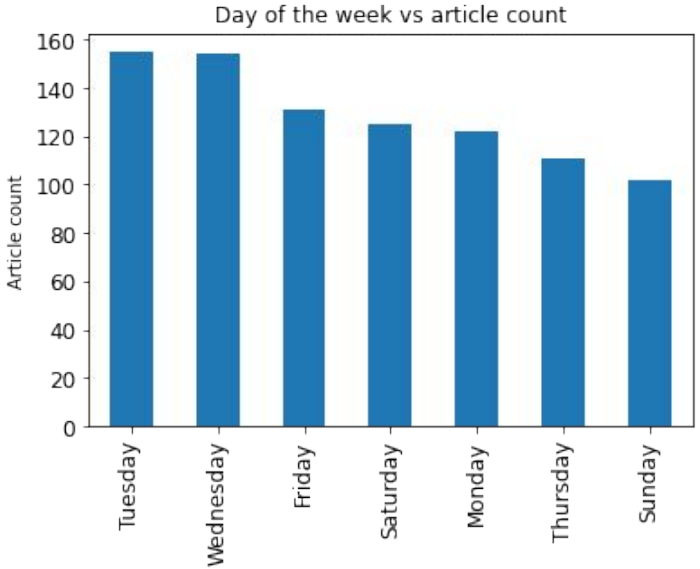
Articles vs Average reading time of user

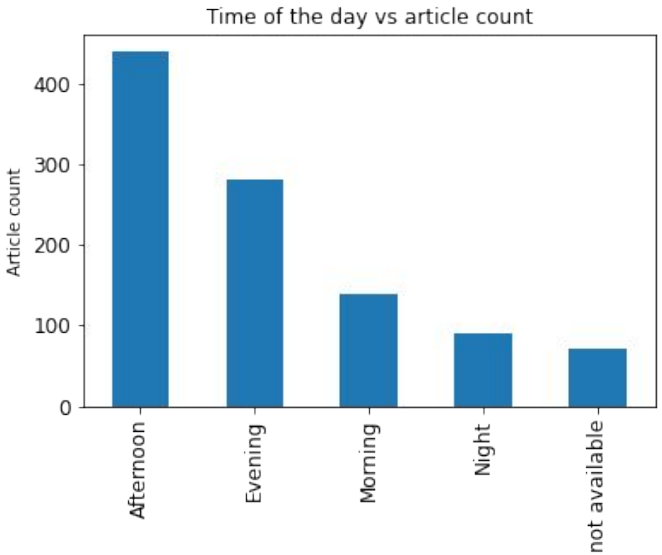


Conclusion

- No. of words in headline can be used to predict average user preference for headline
- Considering average reading speed 250/minute, average reading time predicted can be used to compare with observed reading time by the user.

Variation of article count with time and day





- This analysis is useful in predicting which particular time and day, the probability of readers increasing is more, due to more inflow of articles
- It is also helpful in predicting what subset of news corpus would be recommended to the user

User Clickstream Data

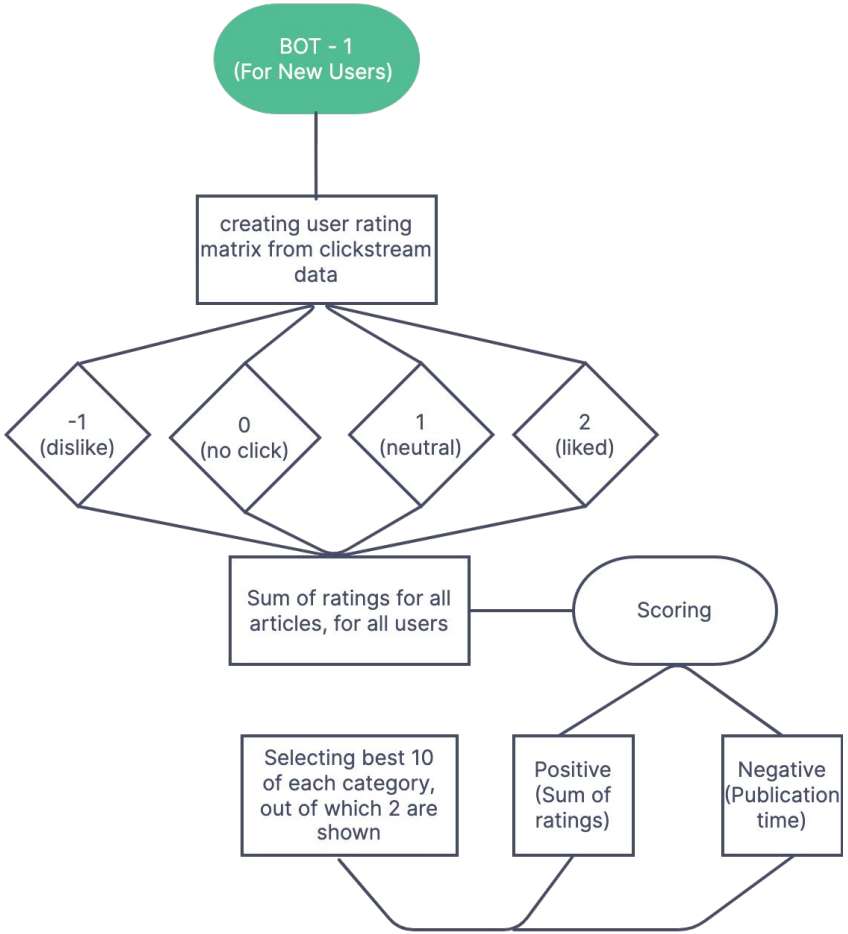
```
click
```

	UserId	SessionId	ArticleId	ArticleRank	Clicked	TimeSpent
0	1	1	1213	1	0	0
1	1	1	1215	2	0	0
2	1	1	566	3	1	64
3	1	1	1544	4	0	0
4	1	1	1917	5	1	54
...	...	...	...	...	...	...
1305	19	6	1393	6	0	0
1306	19	6	2251	7	1	50
1307	19	6	941	8	1	37
1308	19	6	1416	9	0	0
1309	19	6	1696	10	0	0

1310 rows × 6 columns

Bot - 1 (For new users)

- We first generate an average user rating dataframe for articles. Each article can have rating {-1,0,1,2}, 0 is for article not being read, -1 is a disliked article, 1 is a read article with neutral likeness, 2 is read and exceptionally liked article.
- Binomial distribution with probability 0.4 is used, so most articles have 0 rating, as an individual user will not have read most articles.
- Bot 1 recommends articles to new users based on activity of existing users. Takes in two dataframes: Publication date and time of articles, and average user rating of articles.



Reducing Bias

- Finds age of article. Assigns appropriate weight to age and rating, and combining them, assigns a score to the article.
- Takes top ten articles from each category. For bias reduction, two random articles from each top ten are displayed to different users, thus ten articles (2 each for 5 categories). Almost 180 million possible distinct combinations of articles that can be served.

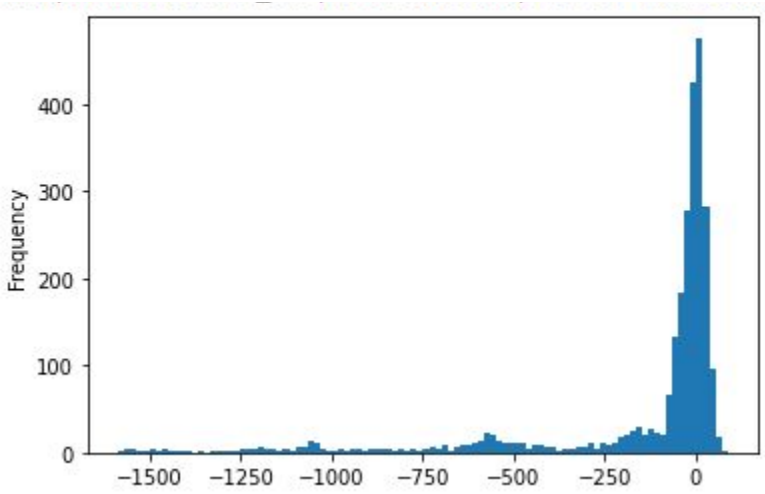
Bot 1 - Results

```
articles,reco = bot1(dt_corpus,trial_user_rating_df)
data.loc[articles]
```

	Category	Sub-Category	Headline	Entire_News	Author
37	general	politics	After Punjab, Congress Now Faces Crisis In Har...	After Punjab, now Congress is divided in Harya...	Swagata Banerjee
29	general	politics	Chandigarh Row: Haryana CM Manohar Lal Khattar...	As the feud between Haryana and Punjab over Ch...	Megha Rawat
249	business	india-business	Vodafone Raises Stake In Vodafone Idea To 47.61%	British telecom major Vodafone has raised its ...	Press Trust Of India
248	business	india-business	Removing Speed Breakers From Path Of Investors...	Pune, Apr 4 (PTI) The Maharashtra government i...	Press Trust Of India
2630	sci-tech	technology	Twitter rival Koo begins voluntary self-verifi...	Indian microblogging site Koo on Wednesday, A...	Yuthika Bhargava
491	sci-tech	gaming	PlayStation Plus Free Games For April 2022 Lea...	Every month, PS Plus subscribers get a new col...	Shikhar Mehrotra
603	sports	football-news	What Is FIFA World Cup 2022 Mascot? Name, Mean...	The draw for the FIFA World Cup 2022 took plac...	Suraj Alva
552	sports	ipl-2022	Gujarat Titans Vs Delhi Capitals, IPL 2022 Hig...	Gujarat Titans win by 14 runs as Lockie Fergus...	Vidit Dhawan
927	entertainment	regional-indian-cinema	'KGF Chapter 2' Trailer Becomes Most Viewed In...	Popular actor Yash-starrer KGF Chapter 2 has b...	Adelle Fernandes

Bot 1 - Results

Scoring of News Corpus to serve to New User

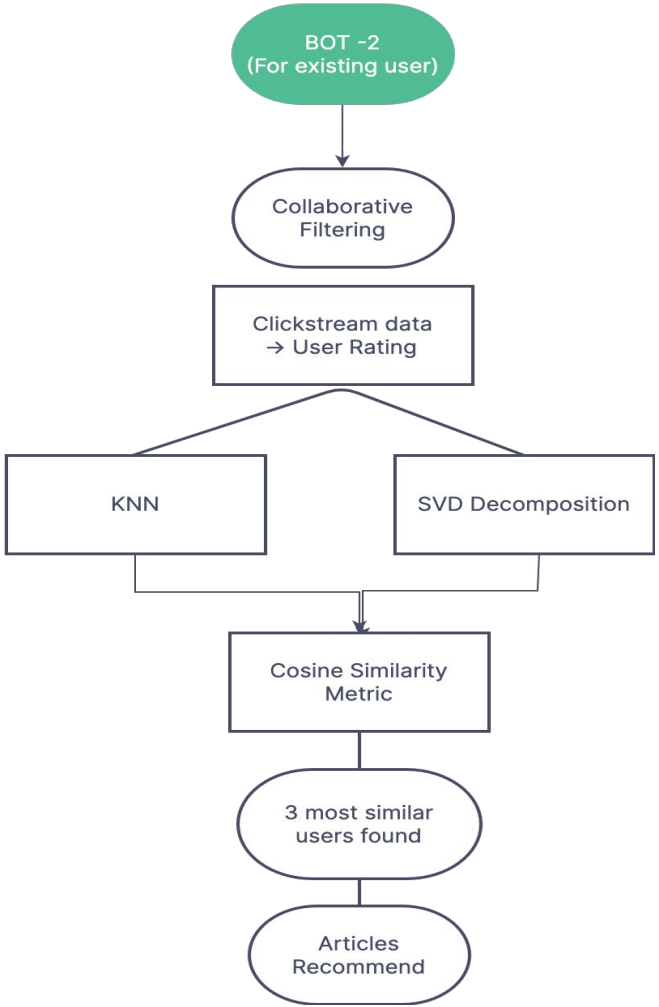


- Graph represents the frequency of articles with a given score
- By scrapping more websites for newer articles, a change in the distribution is expected

Bot -2 (For existing users)

Collaborative Filtering:

- We first prepare a mock data set of user clickstream activity as prescribed in the task.
- This data is used to prepare the matrix of ratings for all articles of the corpus for all users. This is a very sparse matrix.
- From the clickstream data of particular user, we use a KNN approach with n=3 with cosine similarity metric to find similar users.
- From the corpus of articles liked by the KNN users, we randomly select 10 articles to prevent bias and present them to our user.



Bot -2 (For existing users)

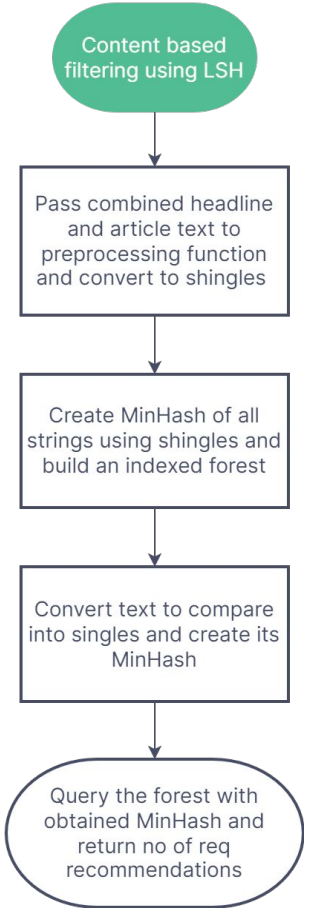
Collaborative Filtering:

- The sparsity of the user ratings matrix could hinter KNN approach.
- We also used a SVD approach to decompose the user ratings to pick up the important features.
- We then find the 3 most similar users to our test user using cosine similarity.
- From the corpus of articles liked by the 3 similar users, we randomly select 10 articles to prevent bias and present them to our user.

Bot -2 (For existing users)

Content Filtering:

- We use two models: LSH and TF-IDF transformation. Locality-Sensitive Hashing involves applying a hashing function on shingled data to reduce its dimensionality. Then nearest neighbour searches are made on the hashed data using Jaccard similarity.
- First we preprocess article text by lowercasing, lemmatizing, and removing stop-words and convert it into unigram shingles. A Minhash signature forest is made of all the text. To find articles similar to an individual one, its Minhash is created and the forest is queries to return a fixed number of recommendations.



Bot -2 (For existing users)

Content Filtering:

- We plan to use HashingVectorizer and TF-IDF tools from Scikitlearn. HashingVectorizer makes working on the vectors faster and more scalable, and TF-IDF punishes more frequent words, thus giving us an ability to truly compare the keywords of two articles.
- We plan to present the user with a set of articles closest to the ones they have read based on similarity scores.

Bot2 - Results

```
Headline of article to compare:
"Sensex drops 200 points, Nifty down to 17,992 amid inflation worries"
Top 10 recommended articles:
2752       Sensex drops 380 points, Nifty slips to 17,044 | Check top performing stocks
2690              Sensex rises 1000 points, Nifty nears 18,000 | Top performing stocks
1890                          Sensex tumbles over 430 points; Nifty slips below 18,000
2724          Sensex rises 700 over points, Nifty closes at 17,498 | Check top gainers
1925        Sensex, Nifty bounce back to close 1% higher on gains in RIL, Infosys, TCS
1895                     Sensex rallies over 1,300 points; Nifty recaptures 18,000 mark
1004   Indian equities extend losses; Sensex down over 300 pts as banks' strike begins
1008                    Indian stocks trade in negative zone, Sensex down by 300 points
2742                 Sensex rises 230 points, Nifty up at 17,222 | List of top gainers
2715                              Sensex tumbles over 100 points, Nifty down at 17,464
```

API

- As of 15/04/2022, We are working on the website
- The website will have logins for old user and sign up feature for new users
- It will serve the users articles based on clickstream data, depending upon type of user

Potential sources of error

There might be errors in categorization of news

- Recommendations to a new user will depend on the initial 10 articles shown to him. The next set of articles will be shown depending on the interaction with this initial set.
- If none of these 10 articles interest the user, the recommendations may be inaccurate. To overcome this, displaying random articles in between the usual feed would be helpful. We have tried to use random articles in between as much as possible
- Hashing data may lead to hashing collisions and loss in comparison accuracy for content based filtering. Optimizing hash size for reduced time and increased accuracy helps overcome this.

Future Plans

- Continue the work on creating bot for existing users
- Create API
- Decrease recommendation time
- Add more news websites to the data set
- Introduce ‘Discover’ section to suggest articles out of users’ interest


That's all Folks!