Building a News Recommender for Jhakaas News Vala

IDC 410 : Coursework
Date : 03/04/2021

Abstract : The task was to build a news recommender app for the company JhakaasNewsVala. The target audience is of working professionals between 21-40 years of age. The objective is to increase retention of the customers by impressing them at the first visit itself. This could be done only if the top 10 stories shown to them are of their interest. The problem is divided into building two intelligent bots- the article recommender and the user profiler.

Introduction :

Every other application we use or website we visit uses some kind of a recommending system to give us a personalised experience. Creating user specific playlists in music apps, suggesting products to a user at an E-store, giving personalised ads are among the many places where recommender systems are put to use. For this purpose, the website and applications store some of our data which includes the searches we have made, links we have visited, locality we live in, etc. The research in the direction to develop a recommender system came to the global platform with the Netflix Prize during 2006- 2009.

An example from our daily life is: when you search for a product on Amazon, you start getting ads related to similar products on Instagram. This is a basic example of a recommender system connecting the two apps two feed each other in order to make the ads shown to us more relevant. Recommender systems can be of vital importance to companies as they could help multiply their income or stand out from their competitors.
There are several different approaches to design a recommender system like content based filtering, collaborative filtering, hybrid filtering,reinforcement learning, session-based systems, mobile systems, risk-aware systems, etc.

The objective is to make a news recommender system that recommends news articles to the user based on the user's interests and relevance to get the maximum click throughs and user-retentivity.

From the perspective of the company, we want to maximize the pages viewed per session by a consumer and increase the length of each session. From the perspective of the customer, we want to serve the content based on the customer's interests to minimize the cognitive load. From the perspective of the app, we want to reduce over-customisation and increase the content consumption equally among the customers.

Approach :

Our approach was to design a hybrid recommender that integrates both content-based filtering and collaborative filtering. This approach would help us alleviate the drawbacks of both the methods and take advantage of the positives.

Content based filtering: This approach uses the features of the items liked by the user to rate other items by comparing them with the ones he has already liked. The clickstream and item consumption time of a user tells the rating of a particular item. In order to extract the features of a document like a news, a commonly used algorithm is tf-idf vectorization.

TF-IDF score consists of two terms the term frequency tf(*t,d*) and the inverse document frequency idf(*t,D*). It is helpful in gauging the importance of a particular word in the document. It is better than a simple "bag of words" representation as it penalizes more common words even if their frequency is higher.

$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}},$$

where $f_{t,d}$ is the *raw count* of a term in a document,

$$\mathrm{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

with

- N: total number of documents in the corpus $N = |D|$
- $|\{d \in D : t \in d\}|$: number of documents where the term t appears (i.e., $\mathrm{tf}(t,d) \neq 0$). If the term is not in the corpus, this will lead to a division-by-zero. It is therefore common to adjust the denominator to $1 + |\{d \in D : t \in d\}|$.

Based on the tf-idf score, a word weighted vector is created for each document. Depending on the user's past history we can recommend new news which matches with the user's interest. The main disadvantage of content based recommenders is that it tends to saturate and recommend the same type of articles. Collaborative recommenders can solve this problem of not having variety in the news.

Collaborative filtering: This approach exploits the similarity between users to recommend items which one likes to the other. The dot product of user vectors is used to calculate similarity between the two users, higher the product, higher is the similarity. Matrix factorization is another better approach to implement collaborative filtering. It involves predicting the unknown rating from the user-document matrix using optimization algorithms like stochastic gradient descent and ALS. This approach is based on a more universal set rather than just the user alone but the

problem of cold start remains. We cannot recommend a new document as there is no data of other users for that particular document.

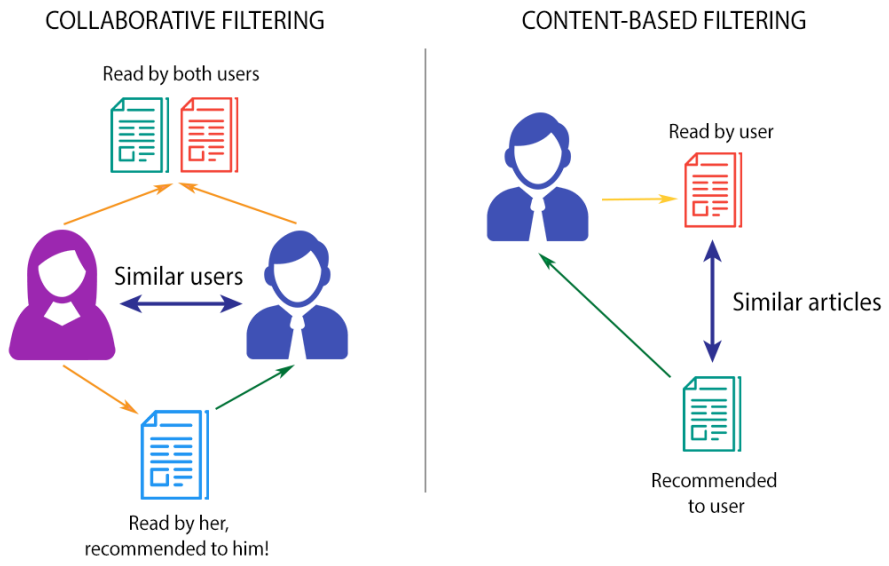

Figure 1

We worked on a hybrid of both content-based and collaborative recommender systems.

Progress :

At first we scrapped the news body from various news websites like NDTV, India Today, Reuters, etc and combined it into one single dataframe.

Preprocessing of data:
We processed the news articles by removing capitalization, numbers, stop words, dropping empty rows and duplicates. Then we applied lemmatization using nltk library on data which brings context to the words for better analysis. Any news having Hindi/Gujrati and other non-ASCII characters were also removed.

Raw news data

	Content
0	Media reports about Swedish bus manufacturer S...
1	Access to COVID-19 vaccines, cooperation on te...
2	After severe criticism over not holding consul...
3	Former Congress president Rahul Gandhi on Thur...
4	The Enforcement Directorate has attached three...
...	...
4589	Over 200 Mughal-era gold coins, dating back to...
4590	China is planning to spend big in Tibet as it...
4591	The Supreme Court Tuesday came out with a solu...
4592	Indian-American Maju Varghese, who previously ...
4593	The Election Commission on Tuesday ordered the...

Pre - processed news

	Content
0	medium report about swedish manufacturer scani...
1	access covid vaccine cooperation technology cl...
2	after severe criticism over holding consultati...
3	former congress president rahul gandhi thursda...
4	enforcement directorate attached three immovab...
...	...
4589	over mughal gold coin dating back early centur...
4590	china planning spend tibet five year plan allo...
4591	supreme court tuesday came with solution stale...
4592	indian american maju varghese previously serve...
4593	election commission tuesday ordered removal vi...

Figure 2 : Preprocessed data visualization

Tf-idf matrix:

	aabad	aadarsh	aadat	aadhaar	aadhar	aadhi	aadmi	aage	aajtak	aakash	aaksha	aamir	aamk	aandolan	aandolanjivi	aane	aapada	aapko
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4589	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4590	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4591	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4592	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4593	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 3 : TF-IDF matrix

Topic modelling with LSA/LDA :

Then we allocated Tf-Idf scores to each article and applied latent semantic analysis (LSA) using sklearn's truncatedSVD to generate 25 latent topics. On applying LDA(Latent Dirichlet Allocation), the recommended news articles to users were of similar nature which is why we switched back to LSA. But if we had more time we would have tried applying LDA first and then truncatedSVD (LSA) over the topics and checked how efficiently the variance of the data is captured.

```
Topic 0:  ['covid', 'case', 'vaccine', 'state', 'farmer', 'minister', 'india', 'government', 'vaccination', 'health']
Topic 1:  ['covid', 'vaccine', 'case', 'death', 'vaccination', 'health', 'total', 'dose', 'reported', 'active']
Topic 2:  ['election', 'assembly', 'party', 'congress', 'vaccine', 'seat', 'poll', 'minister', 'bengal', 'west']
Topic 3:  ['vaccine', 'vaccination', 'farmer', 'dose', 'first', 'covid', 'law', 'modi', 'dos', 'covaxin']
Topic 4:  ['farmer', 'law', 'farm', 'congress', 'party', 'death', 'protest', 'price', 'total', 'hour']
Topic 5:  ['court', 'vaccine', 'supreme', 'justice', 'vaccination', 'petition', 'election', 'bench', 'hearing', 'dose']
Topic 6:  ['police', 'vaccine', 'dose', 'farmer', 'arrested', 'vaccination', 'bengal', 'west', 'accused', 'delhi']
Topic 7:  ['sabha', 'price', 'rajya', 'budget', 'session', 'house', 'congress', 'opposition', 'parliament', 'party']
Topic 8:  ['lakh', 'price', 'sabha', 'bengal', 'west', 'fuel', 'petrol', 'crore', 'rajya', 'fatality']
Topic 9:  ['woman', 'minister', 'international', 'child', 'occasion', 'gender', 'scheme', 'education', 'march', 'pradesh']
Topic 10:  ['congress', 'gandhi', 'leader', 'party', 'minister', 'modi', 'lakh', 'vaccine', 'prime', 'rahul']
Topic 11:  ['rawat', 'minister', 'uttarakhand', 'crore', 'state', 'chief', 'banerjee', 'singh', 'fire', 'bengal']
Topic 12:  ['bengal', 'banerjee', 'west', 'mamata', 'woman', 'nandigram', 'kolkata', 'hour', 'case', 'recovery']
Topic 13:  ['lakh', 'rawat', 'sabha', 'kashmir', 'singh', 'uttarakhand', 'army', 'jammu', 'rajya', 'force']
Topic 14:  ['price', 'petrol', 'kashmir', 'diesel', 'dose', 'jammu', 'congress', 'fuel', 'gandhi', 'army']
Topic 15:  ['minister', 'prime', 'modi', 'tamil', 'delhi', 'nadu', 'budget', 'india', 'narendra', 'court']
Topic 16:  ['crore', 'dose', 'hcws', 'flws', 'dos', 'company', 'beneficiary', 'bank', 'defence', 'accused']
```

Figure 3.1 : Top 10 words inside each latent topic

	Docs	topic_0	topic_1	topic_2	topic_3	topic_4	topic_5	topic_6	topic_7	topic_8	topic_9	topic
0	medium report about swedish manufacturer scani...	0.129416	-0.031665	-0.002044	0.016616	-0.058626	-0.036387	-0.018073	0.009407	0.001573	-0.020310	0.011
1	access covid vaccine cooperation technology cl...	0.241852	-0.033735	0.076402	0.116281	-0.063050	-0.107387	-0.079549	-0.059717	-0.053851	-0.141224	-0.00
2	after severe criticism over holding consultati...	0.187095	-0.047153	-0.003871	0.027602	-0.070254	-0.018248	-0.070942	-0.019196	-0.003131	-0.022006	-0.01
3	former congress president rahul gandhi thursda...	0.177060	-0.026037	0.029299	0.010233	-0.039886	-0.040989	-0.059030	0.027473	-0.033055	-0.081083	0.101

Figure 4 : Documents vs Latent topics

Creating Users and assigning ratings:

We fitted GMM for each topic with 10 components serving as our ratings. Then we generated a user-topic matrix using the fitted GMM based values for each topic for about 50 users. After taking the cosine similarity for every user with a random set of documents we generated ratings matrix. We found this approach better than randomly assigning ratings to the documents or latent topics.

News

Users

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	6.0	10.0	NaN	8.0	5.0	4.0	NaN	1.0	4.0	NaN	3.0	6.0	NaN	NaN	NaN	NaN	5.0
1	3.0	8.0	NaN	5.0	4.0	4.0	9.0	2.0	NaN	NaN	NaN	NaN	NaN	1.0	NaN	3.0	NaN
2	6.0	NaN	8.0	7.0	5.0	NaN	NaN	NaN	NaN	7.0	NaN	7.0	NaN	3.0	2.0	7.0	7.0
3	NaN	NaN	NaN	NaN	6.0	8.0	8.0	5.0	6.0	3.0	3.0	NaN	NaN	NaN	4.0	9.0	NaN
4	NaN	8.0	9.0	7.0	5.0	7.0	8.0	NaN	NaN	NaN	5.0	NaN	NaN	NaN	1.0	6.0	NaN
5	3.0	NaN	3.0	NaN	NaN	NaN	4.0	6.0	3.0	3.0	4.0	9.0	2.0	NaN	1.0	NaN	6.0
6	5.0	9.0	5.0	6.0	4.0	NaN	10.0	NaN	NaN	7.0	NaN	0.0	NaN	NaN	NaN	8.0	6.0
7	NaN	NaN	NaN	5.0	5.0	10.0	NaN	NaN	NaN	NaN	5.0	6.0	4.0	8.0	5.0	9.0	5.0
8	2.0	9.0	NaN	6.0	0.0	4.0	8.0	NaN	7.0	4.0	5.0	9.0	2.0	NaN	NaN	NaN	6.0
9	NaN	NaN	2.0	NaN	7.0	NaN	NaN	6.0	NaN	7.0	NaN	0.0	6.0	2.0	6.0	1.0	3.0
10	5.0	5.0	9.0	7.0	NaN	1.0	NaN	5.0	8.0	NaN	5.0	10.0	7.0	2.0	NaN	2.0	5.0

Figure 5 : Rating matrix

Content based recommender :

Content based recommender chooses the top 10 documents similar to top 5 highest rated documents from the rating matrix based on cosine similarity.

```
selected_docs_content,selected_docs_content_with_ID = content_recommender(rank_matrix,cos_sim)
selected_docs_content
```

	0	1	2	3	4	5	6	7	8	9
0	MoscowIndia s strategic ties with Russia has v...	The first Quadrilateral Security Dialogue ...	The British parliament on Monday discussed fa...	Mumbai Maharashtra India March 8 ANI ...	New Delhi India March 7 ANI Union Minis...	WASHINGTON Reuters U S President Joe Bid...	External Affairs Minister S Jaishankar will on...	The Indian High Commission in London has cond...	Mumbai Maharashtra India March 10 ANI ...	New Delhi India March 7 ANI Prime Minis...
1	The British parliament on Monday discussed fa...	The daily new coronavirus COVID 19 cases fe...	New cases of coronavirus infection in India w...	Maharashtra Kerala Punjab Tamil Nadu Gujar...	Eighteen States UTs including Assam Rajasth...	The Indian High Commission in London has cond...	Daily COVID 19 cases in India registered an i...	The daily new coronavirus COVID 19 cases fe...	Maharashtra Kerala Punjab Tamil Nadu and G...	With several States in the country continuing ...
2	The British parliament on Monday discussed fa...	Congress leader Rahul Gandhi on Sunday compar...	Farmer leader Rakesh Tikait on Monday took a d...	The Indian High Commission in London has cond...	New Delhi India March 5 ANI Prime Minis...	The Indian High Commission in London has cond...	Urging the Centre not to make the agri laws a ...	More than 850 faculty members of various educa...	The High Commission of India in London has con...	The Union Cabinet on Wednesday approved the Pr...

Collaborative based recommender :

In this we use scipy's nnls to optimize and find the missing ratings from the rating matrix. Top best rated documents are chosen. Random masking of known values was done which acts like a training and testing dataset.

	1	2	3	4	5	6	7	8	9	10
0	Access to COVID 19 vaccines cooperation on te...	Lauding India for its vaccine leadership a t...	The Union home ministry on Monday said 200 ...	The National Investigation Agency NIA has ta...	No new COVID 19 fatalities have been reported ...	Maintaining that the new agricultural laws are...	Finance Minister Nirmala Sitharaman lashed out...	Prime Minister Narendra Modi on Friday said p...	Maharashtra Kerala Punjab Tamil Nadu Gujar...	Maharashtra Kerala Punjab Tamil Nadu Gujar...
1	Rajya Sabha Chairman M Venkaiah Naidu on Mar...	India s COVID 19 tally rose to 1 11 92 088 wi...	The Congress is changing the way it selects ca...	No new COVID 19 fatalities have been reported ...	As the farmers protest completed 100 days on t...	A week after the show of strength in Jammu by ...	Maharashtra Kerala Punjab Tamil Nadu Gujar...	Maharashtra Kerala Punjab Tamil Nadu Gujar...	Maharashtra Kerala Punjab Tamil Nadu and G...	India s COVID 19 cases rose to 1 11 24 527 wit...
2	Access to COVID 19 vaccines cooperation on te...	Prime Minister Narendra Modi and his Japanese...	Discussion on the farm laws by the British Par...	Lauding India for its vaccine leadership a t...	The inclusion of India in the United States s...	The High Commission of India in London has con...	Congress General Secretary Priyanka Gandhi Vad...	Agriculture Minister Narendra Singh Tomar on S...	Maintaining that the new agricultural laws are...	Finance Minister Nirmala Sitharaman lashed out...

Hybrid recommender :

Content based filtering suffers from a lack of variety whereas the collaborative filtering suffers from the cold start problem. So using a hybrid approach, we recommend top 5 articles based on the content based approach and top 5 based on the collaborative approach.

Flask app :

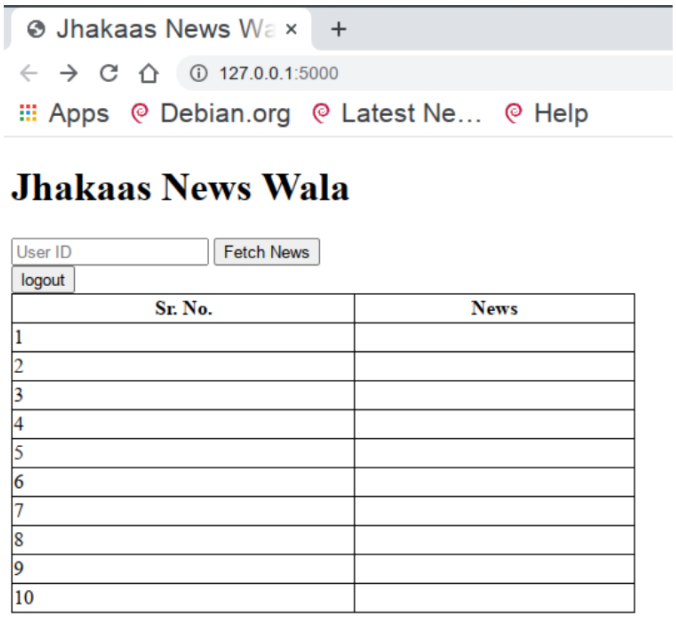

Figure 6 : Home page of our news website.

Whenever the user logs in, top 10 news articles are recommended to him/her using the hybrid approach.

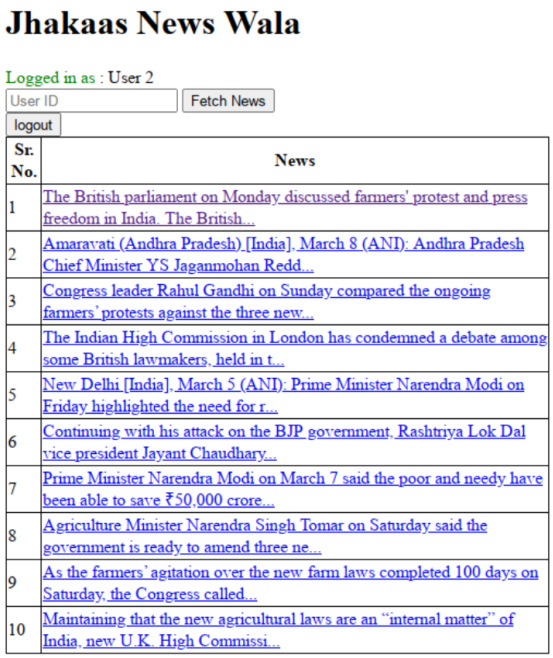

Figure 7 : Recommended articles using hybrid recommender.

As soon as the user clicks on a link, a time counter starts in the user window.
Upon pressing the back button, the total time spent is sent back to the server.

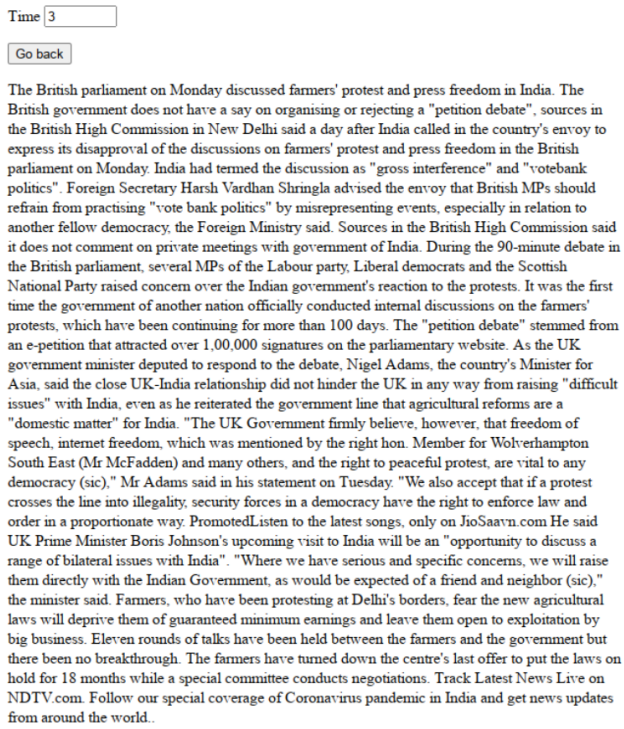

Figure 8 : Opened news article and time counter

The rating of the article is calculated by the time spent and assuming a 200 words per minute reading time. This is used to update the rating of the article in the article rating matrix.
Once the user logs out, the hybrid recommender recalculates the matrices.

Drawbacks & what could have been done better if we had more time ? :

We have finished building a basic recommender system however it is far from perfect. We would like to present few things which we would have done if we had more time :

1. Scrapping more information for the news. The html page contains much more than the news body itself. Display of news in the webpage would have been better had we scrapped the title along with the body. Some news websites also have html tags for location, author, place, etc which would have helped in vectorizing the documents better.
2. Speed up the collaborative recommender system. Currently we are using nnls to optimize the unknown ratings however it is slow and inefficient. Better would have been to implement ALS from pyspark library for matrix factorisation.
3. We can track more things about the user like his/her mac address, ip address, browser fingerprint, operating system (mobile users vs desktop), location, time of login, etc. More

information would help us to build better user profiles. We can find patterns like if a user prefers reading a particular kind of news in one location and a different kind in another.

4. Dynamic updating of the recommender can also be implemented. Currently the recommender only updates the recommender when a user logs out.