News Recommender for “JhakaasNewsWala”

(Course work - 3)
IDC 409 (Machine Learning)

Group members:

1) ***Abhishikta*** - *MS17192*
2) ***Agrima*** - *MS17191*
3) ***Dinesh*** - *MS16067*
4) ***Sarvesh*** - *MS16146*
5) ***Satender***- *MS16069*
6) ***Vishal Kumar*** - *MS17195*

Plan and Progress

Our Objective/ Goals:

(a) Business objective:

(i) Increasing the pages view per session by a user.

(ii) Increasing the session length (Increasing screen time)

(b) Customer Objective:

(i) Serving the preferred news content which would reduce the cognitive load from the readers.

(ii)Increasing the relevance of the news content

Plan and Progress

Our Objective/ Goals (continued):

c) Product Objectives:

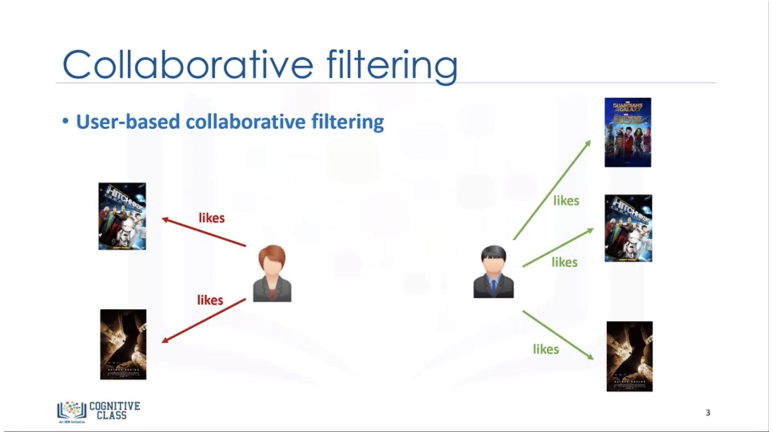
(i) To reduce over customization.

(ii) To increase content consumption equally among the anonymous users, reg

Selecting the right technique:

Both Content based as well as collaborative filtering have their owns set of disadvantages.

We are following the **hybrid model**



Collaborative filtering

- User-based collaborative filtering

Proposed plan:

Building two bots:

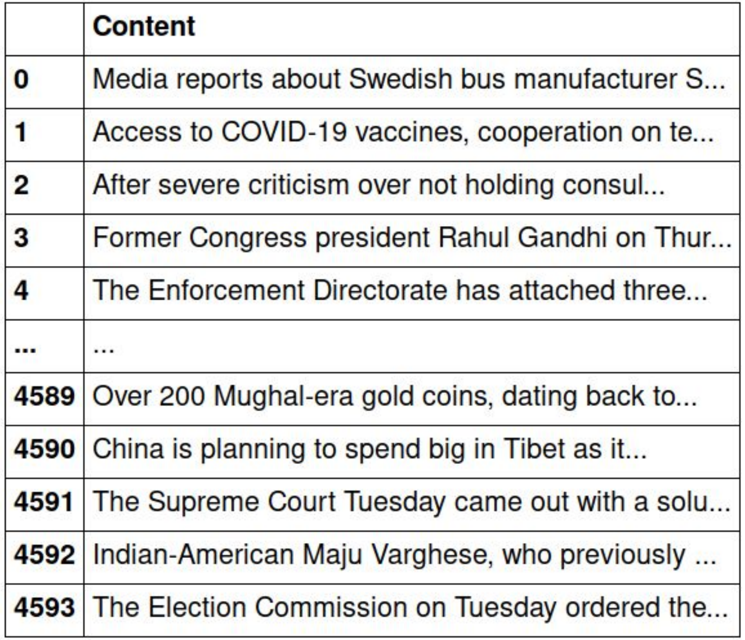
1. Article Recommender
2. User Profiler

The user profiler will be creating user profile based on the clickstream left by users. The article recommender will then be updated with these user profiles.

Preprocessing of the data:

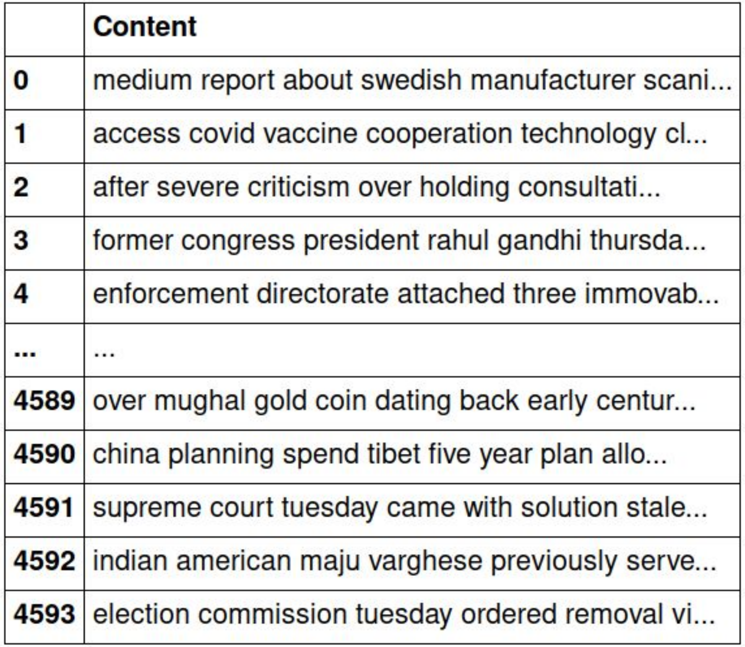
- We scraped the body of news articles from various news websites like The Hindu, Times of India, NDTV, Reuters, India Today, etc.
- We scraped the data using BeautifulSoup library.
- News corpus of all the group members was combined into a single dataframe.
- Then we carried out some basic processing as follows :
  a. removed empty rows
  b. removed punctuation
  c. removed numbers
  d. removed capitalization
  e. removed stop words using nltk library
  f. performed lemmatization.
  g. and non roman scripts like (Gujrati/Hindi) which were leaking from twitter quotes

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...

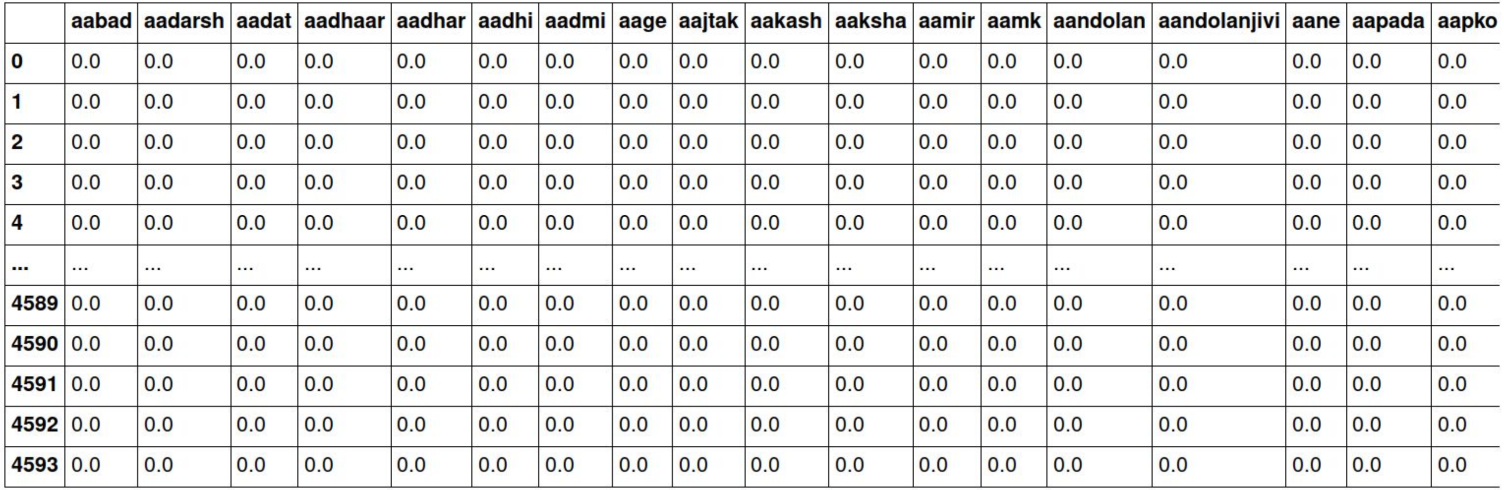
Pre - processed news



	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...

Vectorizing the news corpus

- We used TF-IDF to find the important words in a given document.

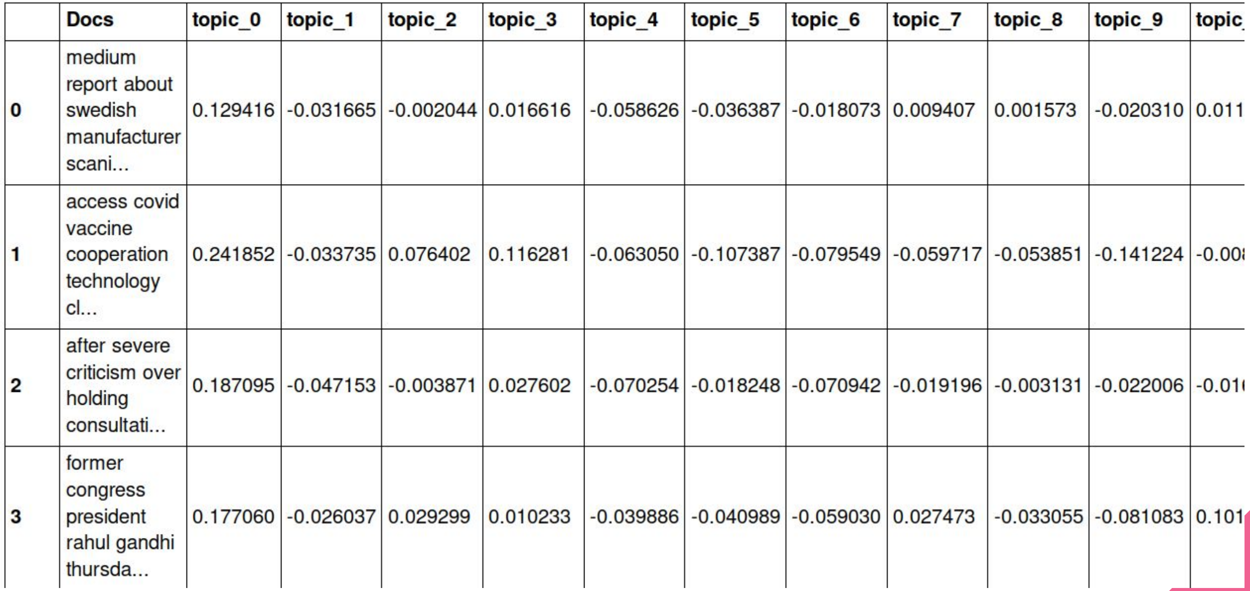


	aabad	aadarsh	aadat	aadhaar	aadhar	aadhi	aadmi	aage	aajtak	aakash	aaksha	aamir	aamk	aandolan	aandolanjivi	aane	aapada	aapko
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	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

- Matrix of 30,000 x 4593

Topic Modelling : Latent Semantic Analysis

- To further reduce dimensions we used sklearn’s truncatedSVD to model 25 topics.



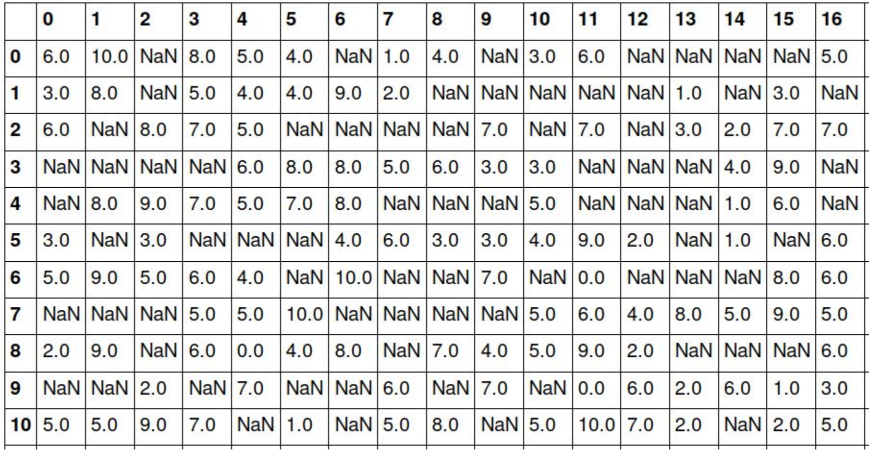
	Docs	topic_0	topic_1	topic_2	topic_3	topic_4	topic_5	topic_6	topic_7	topic_8	topic_9	topic
0	medium report about swedish manufacturer scani...	0.129416	-0.031665	-0.002044	0.016616	-0.058626	-0.036387	-0.018073	0.009407	0.001573	-0.020310	0.011
1	access covid vaccine cooperation technology cl...	0.241852	-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

Generating Initial user ratings (Rating matrix)

- We fitted 10-component GMM to each topic column.
- Then about 50 user profiles were generated using these GMM.
- Found rating for random documents using cosine similarity.

News

Users

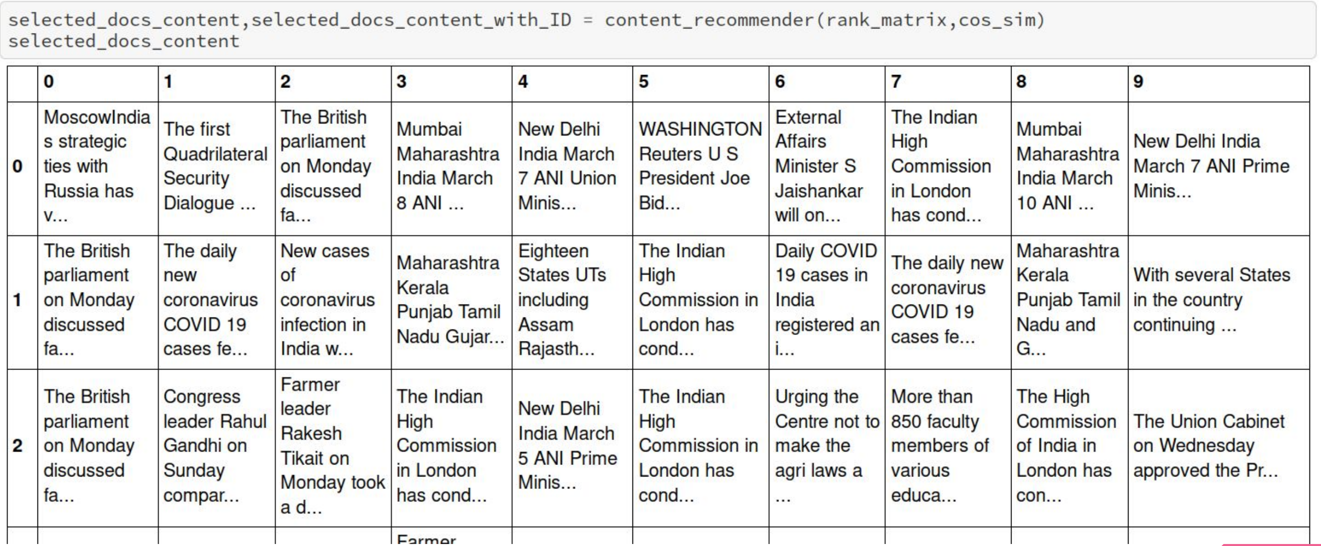


	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	6.0	10.0	NaN	8.0	5.0	4.0	NaN	1.0	4.0	NaN	3.0	6.0	NaN	NaN	NaN	NaN	5.0
1	3.0	8.0	NaN	5.0	4.0	4.0	9.0	2.0	NaN	NaN	NaN	NaN	NaN	1.0	NaN	3.0	NaN
2	6.0	NaN	8.0	7.0	5.0	NaN	NaN	NaN	NaN	7.0	NaN	7.0	NaN	3.0	2.0	7.0	7.0
3	NaN	NaN	NaN	NaN	6.0	8.0	8.0	5.0	6.0	3.0	3.0	NaN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

Rating matrix

Content Based Recommender

- For every user we take the best 5 rated news from the rating matrix and find two similar ones to each one of them using ***Cosine Similarity.***

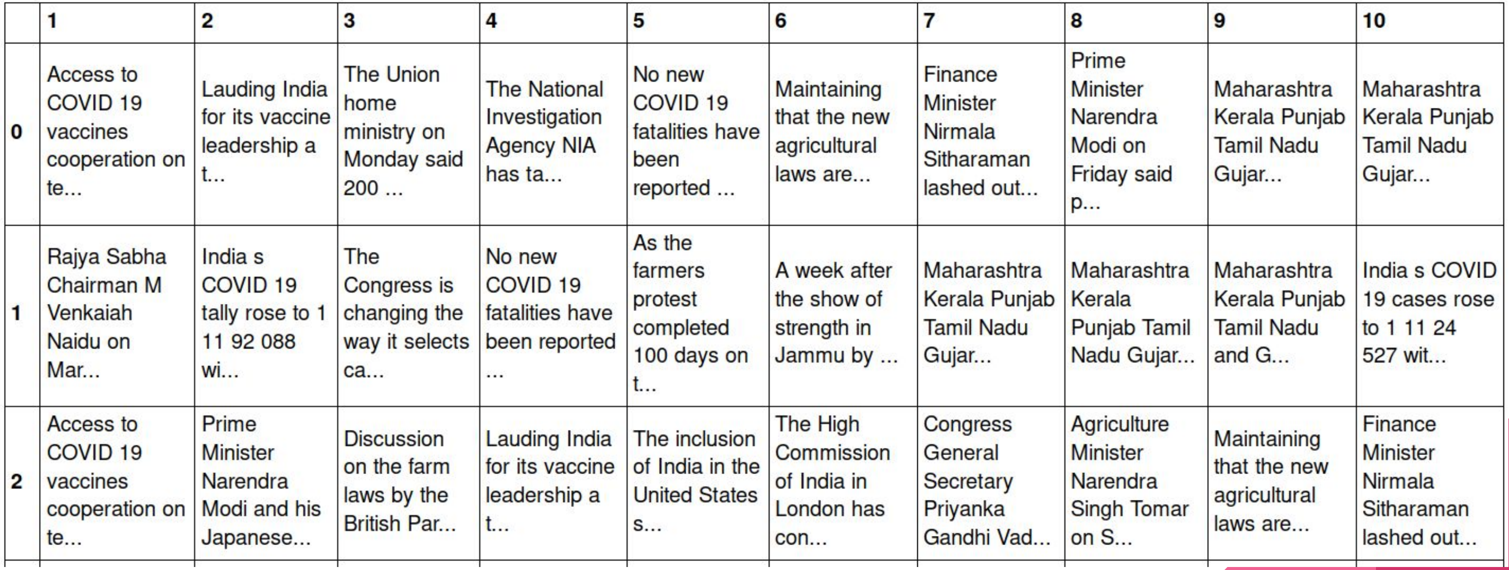


```
selected_docs_content,selected_docs_content_with_ID = content_recommender(rank_matrix,cos_sim)
selected_docs_content
```

	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...
				Farmer						

Collaborative recommender

- We found the missing entries using matrix factorization and ALS optimization.
- Subset selection was done.



	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 recommendation suffers from lack of variety.
- Collaborative recommenders cannot be used to recommend new articles.
- Therefore we are using a hybrid approach.
- Out of the 10 news, our recommender show 5 from the content based recommender and the rest 5 from the collaborative recommender.

Flask App & User Profiler

- User profiler bot is integrated inside the flask app.
    a. Whenever the a user logs in, 10 best news are recommended to him/her using the hybrid recommender.
    b. After the user opens a link a counter starts in the client's browser counting the time.
    c. Once the user presses the back button, the total time is sent back to the server.
    d. Assuming 200 WPM average reading speed, we update the rating of the article in the rating matrix.
    e. Once the user logs out, the hybrid recommender re-evaluates the recommendations.

Next up

- Optimize the hyperparameters like no. of topics.
- Implement ALS for matrix factorization.
- Speed up the collaborative recommender.