

NLP

Final Report

A Report on Assignment

Submitted in Partial Fulfillment of

Building a News Recommender for Jhakaas News Vala

for the Degree of

Bachelor of Technology

In

Computer Science Department

By

Chadalawada Amarnath-1800208C203

Narapareddy Kushal Reddy-1800243C203

Kosana Sai Venkata Pavan Kumar-1800230C203

Bollavaram Golla Riteesh Ram Chander-1800206C203

CSE

3rd year

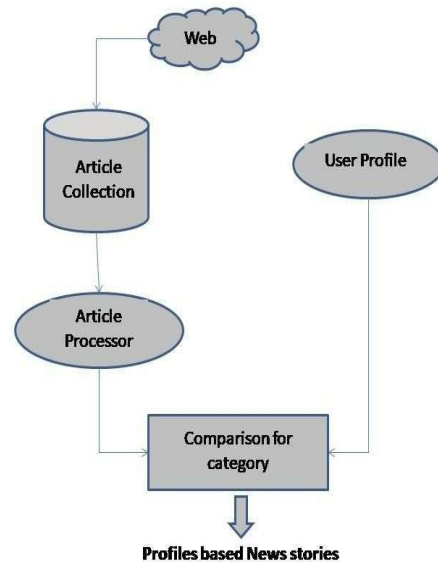


**BML MUNJAL
UNIVERSITY™**

SCHOOL OF ENGINEERING AND TECHNOLOGY

BML MUNJAL UNIVERSITY GURGAON

March,2021



Implemented Strategy

Data Collection:

Through web scrapping of different news websites around 1138 news articles are collected. This scrapped data consists of information like **date of creation, headlines and content of news articles**.

Preprocessing of Data

To get proper data, scrapped data is cleaned by removing punctuations, numbers, special characters and multiple spaces by using regular expressions. After this, data is converted to pandas dataframe to make it structured as shown in figure below

	id	Headlines	Content	date
0	0	India so far contributed over USD 1 million t...	India has till date contributed over USD 1 mi...	PUBLISHED ON APR 08 2021 06:52 AM IST
1	1	AstraZeneca serves legal notice to Serum Inst...	AstraZeneca, the developer of the coronavirus...	UPDATED ON APR 08 2021 06:43 AM IST
2	2	Karan Patel angry with partial lockdown: 'Stu...	Actor Karan Patel has questioned the partial ...	PUBLISHED ON APR 08 2021 06:39 AM IST
3	3	Refund passengers who cancelled flights in lo...	The ministry of civil aviation (MoCA) on Wedn...	PUBLISHED ON APR 08 2021 06:16 AM IST
4	4	Covid-19 prevalence in UK dropped in March, b...	The prevalence of Covid-19 infections in Engl...	PUBLISHED ON APR 08 2021 06:11 AM IST

Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation is used to find the hidden topics in news articles. It is an example of topic model. It builds a topic per document per document model and words per topic model, modeled as Dirichlet distributions. Before doing LDA some preprocessing of text is done like every NLP task

Steps in Preprocessing of Text:

- 1. Tokenization:** Text is split down into sentences and sentences into words.
- 1. Removal Stops words:** Words that add no meaning to sentence and repeat many times are called stop words. These words are removed from text to build good models.
Eg: The, is, to.

2. **Lemmatization:** Words are converted to their root forms. This reduces the redundancy of words. words in third person are changed to first person and verbs in past and future tenses are changed into present.
3. **Stemming:** This technique is used to get base form of respective words by removing affixes from them.

Vectorization of Documents:

After preprocessing of text, vectorization is required since machine can't understand text and only understands numbers. Vectorization is done using 2 techniques and the best one is chosen

1. **Bag of words:** It converts the text into bag of words, which keeps count of total occurrences of most frequently used words. It doesn't consider word sequences and meaning of words. It doesn't differentiate important words and ordinary words.
2. **Tf-idf Vectorization:** This method calculates tf-idf value of each word in the document using term frequency and inverse document frequency of respective words. It tells how important the respective word in that document.

LDA is done by passing Bag of words or Tf-idf vectors of documents

LDA Methodology:

M - number of documents

N - number of words in a given document.

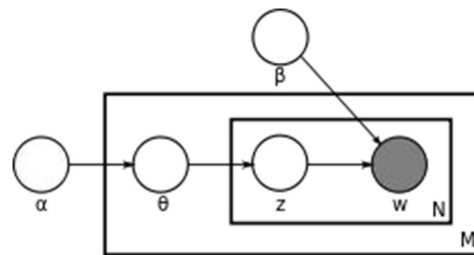
α - parameter of the Dirichlet prior (document topic distributions)

β - parameter of the Dirichlet prior (topic word distribution)

theta - topic distribution for document

z - topic for the word in document

w- is the specific word.



$$p(\theta|\alpha) = \frac{\Gamma \sum_{i=1}^k \alpha_i}{\prod_{i=1}^k \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} \theta_2^{\alpha_2-1} \dots \theta_k^{\alpha_k-1}$$

FIG1.1-DIRICHLET DISTRIBUTION

Above figure is the Probability Density function (PDF) of the Dirichlet distribution it can be viewed as conjugate prior of multinomial distribution. In this, for the generation of the word in a document ,let us say we have **M** number of documents in a corpus , let **alpha** be the hyper parameter for document topic distribution as it helps us in selecting a document with respect to the topics it can be a multinomial or Bernoulli distribution later after selection of a document , within a document **theta** will be multinomial distribution within the topics and **z** is the selection of specific topic and now **beta** will be another hyper parameter and it is the topic to word distribution and **w** will be the specific word selection from beta

distribution. Here **alpha**, **beta** were the Dirichlet parameters. Dirichlet is used here as for a document there will be more than one topic. Let us say an example of sentence as **The tree is in front of the building and behind a car** , now in this sentence the words tree belong to nature and building and car belong to city as we can see that a sentence itself is a mixture of topics so we will need Dirichlet distribution for the distribution of topics in documents and words with respect to topics in that document. Tuning of hyper parameters alpha and beta is done by comparing the generated document with the original document.

$$p(\theta, z, w | \alpha, \beta) = p(\theta | \alpha) \prod_{n=1}^N p(z_n | \theta) p(w_n | z_n, \beta)$$

FIG-1.2: This is a conditional probability of theta topic mix from chosen N topics of z and N words of w (distribution of words) for obtained values of alpha and beta

$$p(w | \alpha, \beta) = \int p(\theta | \alpha) \prod_{n=1}^N p(z_n | \theta) p(w_n | z_n, \beta)$$

FIG1.3: This is the marginal probability of the document from N number of words w

- Here each document is divided into mixture of 20 topics and most probable words for that each topic are discovered.
- LDA is tested with Bag of words and Tf-idf and got better results for Tf-idf. So, LDA using Tf-idf is considered for further steps.

Now, topics are assigned to every news article based on LDA as shown in the figure

FIG1.3: This is the marginal probability of the document from N number of words w

id	Headlines	Content	date	Topic	Merged Content
0	India so far contributed over USD 1 million t...	India has till date contributed over USD 1 mi...	PUBLISHED ON APR 08 2021 06:52 AM IST	[16, 1, 12]	India so far contributed over USD 1 million t...
1	AstraZeneca serves legal notice to Serum Inst...	AstraZeneca, the developer of the coronavirus...	UPDATED ON APR 08 2021 06:43 AM IST	[1, 19]	AstraZeneca serves legal notice to Serum Inst...
2	Karan Patel angry with partial lockdown: 'Stu...	Actor Karan Patel has questioned the partial ...	PUBLISHED ON APR 08 2021 06:39 AM IST	[1, 6, 5]	Karan Patel angry with partial lockdown: 'Stu...
3	Refund passengers who cancelled flights in lo...	The ministry of civil aviation (MoCA) on Wedn...	PUBLISHED ON APR 08 2021 06:16 AM IST	[8, 1, 6, 16]	Refund passengers who cancelled flights in lo...
4	Covid-19 prevalence in UK dropped in March, b...	The prevalence of Covid-19 infections in Engl...	PUBLISHED ON APR 08 2021 06:11 AM IST	[1, 6, 18]	Covid-19 prevalence in UK dropped in March, b...
...
1126	Five more arrested for Mar 27 attack on BJP MLA	BATHINDA The Muktsar police on Monday arreste...	PUBLISHED ON APR 06 2021 01:44 AM IST	[12, 8, 1]	Five more arrested for Mar 27 attack on BJP M...
1127	Experts see no link between Punjab Covid-19 s...	Punjab health department experts say there is...	PUBLISHED ON APR 06 2021 01:41 AM IST	[1, 3, 6]	Experts see no link between Punjab Covid-19 s...
1128	Violation of Covid-19 curbs: Govt seeks	Punjab principal health secretary Husan Lal	PUBLISHED ON APR 06 2021 01:41 AM IST	[1, 14]	Violation of Covid-19 curbs: Govt seeks

Vectorization(Doc2Vec):

Initially while classifying our corpus of news articles into topics with LDA we used Bag of Words and TFIDF vectors. Moreover we need to recommend news articles to our active users so to recommend news articles we used different approaches like Content based filtering, Collaborative and Hybrid filtering etc. In our collaborative methodology we need to represent the documents as vectors of some dimension. So that

would be helpful in creating user-doc matrix and that matrix will be helpful in finding enriched user profile. So to convert documents into vectors we decided to use Doc2Vec. Instead of Doc2Vec we can use Word2Vec but in Word2Vec we have to convert those word vectors of every document as a single document vector and to do that we have to do averaging and summing over all the words so that may cause to lose the semantic and the contextual meaning of the documents. So to overcome this we used Doc2Vec in our approach. In that Doc2Vec we would be using PV-DM which is an extension to CBOW model. Same as CBOW here based on the context words we find the center word but the difference is that we will be inputting paragraph matrix with out context words. In case of doing vectorization we are using Gensim models of Doc2Vec the steps taken are mentioned below:

1. Taking all the documents in to a single list.
2. Mentioning the tags for every document.
3. Initialized the model and built the vocabulary.
4. Finally trained the model.

Finally we done with creating document vectors. The vectors would look like this and the considered vector size of a document is 1131.

```
> 1
  [ 0.07653273 -0.04893406 -0.0406334 ... -0.18526982 0.08646642
    0.08632808]
2
  [ 0.01810286 0.01486883 -0.0049757 ... -0.01611219 0.01998717
    0.00304977]
3
  [-0.08326335 -0.05991447 -0.08997986 ... 0.05433492 -0.02680857
    0.03058095]
4
  [-0.04720877 -0.01497303 0.00198254 ... 0.0017039 -0.00598709
    0.01302787]
5
  [-0.08533983 -0.04860721 0.02012935 ... -0.08096255 0.0384644
    0.02697446]
6
  [ 0.02120355 0.00296967 -0.00096896 ... -0.10363605 0.05139556
    0.04383315]
7
  [-0.02689749 -0.01684707 -0.0197086 ... -0.0058431 -0.00152368
    0.02901435]
8
  [ 0.00569682 0.00800785 -0.03499453 ... 0.0201887 0.00920147
    0.01182081]
9
  [ 0.00057384 -0.02027813 -0.01818098 ... -0.02382979 0.01199658
    0.01654303]
```

User Dummy data creation:

This is stage is one of the major part of our project here we created user profiles of nearly 1000 users. Firstly for every user we randomly assigned some of the interested topics which are derived from LDA. By doing that we can generate user-topic matrix, where for every user the sum of the probabilities of all the topics would be equal to 1.

After creating that user-topic matrix we created doc-topic matrix this matrix will be showing the probabilities of those 20 topics to be appeared in each and every document. Ultimately these two matrices are used in all the three types of filtering approaches in different ways the further more details will be discussed later in those approaches.

Also we created user-doc matrix which will be in 0,1 format where if user read the article its value would be one and zero if not read. Then we generated user-doc-rating matrix where this will be showing the ratings given by user for read articles. Those ratings would be from 0 to 1. Finally we created final-user-doc matrix this matrix is generated by calculating the weighted average sum of users read articles and the weights here are those ratings.

Rating = time spent by the user to read to article/expected time to read the article.

If time spent is greater than or equal to expected time then the rating given would be 1.

Moreover these are introductory steps we done. Till now we done all the preprocessing parts and gathered all the requirements for building our recommendation approaches.

Building the Recommender systems:

Firstly we implemented Content based, after that collaborative and finally hybrid which is the combination of content and collaborative approaches.

1. Content Based: Here the articles need to be recommended based on the content that the user consumed past. So initially we created all the required user profile data like user-topic matrix, this matrix will be showing the users interest on those 20 topics which are derived from LDA. Every user-topic vector will be of (1, 20) dimensions and now we have to find the cosine similarity with the active user profile and with the complete doc-topic matrix these doc-topic vectors are also of (1, 20) dimensions. Hence by finding the cosine similarity with the doc-topic matrix we get the nearby documents with similar topic distribution of active user topics. Therefore we got the cosine similarity values for all the 1131 articles and we extracted top 10 articles

These are the recommended articles based on content filtering for user 3.

536	Andhra cop takes care of one-month-old baby a...
621	Day after Ludhiana factory roof collapse, fac...
26	One more worker dies in Ludhiana factory roof...
1050	Tripura district council polls begin amid tig...
593	Life played a cruel joke, say kin of Ludhiana...
33	HC transfers 367 judicial officers in Punjab,...
590	No relief announced for Ludhiana factory next...
128	Himachal MC polls: Cong bags Palampur, Solan;...
157	Properties get costlier as admin sanctions ci...
968	Galat video: Rubina Dilaik gets fooled in lov...

2. Collaborative Based: In collaborative firstly we find the similar users for our active user and based on the content used by those neighbouring users we recommend articles to our active user. There are two ways to get similar users one is based on topics or second is based on articles. In case of news recommender's the articles will keep on updating so its not a good idea to use articles to find similar users, so we used topics here. By doing the cosine similarity between active user-topic vector and user-topic matrix we can extract the top similar users. In our case we took top 5 similar users. Now the next step is to create enriched user profile. This enriched user profile is calculated by taking the weighted average sum of active and similar users final-user-doc matrix and the weights are those cosine similarities. After getting

that profile we find the similar documents which are generated by Doc2Vec using cosine similarity. Finally we can extract and recommend top 5 articles based on collaborative filtering.

These are top five recommended articles for user 3

```
{567: 0.888200581073761, 721: 0.8631088137626648, 664: 0.85452491044499817, 578: 0.8530427813529968, 488: 0.8527843356132507,
```

3. Hybrid Recommender: Here like content filtering we took user-topic matrix and doc-topic matrix that doc-topic matrix used to find similar documents and like collaborative filtering we found similar users and created enriched user-topic matrix instead of enriched user-doc matrix. Here firstly we took the active users user-topic vector and then found the similar users using cosine similarity, then extracted similar users with similarity score greater than 0.7 and then calculated enriched user profile by doing the weighted average sum of active user and similar users user-topic matrix and the weights are those similarity scores. Finally after getting enriched user profile we done cosine similarity with doc-topic matrix and extracted top 10 articles with high scores.

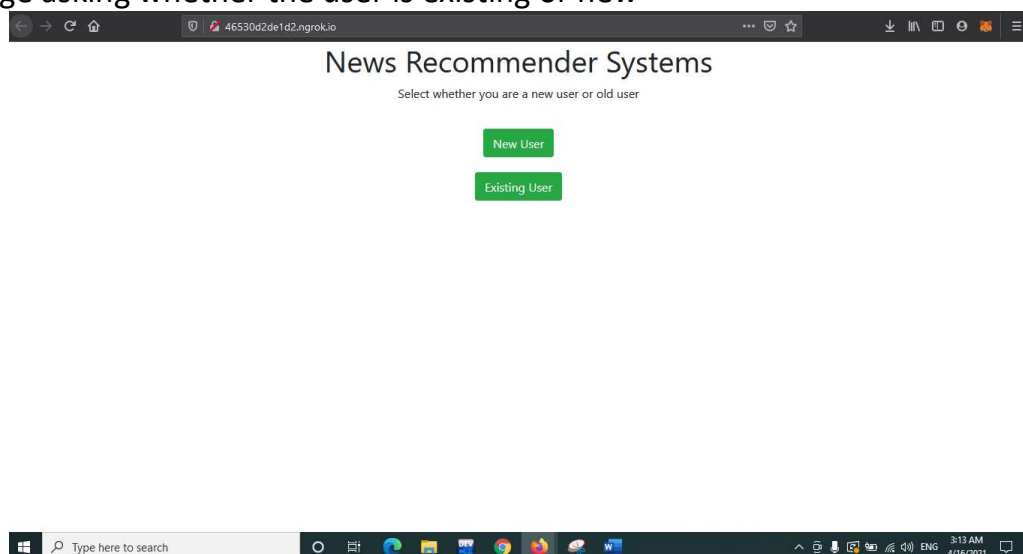
These are the top 10 recommended articles for user 3

```
536      Andhra cop takes care of one-month-old baby a...
1050     Tripura district council polls begin amid tig...
621      Day after Ludhiana factory roof collapse, fac...
26       One more worker dies in Ludhiana factory roof...
593      Life played a cruel joke, say kin of Ludhiana...
865      NEET PG 2021: Application correction deadline...
33       HC transfers 367 judicial officers in Punjab,...
128      Himachal MC polls: Cong bags Palampur, Solan;...
506      JEE Main April 2021 correction window last da...
157      Properties get costlier as admin sanctions ci...
```

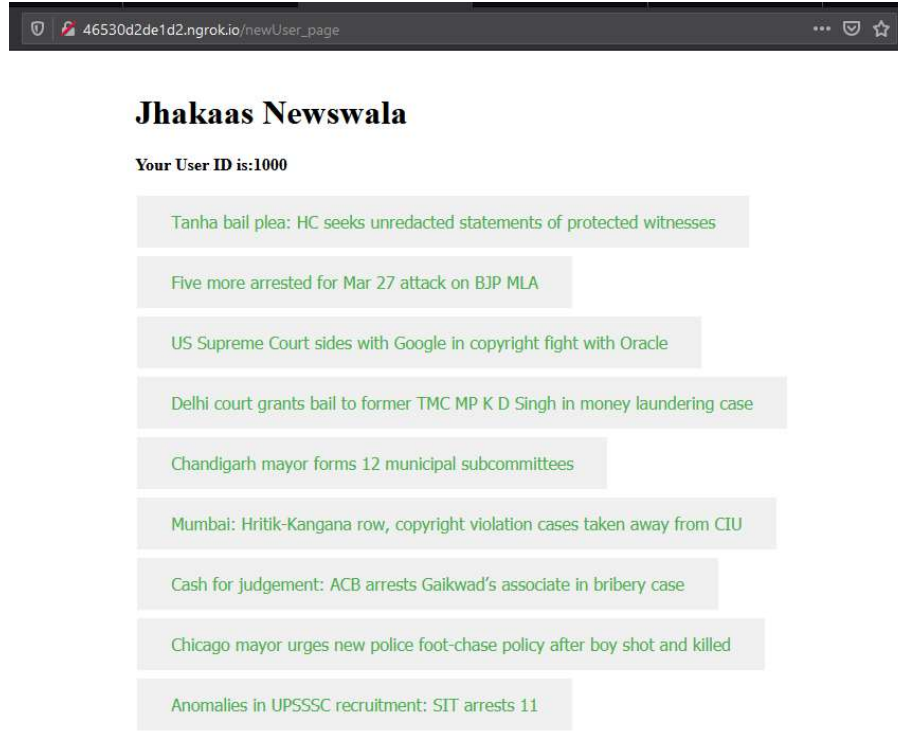
We done with implementing the recommender systems with these three approaches and we selected Hybrid recommender's to deploy in our flask application that we built because that is more optimal than the other two. Finally these are the things that we have worked.

Outputs:

1)Initial page asking whether the user is existing or new



2) New User: New user Id is generated and 10 news articles headlines are shown.



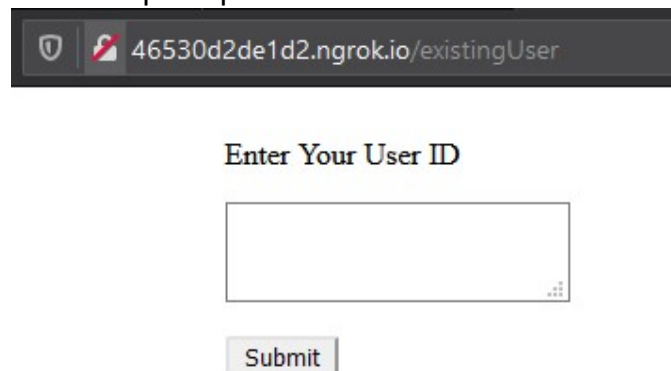
46530d2de1d2.ngrok.io/newUser_page

Jhakaas Newswala

Your User ID is:1000

- Tanha bail plea: HC seeks unredacted statements of protected witnesses
- Five more arrested for Mar 27 attack on BJP MLA
- US Supreme Court sides with Google in copyright fight with Oracle
- Delhi court grants bail to former TMC MP K D Singh in money laundering case
- Chandigarh mayor forms 12 municipal subcommittees
- Mumbai: Hritik-Kangana row, copyright violation cases taken away from CIU
- Cash for judgement: ACB arrests Gaikwad's associate in bribery case
- Chicago mayor urges new police foot-chase policy after boy shot and killed
- Anomalies in UPSSSC recruitment: SIT arrests 11

3) Existing User Id: The user Id is prompted from the user

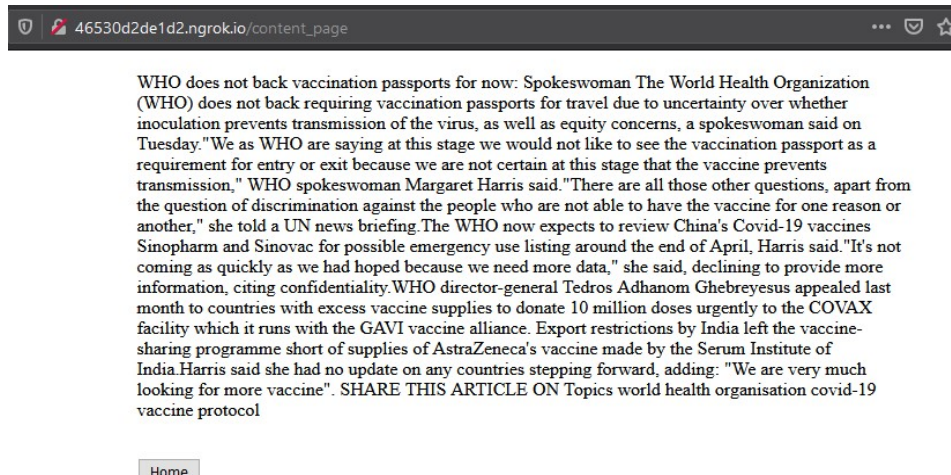


46530d2de1d2.ngrok.io/existingUser

Enter Your User ID

Submit

4) when clicked on the head line of news article:

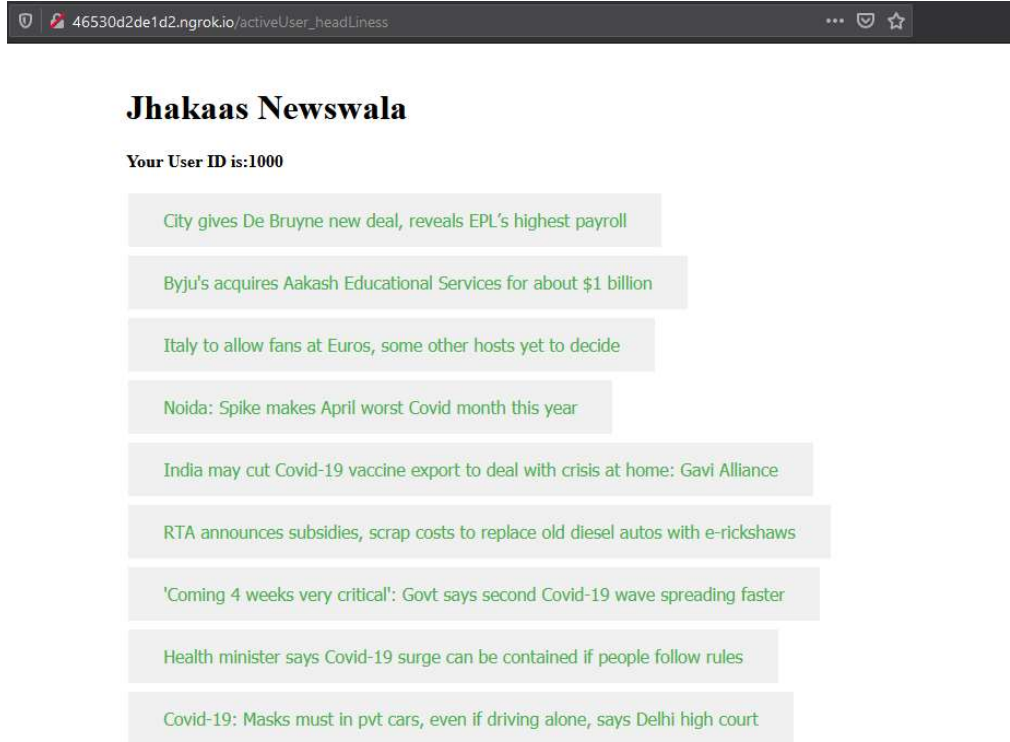


46530d2de1d2.ngrok.io/content_page

WHO does not back vaccination passports for now: Spokeswoman The World Health Organization (WHO) does not back requiring vaccination passports for travel due to uncertainty over whether inoculation prevents transmission of the virus, as well as equity concerns, a spokeswoman said on Tuesday. "We as WHO are saying at this stage we would not like to see the vaccination passport as a requirement for entry or exit because we are not certain at this stage that the vaccine prevents transmission," WHO spokeswoman Margaret Harris said. "There are all those other questions, apart from the question of discrimination against the people who are not able to have the vaccine for one reason or another," she told a UN news briefing. The WHO now expects to review China's Covid-19 vaccines Sinopharm and Sinovac for possible emergency use listing around the end of April, Harris said. "It's not coming as quickly as we had hoped because we need more data," she said, declining to provide more information, citing confidentiality. WHO director-general Tedros Adhanom Ghebreyesus appealed last month to countries with excess vaccine supplies to donate 10 million doses urgently to the COVAX facility which it runs with the GAVI vaccine alliance. Export restrictions by India left the vaccine-sharing programme short of supplies of AstraZeneca's vaccine made by the Serum Institute of India. Harris said she had no update on any countries stepping forward, adding: "We are very much looking for more vaccine". SHARE THIS ARTICLE ON Topics world health organisation covid-19 vaccine protocol

Home

6)When clicked on home button of the content then the head line page is shown again with 10 newly recommended articles based on the last and all previous clicks and similar users.



Things expected but not worked:

Calculating the users ratings to the document: Like we have discussed earlier the ratings should be calculated implicitly by the web application. To calculate that rating we need to find time spent by the user to read an article but we haven't done that part.

Future Works:

1. Calculating those implicit user ratings.
2. Recommend photos of that news along with articles
3. Provide user with more personalized experience