Building a News Recommender for JhakaasNewsVala

Team StrawHat: Nikhil Narkhede(MS17080), Vishal Dafada(MS17090), Gulshan Badole(MS17062), Varun Gautam(MS17010), Ajeet Kumar Singh(MS17056)
Indian Institute of Science Education and Research, Mohali.

CONTENTS

I. INTRODUCTION

In this project we have developed a news recommender system for the target audience of young professionals in the age group of 21-40. This recommender system will also provide a unique reading experience to users with news articles they are interested in by extracting user's interest with the help of the clickstream data.

We needed to overcome some basic challenges in order to get a good news recommender system. These challenges are:

- **Cold start issue for the first time user**: If a user visits the profile for the first time, then we do not have the articles to recommend to the user based on his/her previous clickstream data.
- **User's Bias**: If a user reads news article with a specific ideology, it is common for him to get stuck around the articles related to that ideology. We will try to provide the user with variety of news articles.
- **User's Sentiments**: We try not to hurt user's emotions and therefore we plan to do sentiment analysis.

II. TYPES OF RECOMMENDER SYSTEMS

There are three types of news recommender systems. These are explained in detail below.

A. Content Based Filtering

Advantages of Content Based Filtering:

- It collects the particular interest of the user and recommends items based on the information of items previously liked by the user.
- Recommendations based on the information on previously visited news items using certain key-words.

Challenges in Content Based Filtering technique:

- It is less diverse.
- User's interest can change from time to time.

B. Collaborative Filtering Technique

Advantages of Collaborative Filtering Technique:

- Recommendation of items for a user based on information and reactions of users of similar interest on those items.
- Combines users to create a ranked list of suggestions.
- This model can help users to explore new interest.

Challenges in Collaborative Filtering Technique:

- Cold start problem for new users.
- Sparsity of data
- Requirement of large storage to store information of users and Computation power.

C. Hybrid Filtering Technique

Hybrid Filtering Technique:

- It gives better and optimal recommendations by combining both content based and collaborative based filtering techniques.
- It helps to solve the cold start problem

In our project, we are using Hybrid Filtering technique. Although we are using a quite different approach, we are still trying to solve all the problems faced by the user in a news recommender system.

III. OUR FIRST APPROACH: LDA MODEL

Linear Discriminant Analysis(LDA) is a dimensional reduction technique used as a pre-processing step in Machine Learning and pattern classification applications.

The main objective of LDA is to reduce the dimensions by removing the inessential and dependent features by transforming the components from a higher dimensional space to a space with lower dimensions.

We tried to achieve success in this model by increasing the evaluation score called as Coherence score. We were able to make quite good progress on that but we found the method of Locality Sensitive hashing to be more useful and accurate. So, we used LSH technique later in this project.

we used pyLDAvis library to visualize the data and get better idea of using LDA Model. we generated our LDA model for 10 topics to get the idea of content of the NEWS.

As we can see from the diagram in the Github link at the last of this report, here we are getting some overlapping in topics so we will select optimal model by choosing number of topics.

IV. OUR SECOND APPROACH

A. Named entity recognition(NER)

We are using the Named entity recognition(NER) technique to process and analyse all the web scraped data.

Named entity recognition(NER):- NER represents the technique of chunking, extraction or identification of the particular entities in the data. It helps to identify and categorize critical information in the text. An entity can be any word or series of words that consistently refers to the same thing. Each newly detected entity is classified into a predetermined category.

	Category	Sub_Category	Title	Synopsis	News	tags
0	Sports	Badminton	BWF World Tour Finals: Fighting PV Sindhu lose...	This was PV Sindhu's 16th defeat to Tai Tzu Yi...	World champion shuttler P V Sindhu went down f...	Tai Tzu Ying Taiwan $1.5 million HSBC BWF Worl...
1	Sports	Badminton	World Tour Finals Preview: PV Sindhu, recharge...	With the Indian having played more matches tha...	Carolina Marin (50 total) played 24 tournament...	Carolina Marin PV Sindhu Indian Marin Sindhu O...
2	Sports	Badminton	Satwiksairaj's offence gets neutralised by sav...	Satwiksairaj Rankireddy uses big smash to kill...	One would have to be blind to not figure that ...	Satwiksairaj Rankireddy Neutralising Malaysian...
3	Sports	Badminton	Dream run of indian doubles pairs end with sem...	Up against the world number three Thai pair, S...	The Indian mixed doubles pair of Satwiksairaj ...	Indian Satwiksairaj Rankireddy Ashwini Ponnapp...
4	Sports	Badminton	Satwik-Chirag's impressive run ends with semif...	The Indian pair had participated in Super 1000...	Tokyo Olympics medal contender Satwiksairaj Ra...	Tokyo Olympics Satwiksairaj Rankireddy Chirag ...
...	...	...	...	...	...	...
5918	Entertainment	Box-office-collection	Kesari box office collection Day 8: Akshay Kum...	Kesari box office collection Day 8: Akshay Kum...	Kesari, starring Akshay Kumar in the lead role...	Kesari Akshay Kumar Rs 105.86 crore Taran Adar...
5919	Entertainment	Box-office-collection	Junglee box office prediction: Vidyut Jammwal ...	Junglee box office prediction: Junglee will ha...	Vidyut Jammwal's Junglee has caught the fancy ...	Vidyut Jammwal's Junglee Kesari Notebook Chuck...
5920	Entertainment	Box-office-collection	Luka Chuppi box office collection Day 26: Kart...	Luka Chuppi box office collection Day 26: Krit...	Kriti Sanon and Kartik Aaryan's film Luka Chup...	Kriti Sanon Kartik Aaryan Luka Chuppi Rs 43 la...
5921	Entertainment	Box-office-collection	Badla box office collection day 20: Going gets...	Badla box office collection day 20: Taapsee Pa...	After more than 2 weeks, Taapsee Pannu's film ...	Taapsee Pannu's Badla Badla Rs 1 crore only Rs...
5922	Entertainment	Box-office-collection	Kesari box office collection Day 7: Akshay Kum...	Kesari box office collection Day 7: Akshay Kum...	Akshay Kumar starrer Kesari has become the fas...	Akshay Kumar starrer Kesari Rs 100 crore Gully...

5923 rows × 6 columns

B. Locality-sensitive hashing(LSH)

We are using the technique of Locality-sensitive hashing(LSH) to find out the most relevant news articles to recommend to a user.

Locality-sensitive hashing(LSH):- LSH is a model with the help of which we can sort out the similar input items into the same bucket. Here the number of buckets is small compared to the number of input items. So, we can sort out mostly similar items together in the same bucket and present them in hashtags.

To understand the LSH technique mathematically, let us define a Hash family H.

Now, $P[h(x) = h(y)]$ indicates the probability such that two points $h(x)$ and $h(y)$ in the Hash family H are equal. The Hash family H is locality sensitive if,

$P[h(x) = h(y)]$ is high if x is close to y ,

$P[h(x) = h(y)]$ is low if x is away from y.

The high probabilty indicates that the two points are likely to be included in the same bucket.

The similarity between two points is calculated using the known method of Jaccard similarity. The Jaccard Similarity index is calculated as,

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

Where, $J(A, B)$ is the Jaccard distance between two sets points A and B.

Mathematically, we can write ,

For a Given universe U and similarity $s : U \times U \to [0, 1]$, There exists a probability distribution over some Hash family H such that

$P[h(x) = h(y)] = s(x, y)$

where, $h \in H$ and $s(x, y) = s(y, x)$.

And $s(x, y) = 1 \rightarrow x = y$

We are going to use data scatch library to make LSH model by using MinHash forest you can see the details here in this link. https://github.com/ekzhu/datasketch

We also used MinhashLSH forest to make our model based on NER tags that we have extracted earlier.

Right now, we don't have user data so we will use existing data. This data is from sport section and sub section of football and as we can see the recommendations we are getting some really good results.

```
It took 0.007288932800292969 seconds to query forest.
```

	Category	Sub_Category	Title	Synopsis	News	tags
299	Sports	Badminton	PV Sindhu eyes China Open after World Champion...	PV Sindhu ended India's long wait for a world ...	World champion PV Sindhu will look to reassert...	Indian China Changzhou Sindhu India Basel Swit...
719	Sports	Football	Mount delivers again for Lampard as Chelsea ek...	Four months into the season and Chelsea manage...	Four months into the season and Chelsea manage...	Chelsea Frank Lampard $300 million Mason Mount...
881	Sports	Football	Lionel Messi's salary at Barcelona unsustainab...	Lionel Messi, who sought an exit from Barcelon...	Lionel Messi's salary is too big for Barcelona...	Lionel Messi Barcelona Emili Rousaud Messi Bar...
252	Sports	Badminton	French Open: Satwik/Chirag continue golden run...	French Open: Saina Nehwal suffered a 20-22, 21...	Satwiksairaj Rankireddy and Chirag Shetty stor...	Satwiksairaj Rankireddy Chirag Shetty the Fren...
789	Sports	Football	Lionel Messi returns for Barcelona as Koeman w...	Lionel Messi has stated he will not decide his...	Lionel Messi has recovered from an ankle injur...	Lionel Messi Huesca Barcelona La Liga Eibar Ro...
247	Sports	Badminton	BWF Rankings: Satwik-Chirag reclaim top-10 spo...	PV Sindhu and Saina Nehwal remained static at ...	Indian men's doubles pairing of Satwiksairaj R...	Indian Satwiksairaj Rankireddy Chirag Shetty B...
984	Sports	Football	Lionel Messi pays tribute to Diego Maradona in...	After scoring Barcelona's fourth goal, Lionel ...	Lionel Messi paid a personal tribute to the la...	Lionel Messi Diego Maradona Barcelona Osasuna ...
892	Sports	Football	Lionel Messi fires anxious Barcelona to victor...	The win lifted Barcelona up to eighth in the s...	Lionel Messi dragged a nervous Barcelona to a ...	Lionel Messi Barcelona Levante La Liga Catalan...
797	Sports	Football	As transfer window opens, struggles of Premier...	With the January transfer window opening, Prem...	The high expectations from new recruits aren't...	Tottenham Bale Real Madrid Bale Madrid Harry K...
2206	Sports	Wwe-wrestling	WWE Raw Results: The Boss Sasha Banks returns,...	Sasha Banks reemerged on the Raw after SummerS...	Appearing on WWE programming for the first tim...	WWE WrestleMania Sasha Banks SummerSlam Nataly...

V. CREATING NEW USER PROFILE

Now we have created new and old user profile which we will use to in our recommandation system.

A. New User

Here, we are creating first user interface. The user will see this when he visits for the first time. He will get news from all the section. For the first time visit, we are giving him a choice to select his favorite news categories.

And as we can see, we have selected some sports news and Entertainment news and we are getting recommendations from the same section. so, our recommendation system is working nicely. Even in sports section we've selected cricket and Wwe-wrestling as sub sections and as we can see we are getting recommendations from the same section. This looks wonderful....

```
It took 0.005433082580566406 seconds to query forest.
```

	Category	Sub_Category	Title	Synopsis	News	tags	cat
5792	Entertainment	Box-office-collection	War box office collection Day 15: Hrithik-Tige...	War box office collection Day 15: The YRF acti...	Hrithik Roshan and Tiger Shroff starrer War is...	Hrithik Roshan Tiger Shroff Rs 300 crore Rs 53...	3
2055	Sports	Wwe-wrestling	WWE Hell in a Cell 2020 Live Streaming, Date a...	WWE Hell in a Cell 2020 Live Streaming, Date a...	WWE Hell in a Cell 2020 Live Streaming, Date a...	Universal Championship Jey Uso Cell Jimmy Uso ...	0
649	Sports	Cricket	Navdeep will rise to the occasion if handed a ...	Does the scrawny pacer from Taraori, Haryana, ...	Navdeep Saini is yet to make his Test debut, e...	Navdeep Saini India Umesh Yadav Saini Sydney T...	0
5802	Entertainment	Box-office-collection	War box office prediction: Hrithik Roshan and ...	Releasing in over 4000 screens, Hrithik Roshan...	The Gandhi Jayanti release of the year, War, s...	Gandhi Jayanti War Hrithik Roshan Tiger Shroff...	3
431	Sports	Cricket	Rajasthan Royals release Steve Smith; Chennai ...	Rajasthan Royals have also appointed former Sr...	The IPL player retention/release day doesn't o...	IPL IPL IPL Samson RoyalsRajasthan Royals Stev...	0
3281	Business	Companies	Cox & Kings loaned out Rs 6,071 crore to at le...	The travel firm came under the lens after the ...	A forensic audit of Cox & Kings Ltd, the liste...	Cox & Kings Ltd the Enforcement Directorate th...	1
533	Sports	Cricket	Delhi record second successive win, beat Andhr...	Delhi chased the target with three overs to sp...	Delhi recorded their second straight win in th...	Delhi Andhra Pradesh Elite Group E the Wankhed...	0
2070	Sports	Wwe-wrestling	Triple H: Covid-19 has made this time frame 'i...	Triple H opens up about the difficulties WWE f...	World Wrestling Federation (WWE) has managed t...	World Wrestling Federation WWE WWE Paul Michae...	0
2071	Sports	Wwe-wrestling	Slams, not dunks: WWE replaces NBA at one Flor...	Pro wrestling is replacing pro basketball at t...	Pro wrestling is replacing pro basketball at t...	Florida WWE the Amway Center Orlando NBA The O...	0
606	Sports	Cricket	Syed Mushtaq Ali T20 Trophy 2020/21: Schedule,...	Syed Mushtaq Ali T20 Trophy 2020/21: All you n...	Domestic cricket finally returns to India afte...	India Ranji Trophy Indian Mushtaq Ali T20 Trop...	0

B. Old User

Now we will create a profile for the old user.
We have generated this clickstream data for dummy user.

	Click	UserID	Time_Spent
0	0	1.0	0.000000
1	1	1.0	1.764861
2	1	1.0	0.635004
3	1	1.0	0.000000
4	0	1.0	0.000000
...	---	---	---
149995	1	100.0	8.806663
149996	1	100.0	11.005164
149997	1	100.0	10.909229
149998	1	100.0	17.250826
149999	1	100.0	16.754506

150000 rows × 3 columns

The user profile will look like this.

	Click	UserID	Time_Spent	index	tags
2	1	1.0	17.250826	1002	India Hardik Pandya ODI Australia Virat Kohli ...
27	1	1.0	17.250826	3060	the Reserve Bank of India HDFC Bank overdues H...
14	1	1.0	17.250826	1822	Serena Williams the French Open Kristie Ahn Wi...
49	1	1.0	17.250826	5556	OnePlus Pete Lau Weibo OnePlus 8 OnePlus Rs 42...
24	1	1.0	17.250826	3507	The Finance Ministry Rs 9,879.61 Ministry Tami...
...	...	...	...	...	...
9	0	1.0	0.000000	1457	Lewis Hamilton Briton George Floyd Minneapolis...
54	1	1.0	0.000000	4193	Amazfit Neo Amazfit Neo Caravaan India Rs The ...
43	0	1.0	0.000000	4929	Vivo Vivo four-pixel Vivo V17 Pro's USP Macro ...
17	1	1.0	0.000000	924	Kerala Blasters' FC Goa the Indian Super Leagu...
51	0	1.0	0.000000	4111	India Bose Sony Jabra Sennheiser Amazon Great ...

80 rows × 5 columns

These are some recommendations for the old user based on the clickstream data. As we can see, there is a lot of variety for the selection of news items.

```
It took 0.003461122512817383 seconds to query forest.
It took 0.002943515777587890 seconds to query forest.
```

	Category	Sub_Category	Title	Synopsis	News	tags	cat
2976	Business	Banking-and-finance	RBI proposes to limit tenures of CEOs & whole-...	The RBI has said it is desirable to limit the ...	The Reserve Bank of India has proposed to rest...	The Reserve Bank of India RBI WTD the Departme...	1
3049	Business	Banking-and-finance	Reserve Bank puts on hold two key HDFC Bank ap...	The RBI move, which came over four months afte...	The Reserve Bank of India (RBI) has put on hol...	The Reserve Bank of India Sashidhar Jagdishan ...	1
1540	Sports	Motor-sport	Formula 1: Valtteri Bottas takes US pole, Lewi...	Valtteri Bottas will need a victory on Sunday ...	Mercedes' Valtteri Bottas snatched pole positi...	Mercedes Valtteri Bottas the U.S. Grand Prix L...	0
1822	Sports	Tennis	French Open 2020: Serena Williams reaches seco...	Serena Williams is a three-time French Open ch...	Serena Williams advanced to the second round o...	Serena Williams the French Open Kristie Ahn Wi...	0
2131	Sports	Wwe-wrestling	On Stone Cold Steve Austin Day, Coronavirus cr...	Stone Cold Steve Austin asked for a hell yeah ...	Stone Cold Steve Austin has been one of the mo...	Steve Austin WWE the Royal Rumble Austin 3:16"...	0
5755	Entertainment	Box-office-collection	Dabangg 3 box office collection prediction: Sa...	Considering the box office success of the last...	Chulbul Robinhood Pandey aka Salman Khan is re...	Chulbul Robinhood Pandey Salman Khan Dabangg D...	3
5596	Technology	Techook	The best smart speakers under Rs 20,000 in 2020	If you are looking for a smart speaker to cont...	Smart speakers are in the rage these days as m...	Bluetooth Apple HomePod Apple iOS Apple HomeKi...	2
4573	Technology	Laptops	Acer Swift 7 with thin bezels, compact design ...	CES 2019: Acer's Swift 7 features a high-resol...	At CES 2019, Acer has launched the Swift 7 not...	CES 2019 Acer Swift 7 92 per cent Intel Core 9...	2
4112	Technology	Gadgets	LG's new Ultrafine Ergo 4K monitor can be swiv...	The LG Ultrafine Ergo monitor can be tilted in...	LG has launched its new Ultrafine Display Erg...	LG Ultrafine Display Ergo India The LG 32UN880...	2
2695	Business	Aviation	Air India needs to survive till it is sold: CM...	Civil Aviation Minister Hardeep Singh Puri had...	As the central government is planning to invit...	Air India CMD Facebook Air India Ashwani Lohan...	1

So here, we are getting all the recommendations for one new user and we are also getting quite good recommendations for old users too.

VI. SENTIMENT ANALYSIS

We are doing the Sentiment analysis for our news items. The user will rate our news items on a scale of 0-5 depending on how positively or negatively impact the news have. A rating of 5 stands for the most positive news, and 0 stands for very negative news.

We are using this Sentiment analysis technique to ensure that we don't hurt users' feelings if they seek positivity in the news items recommended by us.

VII. SUMMARY

We have used two powerful algorithms to develop our recommendation system. Which are, NER and LSH

As we can see the results, we are getting better recommendations by using these two algorithms. These model is also quite fast and handy to use. Even for all the 100 users with 150000 data points, this algorithm gives recommendations to each user in less then one minute which is extremely fast..

So even for the small servers we can use this recommendation system to save time and this model will work efficiently.

only slow process in this algorithm is getting NER tags. On normal CPU, it will take too much time but we are using google collab GPUs so we are getting results within 10 min. Once NER tagging is done everything is running fast.

For collaborative filtering, we've sticked to our algorithm and decided not to use any other methods because mathematically speaking, every other method will take more computational time than this model. Even if we are using K-nearest neighbour method, then we will definitely lose more data than this method. For matrix factorization method, we will make a model by using just statistical data, but in this method, we are using news tags by using NER so that we will not lose any details about users' interests.

As this is a theoretical project report, many things don't get covered in this report, such as actually running a program and getting a bunch of news recommendations. **That is why we have added a Github link where you can see and run our program for the news recommender system**. Hope you will like our algorithm for JhakaasNewsVala. You can download this notebook and run it into google collab and check out our amazing algorithm.

This is the Github link for our project.
https://github.com/Kira1690/News-recommandation-system-by-using-LSH-NER

Thank you....