Team Atamnirbhar

Friday, 12.06.2020

Agenda

To develop a news article recommender system for Jhakaas News wala

What is a recommender system?

A Recommender System is a process that seeks to predict user preferences. Recommender systems are algorithms aimed at suggesting relevant items to users, items being movies to watch, text to read, products to buy or anything else depending on industries.The purpose is to suggest relevant items to users. To achieve this task, there exist two major categories of methods : collaborative filtering methods and content based methods.

1. Collaborative Filtering methods :

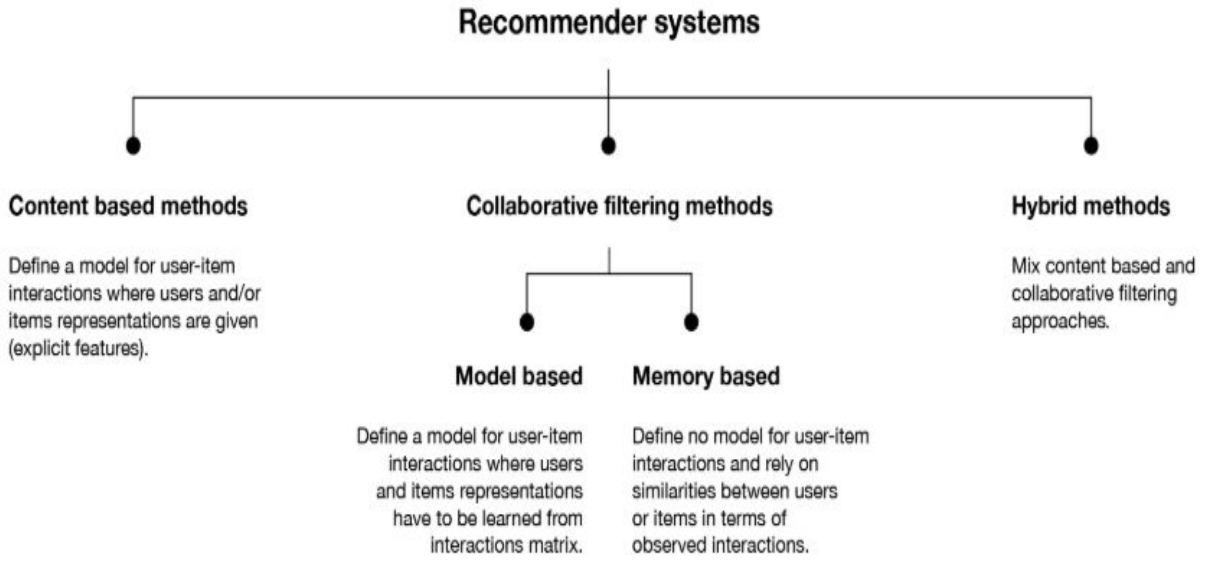
Collaborative methods for recommender systems are methods that are based solely on the past interactions recorded between user and items in order to produce new recommendations. The main idea that rules collaborative methods is that these past user-item interactions are sufficient to detect similar users and/or similar items and make predictions based on these estimated proximities. The class of collaborative filtering algorithms is divided into two sub categories:

1.1- Memory based : approaches directly work with values of recorded interactions, assuming no model and are essentially based on nearest neighbours search. For example finding the closest user of interest and suggesting the most popular items among these neighbours.

1.2- Model based: approaches assume an underlying “generative” model that explains the user-item interactions and try to discover it in order to make new predictions.

2. Content Based methods:

The main idea of content based methods is to build a model, based on the available "features", that explains the observed user-item interactions in other words content based approach uses additional information about user and/or items.



Summary of the different types of recommender systems algorithms.

Problem Statement : COLD START PROBLEM

Cold start concerns the issue that the system cannot draw any inferences for users or items about which it has not yet gathered sufficient information.

USER COLD START PROBLEM

When a recommendation system/engine meets a new user/visitor for the first time as there is no user history and the system doesn't know the personal preferences of the user it is called a user cold start problem

ITEM COLD START PROBLEM

Item cold start problem refers to when items added to the catalogue have either none or very little interactions. If no interactions are available then a pure collaborative algorithm cannot recommend the item.

DATA GENERATION PROCESS

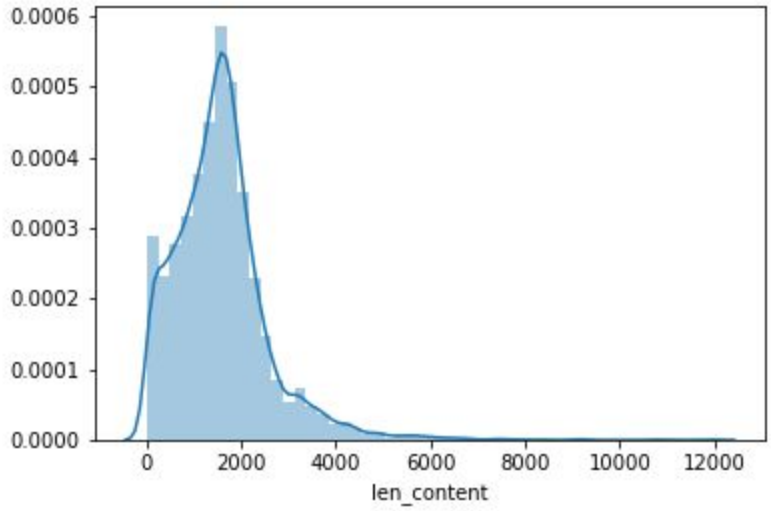
Generating Click Stream Data

1. This process was an integration of various Distributions.
2. For the process of Making **UserId** we made the list of 4000 users using simple Integers ranging from 1 to 4000.
3. For the purpose of making sessions for each user we used Geometric distribution Because Geometric distribution has one analogy that each user must have a minimum 1 session.
   Because Geometric Distribution says the number of failures before you get a success in a series of Bernoulli trials.
4. Minimum Session for each user is 1 and Maximum Session are 27.
5. For the purpose of Generating Clicks Per session for each we used Poisson Distribution because **Poisson Distribution** says:-
   It is a tool that helps to predict the probability of certain events from happening when you know how often the event has occurred. It gives us the **probability of a given number of events happening in a fixed interval of time**.
6. HOW IT WORKS :-
   And Say **poisson distribution** has generated 2 clicks in One session for Anyone user So we have assigned those 2 clicks randomly to 10 news articles assigned to each user in one session. We have randomly assigned these 2 clicks to articles because this data is generated by Distribution not by actual clickstream data.
7. For the purpose of generating Average time spent by the user in reading we used 3 **Gaussian mixture models** with different mean and standard deviation to cater the different reading speeds of each and also to take into account the variable of Length of articles. Which ranges from 250 words to 8100 words.

CONTENT FEATURES GENERATION

For the purpose of generating features from textual content we have tried various approaches.

1. TF/IDF
2. Word2vec, Glove.
3. LONG-FORMER (TRANSFORMER based approach)
   And Instead of Using BERT we used recently released LONG-FORMER Model Because BERT has limitation that we can only pass input to a BERT model Or even it's successors like Roberta,ALBERT,DistillBERT and Electra that we can only pass input's that have maximum length after tokenization equal to 512 tokens.



   But as you can see most of the articles we have are of length more than 500 clearly (more than 90% of articles)soif we have truncated our articles we would have to face information scarcity.So we used LONG-FORMER it is a transformer based model that deals with long docs easily here is the reference https://arxiv.org/abs/2004.05150.

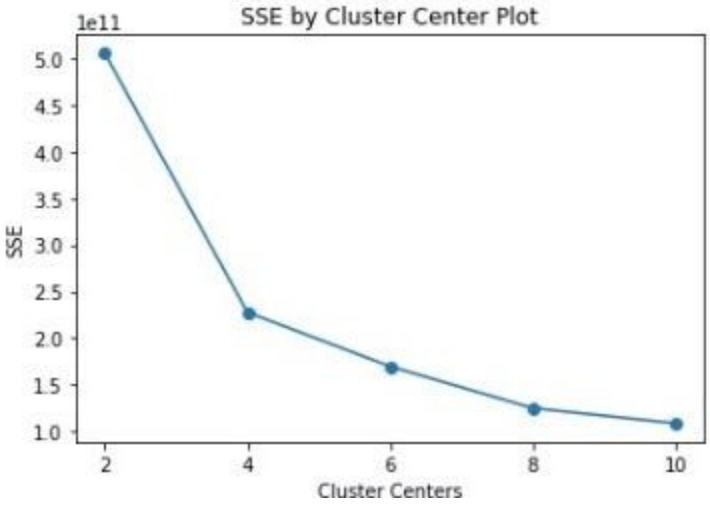
4. Long former has several features that make it better while handling Long Documents.
   4.1 It has no upper cap to set the maximum length which was the case in BERT.
   4.2 It has the concept of manually assigning Local and Global attention to tokens in documents.

5. We used Hugging Face Implementation of Longformer from https://huggingface.co/transformers/model_doc/longformer.html#longformermodel. We know these transformer models generally have 12 encoder layers and according to a research best embeddings are created when we sum the embeddings from the last 4 hidden layers. So we got word embeddings from the last 4 hidden layers and then took a sum of word embeddings to get Document embeddings.
   Because we know **BEST+BEST=BEST. (BEST word embeddings summed give BEST Document embeddings)**

   BLOG that describes the research
   https://jalammar.github.io/illustrated-bert/

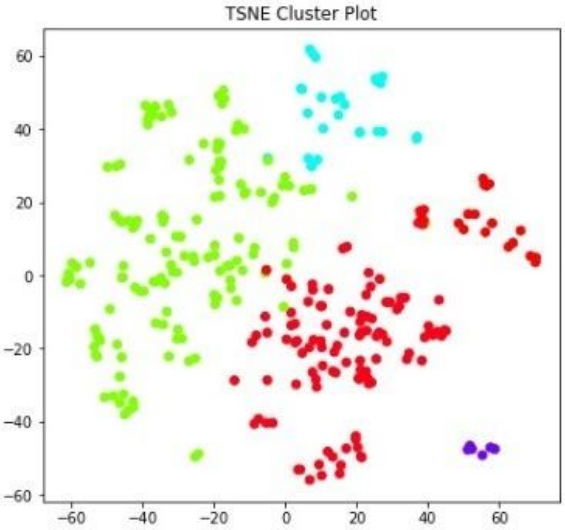
CONCLUSIONS FROM RESEARCH

ELBOW CURVE

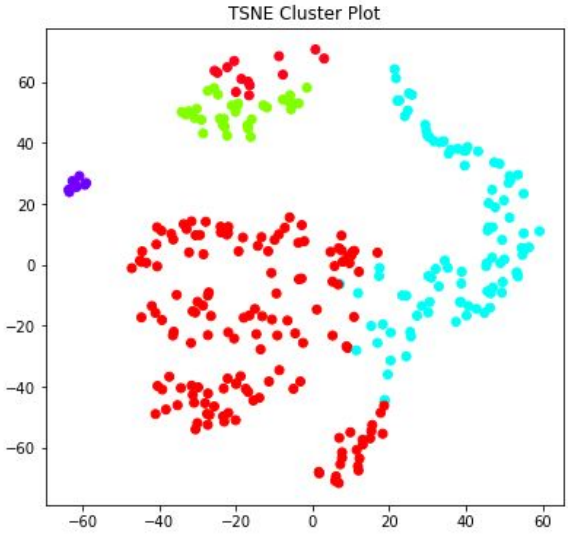


So we choose number of clusters equal to **5**

1. When we applied TF/IDF and did Kmeans clustering we got this result.



2. When we applied LONG-FORMER + Kmeans clustering we got this result.



3. When we were applying word2vec and Glove there was a lot of information because of vocabulary loss. So we dropped the plan.

4. When we eye-balled our top words predicted by each cluster from both techniques **TF-IDF+ Kmeans** and **LONG-FORMER+K Means** we found that **TF-IDF+Kmeans** got us better results. So we choose **TF-IDF+Kmeans** as our go to technique.

5. ONE VERY IMPORTANT CONCLUSION WE MADE FROM BOTH TECHNIQUES ABOVE WAS OUR **DATA IS HIGHLY BIASED** and we have got most of our clusters from **POLITICS** AND **TECHNOLOGY**.

OUR APPROACH FOR COLD START

For the purpose of Handling Cold Start problems we have devised several strategies.

1. To Handle the **COLD START** problem of **New Item**.
   In our case a new item will be a **news article** so we will first make features for that news article using **TF-IDF** then we will assign those features to its corresponding cluster. If it belongs to any of the clusters then it will be assigned to the cluster or otherwise a new cluster will be made.

2. To Handle the **COLD START** problem for **New User**.

   We have made 3 strategies.

   2.1 First we will recommend articles to new users **randomly** from each cluster (2 articles from each) as we have 5 clusters in total as shown by the elbow curve and we can show only 10 items to each user on each screen.

   This is what we got

```
Articles Recommended for you :)
Article  1 :  India hockey team for Rio Olympics 2016 to be announced on July 5  Report
Article  2 :  2016 Rio Olympics  Indian athletes continue to shine ahead of mega event
Article  3 :  Kolhapur municipal polls  Imams slam fatwa asking Muslim women not to contest election
Article  4 :  Indian prisoner lodged in Pakistan jail attacked twice  not taken to hospital
Article  5 :  Xiaomi Mi5 set for international debut at MWC 2016  confirms company
Article  6 :  Xiaomi Redmi Note 3 vs Micromax Canvas Evok vs Lenovo Vibe K5 Plus  Which smartphone should you buy
Article  7 :  Update OnePlus One with Android 5 1 1 Lollipop Custom ROM via PACMAN  How to Install
Article  8 :  Update Samsung Galaxy Note 4 with Android 5 1 1 Lollipop Custom ROM via crDroid  Steps to Install
Article  9 :   Great Grand Masti  faces censor board hassles over  vulgarity
Article  10 :  Nannaku Prematho audio launch live streaming  Watch Jr NTR s film s music release online
```

   2.2 Second we have made recommendations based on recency. That is we will Recommend articles to user based on Time.Example we will make **Top2 recent** Articles from each cluster.

```
Breaking News Recommendation:):
Article  1 :  Batman-inspired Samsung Galaxy Note7 edition images leaked  key design elements revealed
Article  2 :  Gionee S6s with 8MP front camera  LED flash launched in India  price  specifications
Article  3 :  MQM chief Hussain charged with treason for his  Pakistan is cancer for the world  remark
Article  4 :  BREAKING  Vinay Sharma  Nirbhaya rape convict  attempts suicide in Tihar  hospitalised
Article  5 :  Shoot of Salman Khan s  Tubelight  put on hold indefinitely
Article  6 :  Did  Great Grand Masti  actress Urvashi Rautela ditch Dubai Fashion Week organisers for more money
Article  7 :  Samsung Galaxy Note 3 gets Android Marshmallow via OrionOS  unofficial  custom ROM  How to install
Article  8 :  Samsung Galaxy S4 gets new Android Marshmallow update via CyanogenMod CM13  stable  custom ROM  How to insta
ll
Article  9 :  India to have 50 more operational airports by 2019
Article  10 :  Taxman sees multiple benefits in advancing budget presentation to January
```

2.3 Third we will recommend articles based on most clicks. That is user will be Recommended 2 articles from each of the 5 clusters based on most Clicks in each cluster.

```
Frequently viewed News Articles:
Article  1 :  Huawei Honor Bee First Impression  Amazing Budget Smartphone with Decent Specifications at Just  4 999
Article  2 :  Xiaomi MIUI 8 0 release date announced  key features  list of eligible devices
Article  3 :  Anti-India slogans will not be tolerated  Home Minister Rajnath Singh warns over JNU row  students  union pr
esident arrested
Article  4 :  Railways to invite bids for modernising 400 stations  says finance minister Arun Jaitley
Article  5 :   Tamasha  trailer out  Ranbir Kapoor-Deepika Padukone s chemistry captures hearts again  VIDEO
Article  6 :  AP T box office  Dilwale  Bajirao Mastani take a toll on Loafer  Nava Manmadhudu collection
Article  7 :  Sania Mirza Martina Hingis vs Mugurza Suarez Navarro live streaming and TV information  Watch WTA Tour Final
s doubles live
Article  8 :  2016 Rio Olympics  Vijender Singh unclear over representing India
Article  9 :  Update Moto G  aka Falcon  with Tesla Android 5 1 1 Lollipop Custom ROM via GZ Roms  How to Install
Article  10 :  Update Sony Xperia Z with Android 5 1 1 Lollipop Custom ROM via Tesla  How to Install
```

ONCE USER HAS SOME HISTORY

Algorithms used under Collaborative Filtering using Surprise Library

Surprise is a Python scikit building and analyzing recommender systems that deal with explicit rating data. It is a simple python library for building and testing recommender systems.

Basic Algorithms

Normal Predictor: predicts a random rating based on the distribution of the training set, which is assumed to be normal. This is one of the most basic algorithms that do not do much work.

BaselineOnly: predicts the baseline estimate for given user and item.

k-NN Algorithms

KNNBasic : is a basic collaborative filtering algorithm

KNNWithMeans: is a basic collaborative filtering algorithm, taking into account the mean rating of each user.

KNNBaseline: is a basic collaborative filtering algorithm taking into account a baseline rating.

Matrix Factorization-based algorithms

SVD : algorithm is equivalent to Probabilistic Matrix Factorization

SVDpp : algorithm is an extension of SVD that takes into account implicit ratings.

NMF: is a collaborative filtering algorithm based on non-negative Matrix Factorization.

Co-clustering: is a collaborative filtering algorithm based on co-clustering.

Benchmark for the above mentioned algorithms:

Algorithm	test_rmse	fit_time	test_time
BaselineOnly	1.440888	0.489739	0.835327
SVD	1.465503	12.730385	0.944266
SVDpp	1.489882	114.852252	4.759746
CoClustering	1.547328	5.025797	0.817716
KNNBaseline	1.667550	2.230085	4.185500
KNNBasic	1.671622	1.613935	3.353748
KNNWithMeans	1.677092	2.688790	4.223871
NMF	1.686163	14.057996	0.675459
NormalPredictor	1.870662	0.321230	0.785785

We use “RMSE” as our accuracy metric for the predictions,we can use FCP as well.

After tuning SVD parameters with GridSearchCV:

>Best parameters for svd:

{'n_epochs': 5, 'lr_all': 0.002, 'reg_all': 0.6}

> RMSE value reduced from 1.465503 to **1.419594**

After tuning BaselineOnly parameters with GridSearchCV:

>Best parameters for BaselineOnly:

{'bsl_options': {'method': 'sgd', 'n_epochs': 5, 'lr_all': 0.002, 'reg_all': 0.4}}:

> RMSE value reduced from 1.440888 to **1.4265558**

Predictions using SVD and BaselineOnly algorithms:-

```
[Prediction(uid=2, iid=0, r_ui=None, est=2.3550182781660647, details={'was_im
possible': False}),
 Prediction(uid=2, iid=1, r_ui=None, est=2.160331539925972, details={'was_imp
ossible': False}),
 Prediction(uid=2, iid=2, r_ui=None, est=2.4871278858878973, details={'was_im
possible': False}),
 Prediction(uid=2, iid=3, r_ui=None, est=2.1838198878971475, details={'was_im
possible': False}),
 Prediction(uid=2, iid=4, r_ui=None, est=2.331944823352466, details={'was_imp
ossible': False}),
 Prediction(uid=2, iid=5, r_ui=None, est=2.0439856638429985, details={'was_im
possible': False}),
 Prediction(uid=2, iid=6, r_ui=None, est=2.4254878534490816, details={'was_im
possible': False}),
 Prediction(uid=2, iid=7, r_ui=None, est=2.7345355590139118, details={'was_im
possible': False}),
 Prediction(uid=2, iid=8, r_ui=None, est=1.8284207530918508, details={'was_im
possible': False}),
 Prediction(uid=2, iid=9, r_ui=None, est=2.486232319034088, details={'was_imp
ossible': False})]
```

Predictions by SVD

```
[Prediction(uid=2, iid=0, r_ui=None, est=2.708472121464236, details={'was_imp
ossible': False}),
 Prediction(uid=2, iid=1, r_ui=None, est=2.615852436554692, details={'was_imp
ossible': False}),
 Prediction(uid=2, iid=2, r_ui=None, est=2.7177441385545906, details={'was_im
possible': False}),
 Prediction(uid=2, iid=3, r_ui=None, est=2.6507241809536355, details={'was_im
possible': False}),
 Prediction(uid=2, iid=4, r_ui=None, est=2.562330284651921, details={'was_imp
ossible': False}),
 Prediction(uid=2, iid=5, r_ui=None, est=2.4186013549243226, details={'was_im
possible': False}),
 Prediction(uid=2, iid=6, r_ui=None, est=2.7027730301506567, details={'was_im
possible': False}),
 Prediction(uid=2, iid=7, r_ui=None, est=2.7497372044376136, details={'was_im
possible': False}),
 Prediction(uid=2, iid=8, r_ui=None, est=2.282391241175862, details={'was_imp
ossible': False}),
 Prediction(uid=2, iid=9, r_ui=None, est=2.782657240943323, details={'was_imp
ossible': False})]
```

Predictions by BaselineOnly.

FUTURE SCOPE

1. We can make **NER models** that can get locations present in the articles and then we can make recommendations based on it.
2. We can use the **Multilingual Long former** model because we never know our article can have other languages. We could have **Multilingual BERT** models like **XLM Roberta** but the problem is the limit of 512 tokens input in these models so we Have to wait and look for **MULTILINGUAL LONG FORMER** models.
3. We can use the Percentage **Time** column as once a user starts to use our platform we can use this column to recommend articles according to their reading patterns.

References:

https://surprise.readthedocs.io/en/stable/prediction_algorithms_package.html#module-surprise.prediction_algorithms

http://nicolas-hug.com/blog/matrix_facto_4

https://www.youtube.com/watch?v=1JRrCEgiyHM

https://het.as.utexas.edu/HET/Software/Scipy/generated/scipy.stats.poisson.html

HOPE YOU LIKE IT

HAPPY READING :)