TV SHOW POPULARITY ANALYSIS USING DATA MINING

A Project Report of Capstone Project - 2

Submitted by

MANISH KUMAR

1613101370

in partial fulfilment for the award of the degree

of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

Under the Supervision of

Mr.RAVI SHARMA

Asst. Professor

APRIL / MAY- 2020

SCHOOL OF COMPUTING AND SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

Certified that this project report "TV SHOW POPULARITY ANALYSIS USING DATA MINING" is the bonafide work of "MANISH KUMAR" who carried out the project work under my supervision.

SIGNATURE OF HEAD.

Dr. MUNISH SHABARWAL,.
PhD (Management), PhD (CS).
Professor & Dean,
School of Computing Science & Engineering

SIGNATURE OF SUPERVISOR

Mr.RAVI SHARMA
Asst.professor
School of computing science & Engineering

ABSTRACT

The exact and opportune forecast of program prominence is of extraordinary incentive for content suppliers, promoters, and communicate TV administrators. This data can be beneficial for administrators in TV program buying choices and can assist promoters with defining sensible notice speculation plans. In addition, as far as specialized issues, an exact program notoriety forecast technique can improve the entire telecom framework, for example, the substance conveyance arrange procedure and store system. A few expectation models have been proposed dependent on record on-request (VOD) information from YouTube..Be that as it may, existing expectation strategies ordinarily require a huge amount of tests and long preparing time, and the forecast precision is poor for programs that experience a high pinnacle or sharp reduction in prominence. This paper presents our improved forecast approach dependent on pattern discovery. Initial, a unique time traveling separation based K-medoids calculation is applied to bunch projects' fame development into four patterns. At that point, four pattern specific expectation models are assembled independently utilizing irregular woods relapse. As indicated by the highlights separated from an electronic program direct and early review records, recently distributed projects are classified into the four patterns by an inclination boosting choice tree. At last, by consolidating guaging values from the pattern specific models and the classification likelihood, our proposed approach accomplishes better expectation results The test results on a monstrous arrangement of genuine VOD information from the Jiangsu Broadcasting Corporation show that, contrasted and the current expectation models, the forecast precision is expanded by over 20%, and the anticipating time frame is viably abbreviatevly

INTRODUCTION

Unscripted television is the new mantra of TV makers and station administrators. It is the intends to build TRP evaluations and the end is consistently to exceed different channels and the "comparable yet changed to a great extent shows produced by the opposition. The vast majority of the TV programs, which are being broadcast these days, are unscripted TV dramas spend significant time in moving, singing, and acting.

Everything presently is a competition. So wild is the opposition in this section each channel brags of at any rate a few unscripted TV dramas. Some of them are acquired legitimately from abroad, (for the most part and consistently from the USA the Godmother of unscripted tv) or some are modest duplicates of the shows abroad. The Indian unscripted TV dramas have likewise been reliably effective in offering a wide assortment. From Talent Hunt Shows, to move dramatizations, to acting flicks, television shows, talk appears, cookery shows.... the „reality chase list" is perpetual.

- We finish up to assemble such a framework, that will perceive individuals' nostalgic remarks on TV appears.

- The remarks from the watcher will be removed alongside the watcher subtleties, for example, sexual orientation, area, and so forth

- The remarks will be assembled from different sources and the passage will be kept up into the exceed expectations sheet.

- The exceed expectations record will contain people groups name, email id, age, sexual orientation, area and remark.
- In view of people groups remark and notions, the TV Show prevalence will be appraised appropriately.

- Administrator will Login into the framework and can perform errand, for example, Adding pages, looking after passages, seeing charts and printing the diagrams.

- Framework permits administrator to include pages by characterizing the name of the page and connection of that page.

- All the passages from individuals are kept up by the administrator in an exceed expectations sheet.

- The passages may contain name, email id, age, sexual orientation, area, likes-despises and their wistful remark.

- In view of the people groups remark, a chart will be produced by the framework, which will be ordered as age, sexual orientation, area and god or terrible remarks.

- Administrator can likewise print the framework created chart for keeping up a printed copy records.

- Guests can see TV show fame information in a graphical portrayal in pie outlines and bar diagrams.

- Guest can see the famous show rating just as the top show in a nation

MODULES:

The framework contains 2 significant modules with sub-modules as follows:

1.Administrator :-

a.Login: Admin need to login into the framework by contributing the login certifications.

b.Include Pages: Admin will include page subtleties, for example, page name and page interface.

C.Include Entry: Admin can include section for a page by choosing page name and giving different subtleties.

d.View Graph: Admin can create 5 diagram (Pie Chart and Bar Chart) in view of Age, Gender, Location, Comment's Sentiment dependent on people groups survey on every TV Show. Administrator will give the information in exceed expectations group and on the off chance that that information doesn't have any field, at that point administrator will enter it arbitrarily.

e..Print Graph: Can print all the 4 charts which is created by the framework.

2.GUES

T:

a.View Graph: Visitor can see 5 diagram (Pie Chart and Bar Chart) in light of Age, Gender, Location, Comment's Sentiment dependent on people groups survey on every TV Show. Administrator will give the information in exceed expectations group and on the off chance that that information doesn't have any field, at that point administrator will enter it arbitrarily. Guest can likewise see the prevalence of each show.

b.Print Graph: Can print all the 4 charts which is created by the framework.

Software Requirements:

Windows 7 or higher.

SQL 2008

Visual studio 2010

Hardware requirements:

Processor i3

Hard Disk 5 GB

Memory 1GB RAM

Web Connection

Advantages:-

- Nostalgic remark investigating and anticipating fortunate or unfortunate remarks.
- Simple expectation of TV Show drifting dependent on individuals rating.
- Graphical portrayal of TV Show notoriety.
- Arranging of charts by Age, Gender, Location and Good or Bad remarks dependent on people groups surveys or remarks.
- Simple bringing in of information and sending out it into diagram.
- Graphical information in printable arrangement.
- Guest will become acquainted with the show ubiquity.

Disadvantages:-

- May create off base outcomes if information not entered accurately.
- Requires dynamic Internet association.

METHODOLOGY

A.Problem Statement

The program ubiquity expectation issue can be defined as follows. Let c ∈ C be an individual program from a lot of projects C that are seen during a period T. We use t ∈T to depict the age of a program (i.e., the time since it was first distributed) and mark two significant minutes: the sign time ti, which is the time at which we play out the expectation, and the reference time tr, which is the snapshot of time for which we need to anticipate program prominence. Let Nc (ti) be the ubiquity of c from the time a program was distributed until ti and Nc (tr) be the worth that we need to anticipate, i.e., the notoriety sometime in the not too distant future Nc (tr). We define ^Nc (ti, tr) as the forecast result: the anticipated notoriety of program c at time tr utilizing the data accessible until ti. Subsequently, the better the forecast, the closer ^Nc (ti, tr) is to Nc (tr).

B.Method Overview

Our strategy follows 3 stages, as appeared in Fig. 1. The first step is to distinguish prevalence transformative patterns. We figure the DTW removes between chronicled record time arrangement and attempt to bunch the prominence developmental patterns into ideal patterns. Eleven static highlights separated from EPG are acquainted with fortify the aftereffects of bunching. A couple of preliminaries are performed to decide a fitting an incentive for the quantity of prominence patterns (k) for our situation study. For TV program notoriety, there exist various sorts of spread patterns. Diverse engendering patterns have distinctive significant level highlights. On the off chance that we could isolate them and train the model utilizing information from a specific sort of engendering pattern, we could acquire better outcomes for each kind. In this way, our first step is to distinguish the proliferation patterns and separate them into various sorts (groups). For TV proliferation patterns,

normal time arrangement bunching is performed, for which we can utilize DTW-based K-medoids. DTW is a standout amongst other separation estimating apparatuses; later, we will give an increasingly itemized prologue to DTW-based K-medoids. The subsequent advance is to assemble pattern specific expectation models utilizing RF relapse. We split the view records into 4 gatherings as indicated by the above patterns and feed them to the RF relapse model, together with static highlights. As per a few observational examinations, grouping program notoriety into multiple patterns won't improve the precision of the forecast model significantly. In this manner, we choose to group the prevalence developmental patterns of communicate TV programs into 4 expectation models. The third step is to utilize slope boosting choice tree (GBDT) to order the prominence time arrangement of recently distributed projects into the patterns and acquire the final predictionresultsbasedonthepredictionvaluesofthe4models and the classification likelihood.

C.Popularity Trend Detection

In this area, we portray the subtleties of our strategy for K-medoids [27] grouping of program prominence time arrangement with DTW [28] separation. In time-arrangement information investigation, the DTW separation is an exact proportion of the comparability between two worldly signals, which may have various paces. A non-straight mapping of one sign to another is acquired by limiting the separation between the two signs. This methodology is generally utilized in recognizing similitudes between worldly groupings of sound, picture, or video information, or any information that can be changed into a direct succession. Decades back, DTW was acquainted in the scholastic network with unravel for various talking speeds in programmed discourse acknowledgment issues. To find an ideal match between double cross arrangement successions, a "distorted" way limits the twisting expense to decide a proportion of their likeness that is autonomous of certain non-direct varieties in the time measurement. An ideal arrangement and separation between two groupings P = (p1, p2..., pn) and Q = (q1, q2...qm) can be resolved as follows:

$$DTW(P, Q) = \sqrt{dist(pn, qm)}, \quad (1)$$

$$dist(pi-1, qj)$$

$$dist(pi, qj) = (pi-qj)2 + \min \; dist(pi, qj-1) \quad (2)$$

$$dist(pi-1, qj-1).$$

The DTW separation is determined through unique programming to decide the base aggregate separation of every component in a n × m grid. What's more, the twisting way between two successions can be found by following back from the last cell. In this work, the DTW separation is utilized to gauge the closeness between each program's prevalence time arrangement information and group focuses to give

increasingly precise outcomes. The K-medoids calculation is like the notable K-implies calculation for performing bunching investigation. Be that as it may, these two strategies contrast by they way they update the middle area for a specific group. In the K-implies approach, the focal point of a group is virtual on the grounds that it speaks to the mean situation of the individuals that are at present inside the bunch. Notwithstanding, the K-medoids technique regards the inside as the middle of the bunch; hence, the inside corresponds with one of the individuals. Inferable from this distinction, the K-medoids calculation is increasingly vigorous to exceptions in the dataset.

The K-medoids calculation dependent on DTW calculation is depicted briefly as Algorithm 1. To start with, we subjectively pick k programs in D as the underlying medoids and dole out each residual program to the bunch with the closest medoids. At that point, we arbitrarily select a non-medoid program to process the new DTW separation of the patterns. In the event that the new DTW separation is not exactly the past one in the wake of trading, we trade to frame another arrangement of k medoids. The above advances are rehashed until there is no difference in programs in each pattern.

Algorithm 1 K-Medoids Based on the DTW Algorithm (KMDTW(D,C))

1. D: the data set containing program popularity time series
2. C: the number of trends
3. K: the set of trend centers
4. M: the set of popularity sequences in each trend
5. initialize C as trend centers of K
6. do
7. for i =1:size(D)
8. for k = 1:K
9. $Dist_{Di,Ck} = DTW(D_i,C_k)$
10. end for
11. if($Dist_{Di,Ck}$ is min)
12. assign D_i into M_k
13. end if
14. end for
15. while (the cluster membership changes)
16. return K, M

DATASETS

Datasets we are using in this project to train machine learning models is Obtained from IMDB site.This dataset contains motion picture surveys alongside their related parallel opinion extremity marks. This dataset contains 25k review of train sets and 25k review of test sets with their scores. In the whole collection, only upto 30 reviews are taken for same movie since reviews for the same movie will have associated ratings and it contains Entries for reviews with twofold names positive and negative. We incorporate as of now tokenized sack of words (BoW) includes that were utilized in our analysis.

ALOGRITHMS USED

we used are “Decision tree”, “Random Forest” , “K-nearest neighbors algorithm” , “Support vector clustering” , “Naïve bayes classifier” , “Stochastic Gradient Descent”. We will first check F1 score, precision score and accuracy for all algorithms by using it on test sets. Out of all these whichever algorithm gives us highest overall score we will use that algorithm to predict the statement whether it’s a positive or negative.

1. Decision tree – A decision tree is graphical structure of all possible outcomes to a decision based on various conditions. It starts with a root node than goes till the leaf node, leaf nodes contains the number of solutions. .

2.Random Forest – This calculation makes a woods with various choice trees. When all is said in done the more trees in the woods the more precise the forecast. It can perform relapse and grouping. .

3. K nearest neighbour—KNN algorithm identifies the k nearest neighbours of any element.it helps us to estimateits class. It can be used for both classification and regression.

4. Support vector clustering – SVM designs a hyperplane that characterizes all preparation vectors in various classes. Best decision of hyperplane that departs the most extreme edge from the two classes.

5. Naïve bayes – Naïve bayes algorithm is generally used when we have very less data in our training set. It is used for classification problems, mainly used for text classification involving high dimensional training data sets. For example—spam filtering.

.6. Stochastic Gradient Descent—it is used to build a predictive models. It is used to find optimal solution to a linear regression problem. It involves “loss function”, ”weak learner”, ”additive model”.F1 score is a formula to compute the score of precision and recall the higher the f1 score is the better prediction will be.Precision tells us what fraction of your outcome is relevant And recall tells us the fraction of total relevant results correctly predicted by your model.Stemming is a process in which different forms of word are converted to their root word for ex.

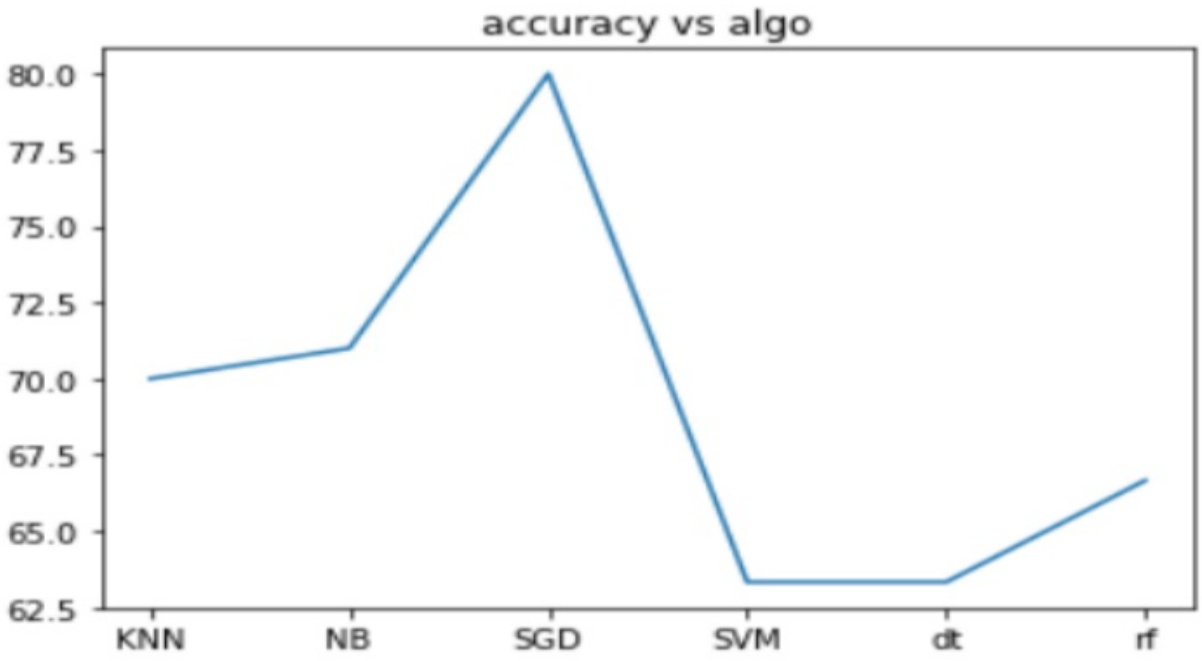
Going, goes go.Lemmantisation is process in which different forms of words are taken so they can be analysed as a single term by their dictionary form.Cosine similarity measures the similarity between two elements like in this project it is measuring the similarity between positive words and negative words.Bag of words is a representation that tells us how many times text comes in different entries. It

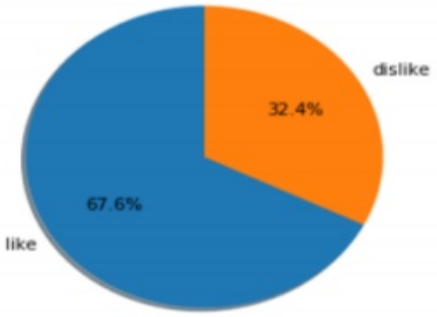
used in natural language processing

```
array([[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]])
```

	DT-C	**RF-C**	**KNN**	**SVM**	**NB**	**SGD**
	BGW	BGW	BGW	BGW	BGW	BGW
F1	61.22	47.91	55.81	38.77	55.81	76.19
Precision	61.11	82.75	83.92	31.66	84.92	81.98
Recall	61.48	54.54	59.09	50.00	56.32	74.64
Accuracy	63.33	66.66	70.00	63.33	71.00	80.0

OBSERVATIONS AND RESULT





CONCLUSION

In this paper we have introduced a prescient model to anticipate the prominence of television programs dependent on client remarks from web based life. We can get critical outcomes over the gave datasets. The model uses assessments of the watchers and can be actualized for any sort of program. Results are profoundly precise dependent on the qualities we've acquired utilizing information mining and AI. In this paper we have introduced a prescient model to anticipate the prominence of television programs dependent on client remarks from web based life. We can get critical outcomes over the gave datasets. The model uses assessments of the watchers and can be actualized for any sort of program. Results are profoundly precise dependent on the qualities we've acquired utilizing information mining and AI.

REFERENCES

1. Yu-Hsuan Cheng, Chen-Ming Wu, Tsun Ku, Gwo-Dong Chen. A Predicting Model of TV Audience Rating Basesd on Facebook, 2013.

2. Yusuke Fukushima, Toshihiko Yamasaki, Kiyoharu Aizwa. Audience Ratings Prediction of TV Dramas Based on the Cast and their Popularity, 2016.

3. Nicolai H. Egebjerg, Niklas Hedegaard, Gerda Kuum, Raghava Rao M,Ravi Vatrapu.Big Social Data Analytics in Football: Predicting Spectators and TV Ratings from Facebook Data, 2017.

4. Chengang Zhu, Guang Cheng, Kun Wang. Big Data Analytics for Program Popularity Prediction in Broadcast TV Industries ,2017.

5. Javaria Ahmed, Prakash Duraisamy, Amr Yousef, Bill Buckles. Movie