ZKG INTERNATIONAL ISSN: 2366-1313

MOVIE LENS DATA ANALYSIS

Mrs. k. Brahmeswari madam, K. Maha Brahamachari , J. Keerthi Angel, V. Shyam, K. Jayadev

Associate Professor, Priyadarshini Institute of Technology&Science ,AP,india

Under Graduate, Priyadarshini Institute of Technology & Science, AP, India.

ABSTRACT:

The Movie Lens datasets are widely used in education, research, and industry. They are downloaded hundreds of thousands of times each year, reflecting their use in popular press programming books, traditional and online courses, and software. These datasets are a product of member activity in the Movie Lens movie recommendation system, an active research platform that has hosted many experiments since its launch in 1997. This article documents the history of Movie Lens and the Movie Lens datasets. We include a discussion of lessons learned from running a long-standing, live research platform from the perspective of a research organization. We document best practices and limitations of using the Movie Lens datasets in new research.

Keywords: Bigdata, machine learning, movie lens.

1.INTRODUCTION:

Human has a long history with basic data visualization, and data visualization is still a hot topic today. The history of visualization has to some extent been shaped by technology available and by the pressing needs of the time probably, including those of: Prehistoric paintings on clays, stones, maps on walls, images, table of numbers (with rows and columns), all these are a kind of data visualization - although we may not call them under this name at the time. Visualization is the graphical presentation of information in an attempt to give the viewer a qualitative knowledge of the actual content of the information. It is also the process of

trying to transform objects, concepts and numbers into a visible and familiar form for human eyes. We can indeed refer to information, procedures, relationships or concepts also when we say "information".

ISSN: 2366-1313

Data visualization is simply the representation of information in the form of charts, graphs, images, etc. It helps in understanding the patterns and trends out of the datasets. Data visualization is all about understanding ratios and numerical relationships. Not to understand individual numbers, but to understand the patterns, trends and relationships in groups of numbers

Seeing and understanding images is one of the basic instincts of human beings and understanding the numerical data requires training skills from the schools, and yet many people are still not good at numerical data. It is much easier to identify trends, patterns and relationships from a well - drawn picture. Because graphical presentation of relevant information takes full advantage of the human eye's huge and often underutilized ability to measure picture and illustration information, the visualization of information shifts the load from numerical to visual. Gathering information from pictures probably saves somewhat more time than looking through original text and actual numbers - that is why many decision - makers probably prefer
to have information provided to them in graphic form rather than in written or text form.

The difficulty of evaluating recommender systems is often pointed out. Recent work showed that many papers overestimate the performance of new algorithms [2, 4]. Pre processing also is a problem for evaluation, but the diversity of pre processing often reflects the diversity of recommender systems applications: most datasets are private and have very different properties. As an example, for session-based recommender systems, researchers often use pre processing to transform ratings datasets such as Movie Lens [3] to this specific case. We argue that the lack of guidelines at this step makes evaluation and comparison of algorithms harder. In this paper, we explicit the diversity of pre processing protocols and use it to extract information about datasets. We analyze how metrics vary across setups. Our key contributions are the following:

• we define a robustness metric evaluating how much a performance metric varies against pre processing protocols of a dataset,

• we define a signature for pre processed datasets,

• we propose a principled way of selecting a pre processing protocol for publishing results.

2.RELATED WORK:

ISSN: 2366-1313

RELATED WORK:

Recommender system is used to make enhanced and supportive user understanding as it is an intelligent method. By streaming data processing, calculation performed not as a batch but in Real Time when data appear. For various great organizations, real-time processing and analytics are becoming important part of big data approach. In the world of big data, predictive systems are flattering trendier because this automated tool unites clients toward the products that are best well matched to them by linking product content and uttering feedback. Clients may find it hard-hitting to choose the best package that encounters their specific interest and necessity. Collaborative Filtering is to be used in our paper for the construction of recommendation system in real world as it is one of the best efficacious technique and automatic predictions about costumer's concern can be measured by using algorithms in Collaborative Filtering. It can predict the preferences of additional indefinite customers, and offer personalize

recommendations by analysing the preferences of current parts of customers and has been broadly recycled in profitable websites comprising Amazon, Netflix, eBay, Taobao, Hulu etc [1,2] [Khorasani, E. S., et al 2016] [Yang, Z. , et al 2016]. Recently many efficacious fallouts attained using collaborative filtering algorithm with Hadoop. For unstructured data analysis, Hadoop is best as it doesn't request for definite data type and inexpensive product hardware is used as data knobs for storing data and calculations of data. But the unstructured data cannot be managed by Hadoop in real-time. So that Hadoop architecture introduced Spark as its extension that is having characteristics of both Storm and Hadoop. Many computer languages such as Java, Python, R and Scala are supported by Spark and Python is one of the firmest evolving language having infinite lending library maintenance that is why we rely on this language. In my paper, for analysing movie review dataset, the PySpark Data frame Transformation and

Actions is used. For the accomplishment of proficiency, RDD is used by Spark. It focuses on commemoration of operations that caused in modern RDD and attains fault tolerance to acquire RDD.ALS (Alternating Least Squares) model is collaborative filtering model that will be used in our work. This effort offers a new instruction to the user based on Apache Spark and Spark RDD strategies combined with the mutual size reduction and clustering of the Spark MLIB algorithms. MLlib is machine learning library that is accessible by Spark and contain collective algorithms and services, these are Classification, Clustering,

ISSN: 2366-1313

Collaborative filtering, Regression, size reduction and also the essential optimization primitives.

Similar to my research, there are also many previous researches that relate to movie rating appeal recommendation system by using Apache Spark. Here previous research work will be discussed. [Ruining He, & Julian McAuley. 2016][3] They proposed that to construct a recommender system, the understanding about client's performances and emotions are mandatory. To construct such type of system, the dataset is rare. With the help of previous feedback, a new modified recommender system is established. Over a period of time the journalist correspondingly search the changing fashion tendency.

[Li Zhuang, Feng Jing, & Xiao-Yan Zhu. 2006][4] They proposed that, for the companies, reviewers and readers of review, the reviews are very advantageous. The companies attract the clients by the reviews and reviewers are given the incentives and the readers will be able to read the reviews and distinguish that the product that they want to buy is comparatively better than other product or not. [Lina L Dhande and Girish K Patnaik. 2014][5] In their paper they discussed to know about movie is worthy or ruthless, simple classifier along with neural network is applied and for the grouping, the unigram feature is recycled. [Callen Rain. 2013][6] The author said, when 15 products of Amazon gets 50000 reviews then It can be classified that if the clients loved the products or not. In this research simple classifier is taken for grouping. Kindle is a product that is the focus of study. [Neelu Rani, Nishant Singh, SujayPawar, et al. 2017] [7] They classified positive or negative opinions of people by SVM based

classifier on movie review data. [Xiaomeng Su.] [8] Showed a diagrammatic explanation in his paper. [J. Sangeetha and Dr. V. Sinthu Janita Prakash. 2019] [13] in proposed research method MoCFCR Modified Collaborative Filtering and Clustering with Regression is presented for the accuracy of movie review classification. K means algorithm is used for clustering process. In this way the movie review is done in an easy way. The overall methods performed in Hadoop so proving that the research is giving better outcome than previous research paper's outcome.

ISSN: 2366-1313

[G. Bathla, 2017] [14] the researchers now-a-days are arguing on the societal systems and also about the big data. By the usage of societal system graphs, the journalist deliberated content based and collaborative filtering (CF) in encouragement or guidance of societal reliance.New methods of recommendations are projected using item rating matrix, by use of Pearson correlation coefficient the comparison between sets are to be intended.

Since analyzing the characteristic of the popularity of online content is an essential step to improve the effectiveness of strategies, such as caching, advertising, recommendation, the field has been widely researched for a couple of decades. For example, the early study of Gill et al. [17] who examined the usage patterns, file properties, popularity and referencing characteristics, and transfer behaviors of YouTube in comparison to other traditional webs and media streaming. Moreover, the authors also discovered that the popularity distribution followed a Zipf-like distribution [31]. In the later research, Zink et al. [43] analyzed how content popularity distributes in YouTube as well as conducted several measurements on YouTube traffic in a large university campus network. In another considerable study, Cha et al. [8] analyzed the popularity of videos on You tube, especially some popular Korean videos. They described the group popularity of those videos as a power law [14] with an exponential cut-off.

Notably, Cheng et al. [11] provided an insight into not only the popularity distribution but also the access pattern, popularity trend and social networking of videos on You tube, by performing a long-term observation. In this work, the author also explored that the Video on Demand (VoD) accesses follow almost the same as the gamma distribution [38], more than other distributions, such as Zipf [31] and Weibull [37]. Unlike the above studies, the recent research of Li et al. [25] examined the data collected from the online video service provider Youku [1]. The research exploited the whole log server of Youku to extract some valuable knowledge r elated to the long-term popularity, the video lifetime, the popularity evolution pattern, and the early stage popularity.

In summary, all above studies have provided essential insight into the popularity of online content, the network traffic, they have also shown the evidence that the video service providers can exceedingly benefit from such analysis by customizing the caching algorithms,

ISSN: 2366-1313

the content distribution. However, most of the studies focus on the long-term popularity of a specific group of videos in a particular region. In this work, we perform some analysis on the whole Movie Lens dataset and examine the access patterns of some popular You tube videos in 20 countries. As we mainly focus on analyzing the popularity evolution in the short periods, our data is collected more frequently, which is also used to evaluate the effectiveness of our predictive model.

videos on You tube, especially some popular Korean videos. They described the group popularity of those videos as a power law [14] with an exponential cut-off. Notably, Cheng et al. [11] provided an insight into not only the popularity distribution but also the access pattern, popularity trend and social networking of videos on You tube, by performing a long-term observation. In this work, the author also explored that the Video on Demand (VoD) accesses follow almost the same as the gamma distribution [38], more than other distributions, such as Zipf [31] and Weibull [37]. Unlike the above studies, the recent research of Li et al. [25] examined the data collected from the online video service provider Youku [1]. The research exploited the whole log server of Youku to extract some valuable knowledge related to the long-term popularity, the video lifetime, the popularity evolution pattern, and the early stage popularity.

In summary, all above studies have provided essential insight into the popularity of online content, the network traffic, they have also shown the evidence that the video service providers can exceedingly benefit from such analysis by customizing the caching algorithms,
the content distribution. However, most of the studies focus on the long-term popularity of a specific group of videos in a particular region. In this work, we perform some analysis on the whole Movie Lens dataset and examine the access patterns of some popular You tube videos in 20 countries. As we mainly focus on analyzing the popularity evolution in the short periods, our data is collected more frequently, which is also used to evaluate the effectiveness of our predictive model.

ISSN: 2366-1313

In the beginning, the field of predicting online contents popularity was first proposed by the initiative of Szabo and Huberman[39]. The authors demonstrated the strong linear correlation between the long-term popularity and the early popularity of online contents on the logarithmic scale. Based on the property, they proposed a simple log-linear model to predict the overall popularity of given online content in the future. The experiments were conducted on various datasets, including You tube videos[2] and Digg stories[39]. Inspired by this idea, Pinto et al. [35] proposed two multivariate regression models which were able to predict the popularity of online contents using the daily samples of their popularity measured up to the given reference date. Their empirical results showed that those models achieved reasonable accuracy on You tube dataset. Notably, in the recent study, Li et al.[25] proposed a novel model that captures the popularity dynamics based on early popularity evolution pattern

and future popularity burst prediction. The authors used some basic early popularity measurements as well as considered the characteristic of individual video and its popularity evolution pattern as the input of their model to increase the accuracy. In addition to the traditional regression-based methods, some other techniques such as reservoir computing[42], time series analysis[18] are also applied to improve the performances. Their approach was evaluated on an exhaustive real-world data collected from Youku [1] and achieved significant decreases in relative prediction errors. Despite achieving initial results, the aforementioned studies mainly focused on predicting the long-term popularity of the given content. To address the problem in both long-term and short-term, a study of Hush chyn et al. [23] proposed a simple artificial neural network (ANN) to predict the popularity of scientific datasets at the Large Hadron Collider at CERN. Since the model is not robust enough to accurately predict the popularity of an item, it is mainly used for classification. Moreover, in our previous work [32], we also introduced a superior model called Attention-based Non-recursive Neural Network (ANRNN) to address predicting the popularity of online content in the near future. Although ANRNN model can outperform some state-of-the-art predictive models, the computation cost could only be improved about 1.8 to 4 times. Thus, in this study, we intend to provide another predictive model which significantly reduce the amount of computation, we then evaluate its performance in both accuracy and inference time aspects in comparison to other existing methods.

3.METHODOLOGY:

Dataset

ISSN: 2366-1313

The problem we'll solve is a binary classification task with the goal of predicting an individual's health. The features are socioeconomic and lifestyle characteristics of individuals and the label is 0 for poor health and 1 for good health

Preprocessing:

Generally, 80% of a data science project is spent cleaning, exploring, and making features out of the data. However, for this article, we'll stick to the modeling. an imbalanced classification problem, so accuracy is not an appropriate metric.

Feature Importance:

random forest algorithm is that it is very easy to measure the relative importance of each feature on the prediction. Sklearn provides a great tool for this that measures a feature's importance by looking at how much the tree nodes that use that feature reduce impurity across all trees in the forest. It computes this score automatically for each feature after training and scales the results so the sum of all importance is equal to one.

Feature Selection of Random Forest Algorithm

Feature selection is the process of using a series of rules to calculate the relative relationship of importance of the characteristics and to rank the characteristics of the data. Feature selection techniques are often used to transform the data in classification analysis, so as to improve the accuracy of classification. In general, feature selection techniques in machine learning are mainly divided into three categories: filter method (Filter), encapsulation method (Wrapper) and integration method.

The following is the introduction to the filtering method. Filtration method is through the statistical method, to give the characteristics with a weight, to carry out feature ranking according to the characteristics of the weight, and then apply some rules to set a threshold, the feature whose weight is greater than the threshold value is retained, otherwise deleted. The feature selection process of the filtering method is operated according to the feature of the data set, which is independent of the specific classification algorithm. There are many

ISSN: 2366-1313

common filtering methods, such as Fisher ratio, information gain, and Relief, T-test and variance analysis. In the following, variance analysis will be briefly introduced.

When using variance analysis to test the characteristics of the screening method, calculate the test statistic.

/Using Out-Of-Bag Estimates to Monitor Error, Strength, and Correlation

In my experiments with random forests, bagging is used in tandem with random feature selection. Each new training set is drawn, with replacement, from the original training set. Then a tree is grown on the new training set using random feature selection. The trees grown are not pruned. There are two reasons for using bagging. The first is that the use of bagging seems to enhance accuracy when random features are used. The second is that bagging can be used to give ongoing estimates of the generalization error (PE*) of the combined ensemble of trees, as well as estimates for the strength and correlation.

Random Forests Using Linear Combinations of Inputs

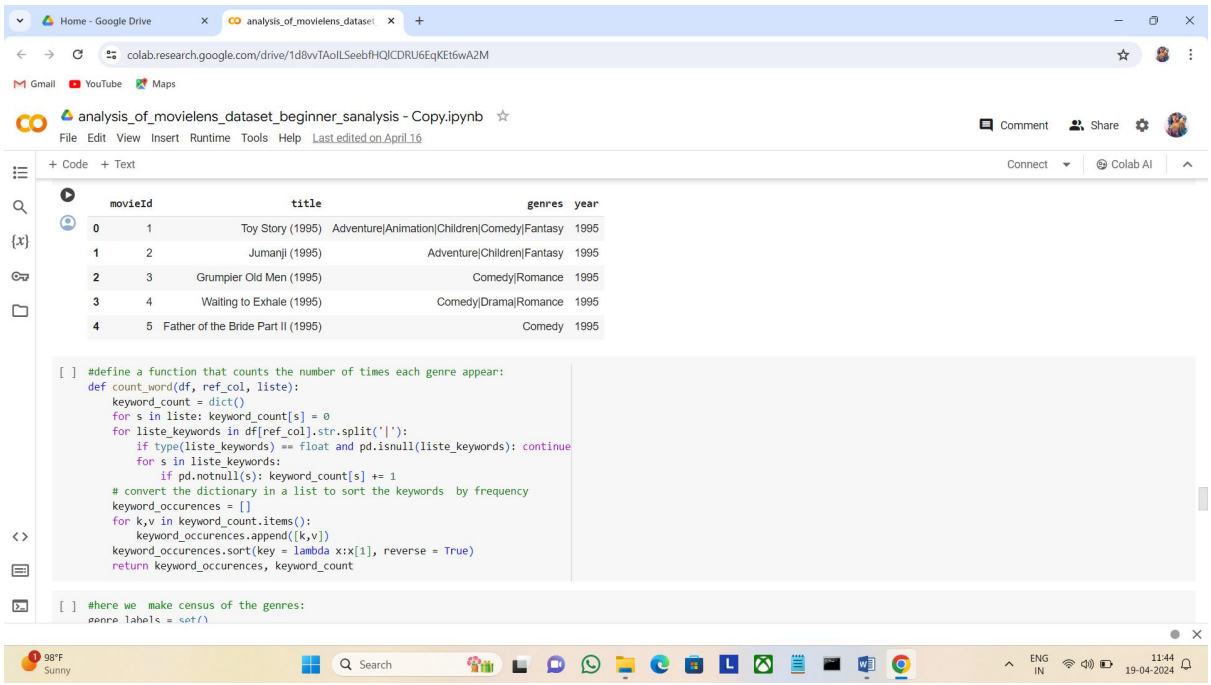
If there are only a few inputs, say M, taking F an appreciable fraction of M might lead an increase in strength but higher correlation. Another approach consists of defining more features by taking random linear combinations of a number of the input variables. That is, a feature is generated by specifying L, the number of variables to be combined. At a given node, L variables are randomly selected and added together with coefficients that are uniform random numbers on [-1,1]. F linear combinations are generated, and then a search is made over these for the best split. This procedure is called Forest-RC .

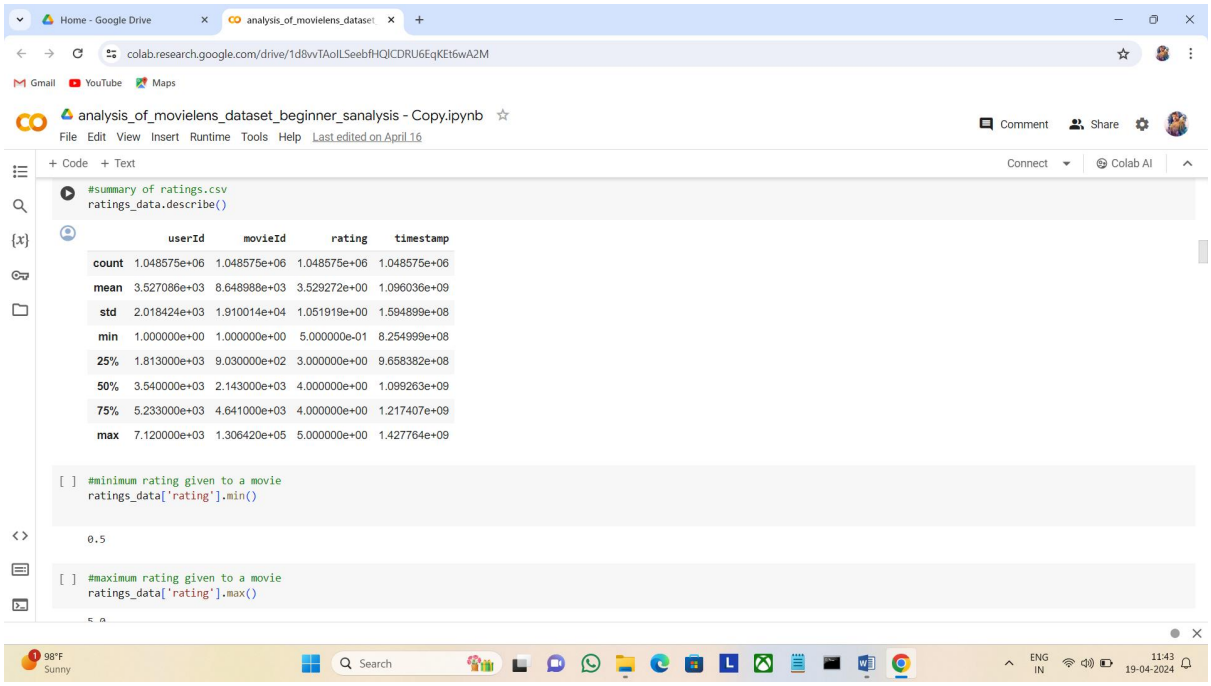
Results

The final testing ROC AUC for the random forest was 0.87 compared to 0.67 for the single decision tree with an unlimited max depth. If we look at the training scores, both models achieved 1.0 ROC AUC, which again is as expected because we gave these models the training answers and did not limit the maximum depth of each tree..

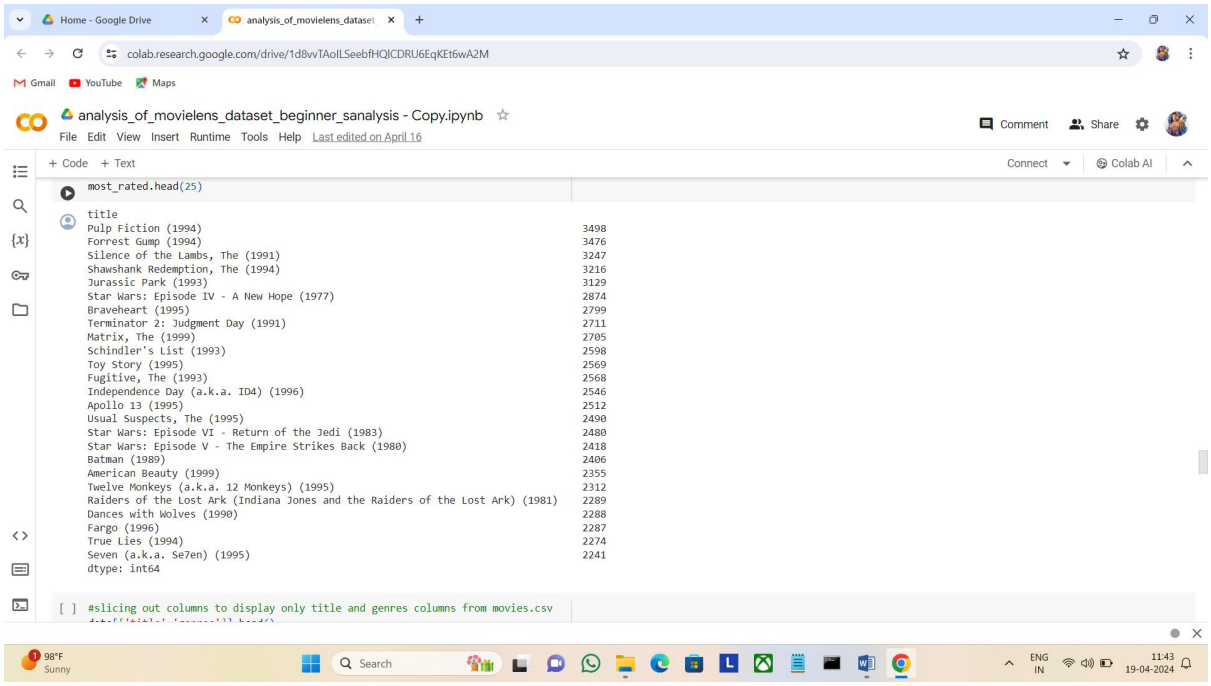
ISSN: 2366-1313

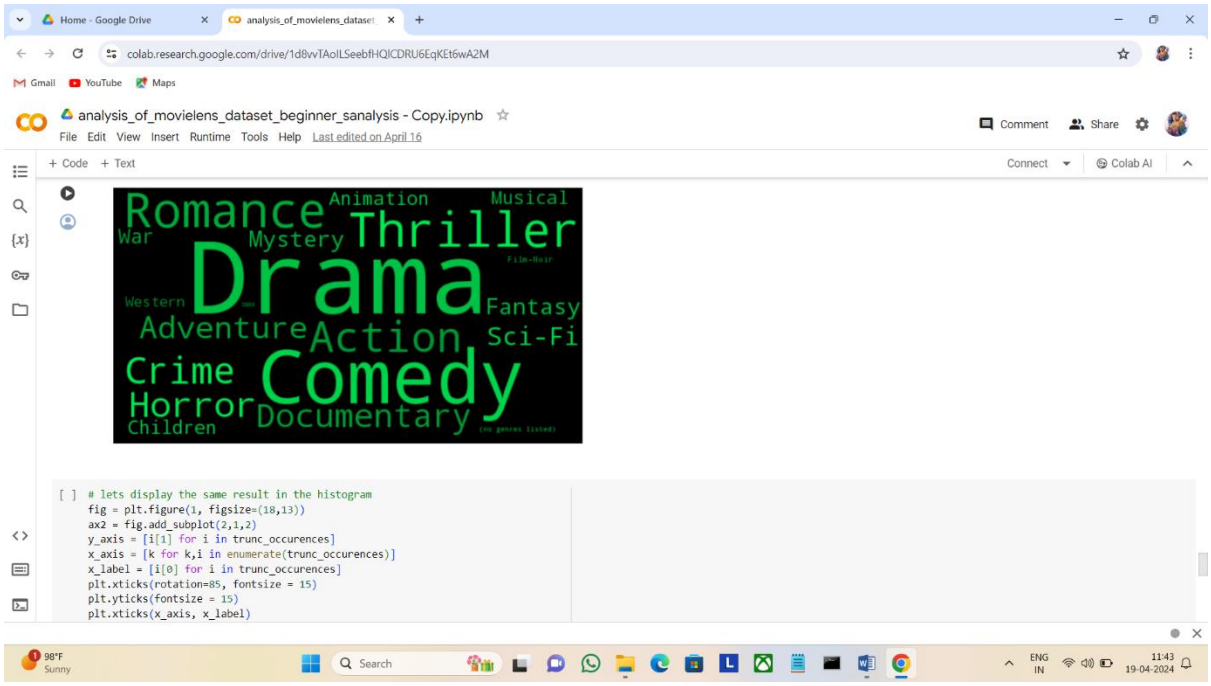
4.RESULT:



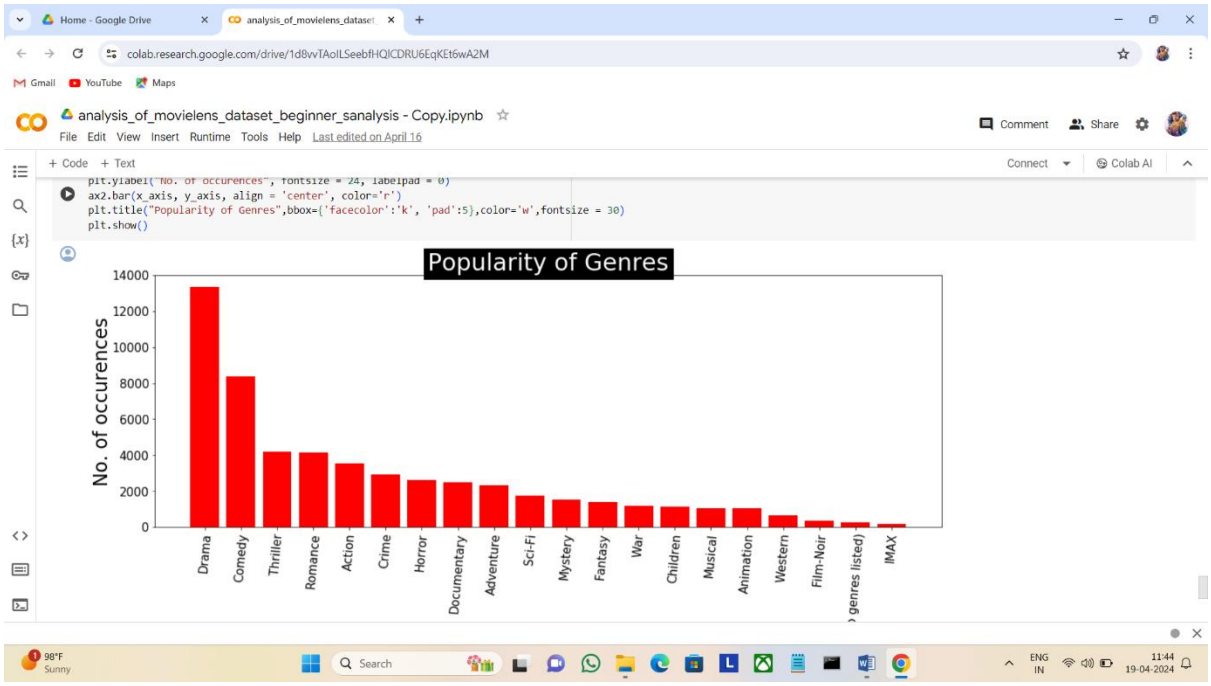


ISSN: 2366-1313





ISSN: 2366-1313



5.CONCLUSION:

The importance of dataset preprocessing cannot be underestimated. If a dataset has low robustness, we show that the preprocessing can change the conclusions of the experiments. We propose a transparent method to select a protocol fitting to the target application. The longer version of this paper will include more extensive experiments to analyse the value of the proposed signature, and a study of how signature similarity generalizes to new algorithms performance similarity. It would be ideal to have a single benchmark dataset for all recommender systems. We argue that it is unlikely that the same benchmark may cover the very different use cases of the industry and instead propose to have a transparent definition of the preprocessing protocol.

7.2 FUTURE ENHANCEMENTS

Since the rate at which new content is uploaded to the Internet has reached unprecedented marks, understanding how user attention distributes on online contents is of importance for network management, infrastructure design, advertising planning and many other services. In this paper, we provide some essential insight into the characteristics of the popularity of videos on one of the biggest video service providers, You tube, and movies within Movie Lens server. Specifically, our analysis reveals the changes of content popularity within short periods as well as the difference among user access patterns in many countries.

ISSN: 2366-1313

6. REFERENCE:

[1] 2016. Youku Tudou Partners With Xiaomi to Accelerate Multi-Screen Ecosystem Development. http://ir.youku.com/phoenix.zhtml?c=241246&p=irolnewsArticle_print&ID=1988630. Last Accessed: March 10, 2016.

[2] 2019. You tube application programming interface. https://developers.google. com/you tube/. Last Accessed: March 18, 2019.

[3] 2019. youtube.com Competitive Analysis, Marketing Mix and Traffic. https://www.alexa.com/siteinfo/youtube.com. Last Accessed: March 18, 2019.

[4] Hervé Abdi. 2007. The Kendall rank correlation coefficient. Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA (2007), 508–510.

[5] Dimitros Asteriou and Stephen G Hall. 2011. ARIMA models and the Box–Jenkins methodology. Applied Econometrics 2, 2 (2011), 265–286.

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014).

[7] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In Noise reduction in speech processing. Springer, 1–4.

[8] Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. 2009. Analyzing the video popularity characteristics of large-scale user generated content systems. IEEE/ACM Transactions on networking 17, 5 (2009), 1357–1370.

[9] Gloria Chatzopoulou, Cheng Sheng, and Michalis Faloutsos. 2010. A first step towards understanding popularity in YouTube. In 2010 INFOCOM IEEE Conference on Computer Communications Workshops. IEEE, 1–6.

[10] Xu Cheng, Cameron Dale, and Jiangchuan Liu. 2008. Statistics and social network of youtube videos. In 2008 16th Interntional Workshop on Quality of Service. IEEE, 229–238.

[11] Xu Cheng, Jiangchuan Liu, and Cameron Dale. 2013. Understanding the characteristics of internet short video sharing: A youtube-based measurement study. IEEE transactions on

multimedia 15, 5 (2013), 1184–1194.

[12] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014).

[13] Cisco Visual Networking Index Cisco. 2016. The zettabyte eraâĂŤtrends and analysis, 2015–2020. white paper.

[14] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. Power-law distributions in empirical data. SIAM review 51, 4 (2009), 661–703.

[15] Eugen Diaconescu. 2008. The use of NARX neural networks to predict chaotic time series. Wseas Transactions on computer research 3, 3 (2008), 182–191.