Android PlayStore Data Analysis

Prof Sumit Kalra

Divya Meghana Mamidipalli

B20CS032

Contents:

1.Introduction

2.Data Scraping

- 2.1 Scraped Data For Educational Apps
  - 2.1.1 Exploratory Data Analysis
- 2.2 Scraping the Reviews Data For Each App
  - 2.2.1Analyzing the Reviews Data

3.Topic Modeling the Reviews

- 3.1 Preprocessing the data
- 3.2 Creating a topic modeling model
  - 3.2.1 Latent Dirichlet Allocation
- 3.3 Printing and Analyzing the topics

4.Results Obtained

5.Conclusions

1. INTRODUCTION

Android Play Store Data Analysis, as the name says, is analyzing the data extracted from Android Google play store. We know there are more than 3.04 million apps found on Google play store. With this project I am going to analyze various apps found on play store with the help of different python libraries.

2. Data Scraping

Data scraping, also known as web scraping, is the process of importing information from a website into a spreadsheet or local file saved on your computer. It is a practice that can automatically extract data from websites, databases, enterprise applications, or legacy systems. It's one of the most efficient ways to get data from the web, and in some cases to channel that data to another website.

There are a lot of techniques for web scraping, out of which I used ParseHub for scraping the data. ParseHub is a free web scraping tool.

2.1 Scraped Data for Educational Apps

The data of educational apps is scraped using ParseHub which contains our required information about every app in the form of a csv/excel file. I collected the data of 30 educational apps.

Names of those 30 educational apps:{ ABC Kids, ABCmouse, BYJU'S, Brainly, Coursera, Diksha,DoubtNut, Duolingo, Embibe, Eduaraa, Epic, ExtraMarks, GreatLearning, KhanAcademy, KhanKids, Kutuki, MathKids, Memrise, Oda, Prodigy, TeachMint, ThirdFlix, TinyTap, Toppr, Udemy, Unacademy, Vedantu, ePathshala, edX, uLesson}

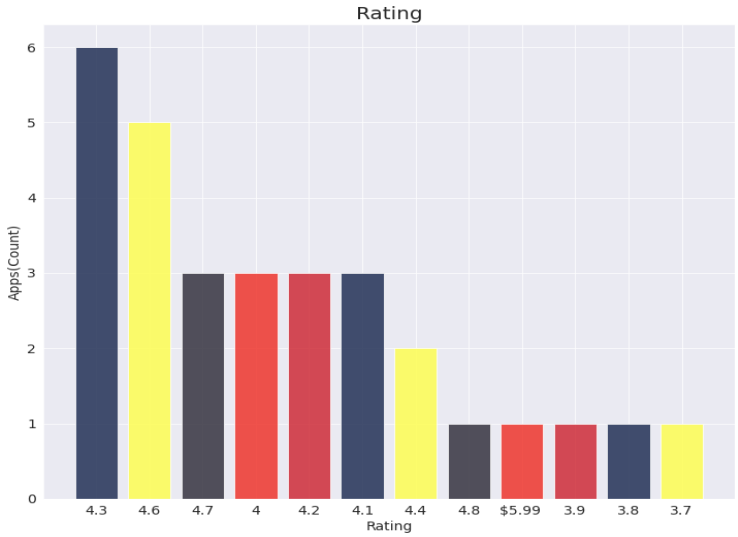
DataSet: The dataset contains the details of 30 educational apps like app name, number of downloads for that app, rating, number of reviews. The data is found in

Data_Educational_Apps

2.1.1 Exploratory Data Analysis:

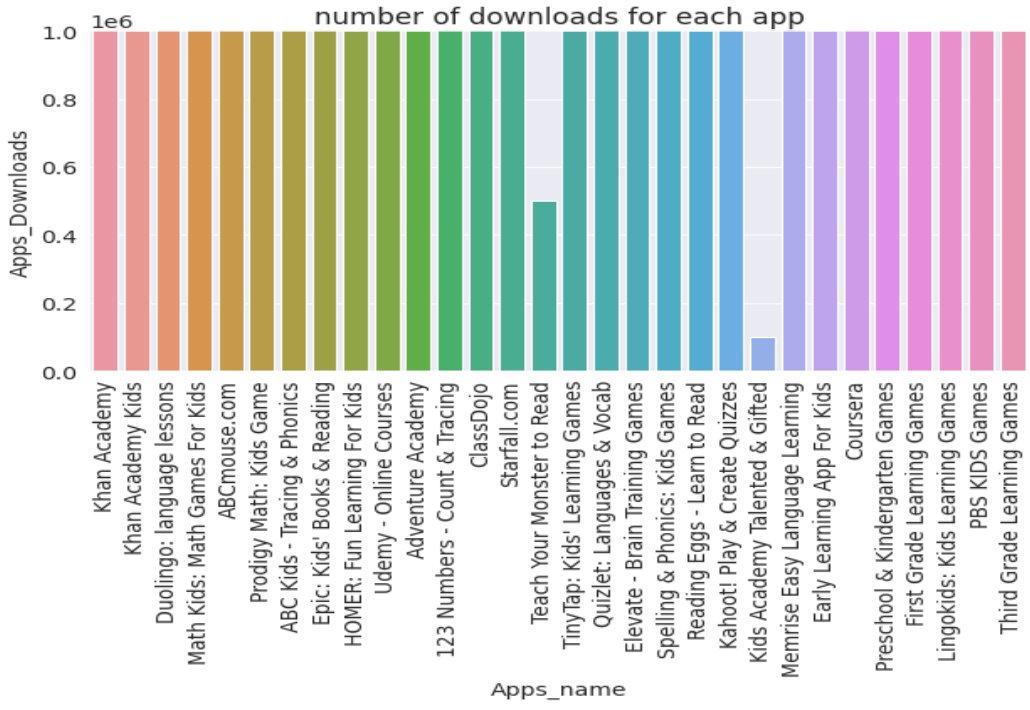
For exploratory data analysis after loading the data in a collab file, first some preprocessing is done to the data where in some columns some special characters are present,

they are removed and then those columns are converted to int or float data types.



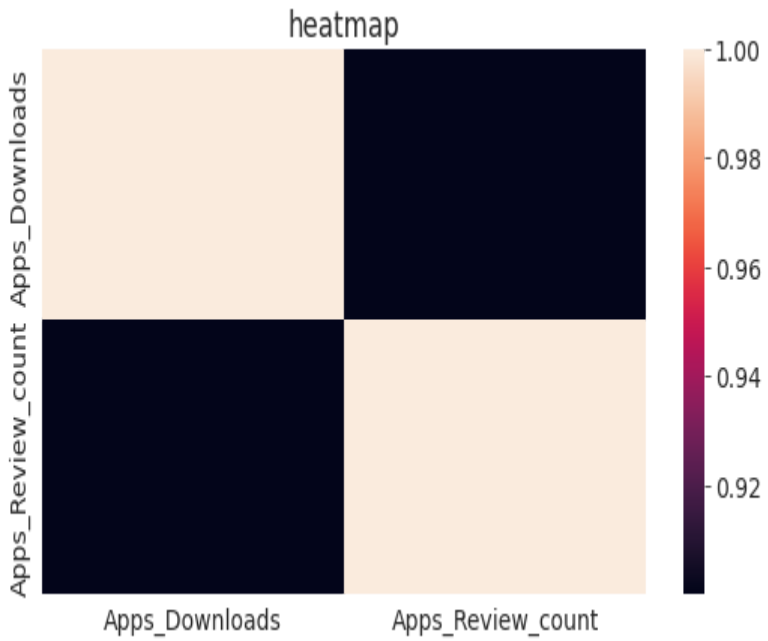
Graph showing the rating variation of different apps

From this graph, we can tell that the apps have ratings in the mid range of 3.5 - 4.7. Out of them most of the apps were given a rating of 4.3 or 4.6.



Box Plot between Apps and the number of downloads

From this graph, we can tell that most of the apps have more than 10M.



Heat map showing corelations

After that, we found the app which has more number of downloads using

```
Apps_with_Highest_rev = data.sort_values(by='Apps_Review_count',
ascending=False).head(20)
```

This code arranges our apps in descending order which gives us that the app with more number of downloads is Duolingo according to the data.

2.2 Scraping the reviews data for Each app

For all the 30 apps in our original dataset, we scraped the reviews data for each and every app using python library- google - play-scraper.

Dataset: EducationalAppsDataScraped.xlsx

For each app out of these 30 apps, I scraped 1000 most relevant reviews along with their rating and thumbs up count. The data for each app is scraped in separate excel sheets.

2.2.1 Analyzing the Reviews Data

I have written a function which plots the countplot of score i.e., rating for each review.

The function is :

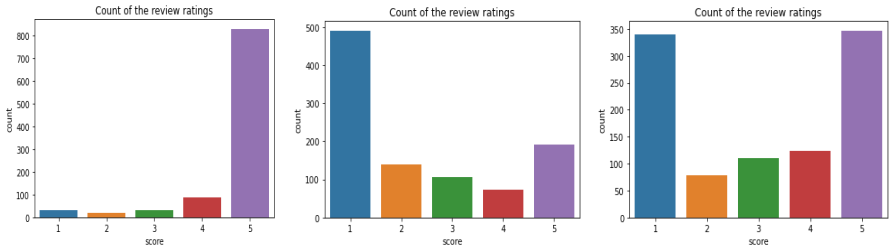
```
def plot(data):

  sns.countplot(data['score'])

  plt.title('Count of the review ratings')

  plt.show()
```

Using this function all the apps' count plots are plotted.



These are the count plots for three apps.

3. Topic Modeling the Reviews:

Topic modeling is recognizing the words from the topics present in the document or the corpus of data. This is useful because extracting the words from a document takes more time and is much more complex than extracting them from topics present in the document. For example, there are 1000 documents and 500 words in each document. So to process this it requires 500*1000 = 500000 threads. So when you divide the document containing certain topics then if there are 5 topics present in it, the processing is just 5*500 words = 2500 threads.This looks simpler than processing the entire document and this is how topic modeling has come up to solve the problem and also visualizing things better.

After importing the required libraries and the data as a dataframe, some preprocessing

must be done to the data to make the modeling technique easier.

3.1 Text Preprocessing

Following steps are performed during the text preprocessing.

- Tokenization - Splitting the text into sentences and sentence into words, lowercasing the words and removing punctuation marks.
- Words smaller than size 3 are removed.
- Stop Words are removed.
- Lemmatization - Words present in third person are converted to first person and words in future tense and past tense are converted into present tense.
- Stemming - Words are converted to their root forms.

A function is being defined to do all the above preprocessing steps for the reviews data of each and every app when it is being called.

```
def lemmatize(text):

    return WordNetLemmatizer().lemmatize(text,pos='n')

#tokenize and lemmatize

def preprocess(text):

    result = []

    for token in gensim.utils.simple_preprocess(text):

        if token not in gensim.parsing.preprocessing.STOPWORDS and len(token) > 3:

            result.append(lemmatize(token))

    return result
```

```
def preprocessing(data):

  processed_docs = []

  for doc in data['content']:

    processed_docs.append(preprocess(doc))

  return processed_docs
```

These are the two functions that are being defined to preprocess the app's review data.

3.2 Creating a Topic Modeling model

After preprocessing the data , a model is created for topic modeling.

Topic modeling is done using LDA(Latent Dirichlet Allocation). Topic modeling refers to the task of identifying topics that best describe a set of documents. These topics will only emerge during the topic modeling process (therefore called latent). And one popular topic modeling technique is known as Latent Dirichlet Allocation (LDA).

Topic modeling is an unsupervised approach of recognizing or extracting the topics by detecting the patterns like clustering algorithms which divides the data into different parts. The same happens in Topic modeling in which we get to know the different topics in the document. It is done by extracting the patterns of word clusters and frequencies of words in the document.

So based on this it divides the document into different topics. As this doesn't have any outputs through which it can do this task hence it is an unsupervised learning method. This type of modeling is very much useful when there are many documents present and when we want to get to know what type of information is present in it. This takes a lot of time when done manually and this can be done easily in very little time using Topic modeling.

3.2.1 Latent Dirichlet Allocation

In LDA, latent indicates the hidden topics present in the data then Dirichlet is a form of distribution. Dirichlet distribution is different from the normal distribution. When ML algorithms are to be applied the data has to be normally distributed or follows Gaussian distribution. The normal distribution represents the data in real numbers format whereas Dirichlet distribution represents the data such that the plotted data sums up to

1. It can also be said as Dirichlet distribution is a probability distribution that is sampling over a probability simplex instead of sampling from the space of real numbers as in Normal distribution.

I created an LDA model creation function using gensim library and when the function is called after the apps are loaded, the function will print the topics identified based on the number of topics given by us.

```
def lemmatize(text):

    return WordNetLemmatizer().lemmatize(text,pos='n')



def preprocess(text):

    result = []

    for token in gensim.utils.simple_preprocess(text):

        if token not in gensim.parsing.preprocessing.STOPWORDS and len(token) > 3:

            result.append(lemmatize(token))

    return result
```

After printing the topics identified for the model for all the 30 apps, we analyze those topics and note the key point present in that and note it.

From those topics we take the issues present in the apps and identify the technical bugs from those issues if any are present and conclude them as the previous bugs of that app.

RESULTS

All the topics are analyzed and noted.

App-1=>ABC Kids

- Topic-1 Phone Version not available for all settings-Technical
- Topic-2 Teachers misspelling the words-NonTechnical
- Topic-3 More time taking for downloading-Technical
- Topic-4 Perfect English, Very Helpful-NonTechnical
- Topic-5 Issue occuring while animation writing.Technical
- Topic-6 Wants even better characters.NonTechnical
- Topic-7 Problem in animation button.-Technical

App-2=>ABC mouse

- Topic-1 Screen freezing in the middle.-Technical
- Topic-2 Slow Loading-Technical
- Topic-3 Charged for one month free trial also-NonTechnical
- Topic-4 Pointer like mouse is annoying-Technical
- Topic-5 Subscription issues-NonTechnical
- Topic-6 Login issues-Technical
- Topic-7 Payment related issue-Technical

App-3=>BYJU'S

- **Topic-1 Starting Problem-**Technical
- **Topic-2 Problem coming in the middle of video-**Technical
- **Topic-3 Extra content more than subject-Non**Technical
- **Topic-4 Issue when receiving a call-**Technical
- **Topic-5 Blank Screen after updating-**Technical
- **Topic-6 Issues regarding calling and talking-Non**Technical
- **Topic-7 Access issues-**Technical

App-4=>Brainly

- **Topic-1 Starting Problem-**Technical
- **Topic-2 Connection Problem-**Technical
- **Topic-3 Downloading issue-**Technical
- **Topic-4 Scanning problem during question answering-**Technical
- **Topic-5 Ranking issues-Non**Technical
- **Topic-6 Problem in the question asked-Non**Technical
- **Topic-7 Asking for the subscription money-Non**Technical

App-5=>Coursera

- **Topic-1 Poor quality of videos-**Technical
- **Topic-2 Internet connection issues-**Technical
- **Topic-3 More Time taking for downloading-**Technical
- **Topic-4 eMail receiving issues**Technical
- **Topic-5 Poor quality certificate-Non**Technical
- **Topic-6 Time taking for receiving the certificate-Non**Technical

App-6=>Diksha

- **Topic-1 Login Issues-**Technical
- **Topic-2 Certificate Downloading problem-Non**Technical
- **Topic-3 Time taking for module loading-**Technical
- **Topic-4 Problem in showing the amount of content completed-**Technical
- **Topic-5 Server Issues-**Technical

App-7=>DoubtNut

- **Topic-1 Loading Notification issues-**Technical
- **Topic-2 Showing more notifications-Non**Technical
- **Topic-3 No language option-Non**Technical
- **Topic-4 Giving notification for the some problem solved a long ago-**Technical
- **Topic-5 Searching time more-**Technical
- **Topic-6 Video Solutions only present, not text solutions-Non**Technical
- **Topic-7 No proper solution suggestions-Non**Technical

App-8=>Duolingo

- **Topic-1 Screen Freezing-**Technical
- **Topic-2 Unable to find spanish Language-Non**Technical
- **Topic-3 Payment deleted issues-**Technical
- **Topic-4 Correcting mistakes well-Non**Technical
- **Topic-5 Mistake in making words form sentences-Non**Technical
- **Topic-6 Good practice along with learning-Positive**

App-9=>Embibe

- **Topic-1 Less study content-Non**Technical
- **Topic-2 Felt Useless due to bugs-Non**Technical
- **Topic-3 Less time for problem solving-Non**Technical
- **Topic-4 All learning subjects like science not available-Non**Technical
- **Topic-5 Number of characters in password issue-**Technical
- **Topic-6 Video lagging issues-**Technical

App-10=>Eduauraa

- **Topic-1 Waste of Subscription money-NonTechnical**
- **Topic-2 Time taking for video-**Technical
- **Topic-3 Content available without registration also-**Technical
- **Topic-4 Felt Fake-Non**Technical
- **Topic-5 Good Video-Positive**

App-11=>Epic

- **Topic-1 Payment related issues-**Technical
- **Topic-2 More free time-Non**Technical
- **Topic-3 Login related issue-**Technical
- **Topic-4 Freely available but did payment-**Technical
- **Topic-5 Unlimited access**-Technical
- **Topic-6 Amount paid for trial period also-Non**Technical

App-12=>ExtraMarks

- **Topic-1 Trust related issues, Felt Fraud-Non**Technical
- **Topic-2 Attempted install-Non**Technical
- **Topic-3 Speed less between web pages-**Technical
- **Topic-4 Want video content more-Non**Technical
- **Topic-5 Problem in presenting slides-**Technical
- **Topic-6 Showing correct password even when wrong password is entered-**Technical

App-13=>GreatLearning

- **Topic-1 Unable to review-Non**Technical
- **Topic-2 Logging in slow-**Technical
- **Topic-3 Updated free course-Non**Technical
- **Topic-4 Available offline-Positive**
- **Topic-5 Lagging Issue while taking quiz-**Technical
- **Topic-6 Downloading certificate issue-Non**Technical

App-14=>KhanAcademy

- **Topic-1 Suggesting to test before learning-Positive**
- **Topic-2 Error in loading task-**Technical
- **Topic-3 Video Crashing-**Technical
- **Topic-4 Need classification of topics more-Non**Technical
- **Topic-5 Zooming problem-**Technical
- **Topic-6 Connection problem-**Technical
- **Topic-7 Couldn't star video in phone-Non**Technical

App-15=>KhanKids

- **Topic-1 Time taking to swipe screen-**Technical
- **Topic-2 Many similar apps available-Non**Technical
- **Topic-3 eMail update option unavailable after updating-**Technical
- **Topic-4 Frozen Screen , On reset crashed-**Technical
- **Topic-5 Unavailable to get password recovery mail for Samsung phones-**Technical

- **Topic-6 Crashing during loading-**Technical

App-16=>Kutuki

- **Topic-1 Language Problem-Non**Technical
- **Topic-2 Not able to change baby year-Non**Technical
- **Topic-3 Choosing language option not present-**Technical
- **Topic-4 Not getting report of kid-Non**Technical
- **Topic-5 Unable to download app freely-**Technical
- **Topic-6 Financial error-Non**Technical
- **Topic-7 Feedback sending issue-**Technical

App-17=>MathKids

- **Topic-1 Hard for children-Non**Technical
- **Topic-2 Pronunciation language changing option available-Positive**
- **Topic-3 Very Useful for kids-Positive**
- **Topic-4 Helping kids stop playing games-Positive**
- **Topic-5 Followed day to day content-Positive**
- **Topic-6 Learning through game-Positive**
- **Topic-7 Math learn from games-Positive**

App-18=>Memrise

- **Topic-1 Good looking Background effects-Positive**
- **Topic-2 Latest version is good with subscription-Positive**
- **Topic-3 Free Audio Option-Positive**
- **Topic-4 Lesson and also memorizing well-Positive**
- **Topic-5 Looking like free but requires subscription-Non**Technical
- **Topic-6 Subscription issues-**Technical
- **Topic-7 Good content-Positive**

App-19=>Oda

- **Topic-1 Thinking there is some fraud- trust issues-Non**Technical
- **Topic-2 Bridge course was not so good-Non**Technical
- **Topic-3 Accepting fake applications-Non**Technical
- **Topic-4 Wastely paid money for the bridge course-Non**Technical
- **Topic-5 Refund issues-Non**Technical

- **Topic-6 Worst Experience-Non**Technical
- **Topic-7 Teaching the same course again again-Non**Technical

App-20=>Prodigy

- **Topic-1 Account related issues-**Technical
- **Topic-2 Asking for membership-Non**Technical
- **Topic-3 Children learning math with game-Non**Technical
- **Topic-4 Time related issues-Non**Technical
- **Topic-5 Nice idea of pet monster-Positive**
- **Topic-6 Membership issue-Non**Technical
- **Topic-7 Hard work required-Non**Technical

App-21=>TeachMint

- **Topic-1 Good problem solving along with starring-Positive**
- **Topic-2 Internet connection issues-**Technical
- **Topic-3 Class opening issue-**Technical
- **Topic-4 No notification option-**Technical
- **Topic-5 Long live classes-Non**Technical
- **Topic-6 On reviewing start form the beginning-Non**Technical
- **Topic-7 Recording feature available-Positive**

App-22=>ThirdFlix

- **Topic-1 Good Problem experience-Positive**
- **Topic-2 Login problem-**Technical
- **Topic-3 Video downloading taking more data-**Technical
- **Topic-4 Problem with interface-Technical**
- **Topic-5 Video timing issues-Non**Technical
- **Topic-6 Chat issues in zoom-Technical**
- **Topic-7 Time wasting sessions-NonTechnical**

App-23=>TinyTap

- **Topic-1 Refund related issues-Non**Technical
- **Topic-2 Subscription related issue-Non**Technical
- **Topic-3 Account subscription canceled-Non**Technical
- **Topic-4 Freezing game-**Technical

- **Topic-5 Time taking game-Non**Technical
- **Topic-6 Card payment related issues-Non**Technical
- **Topic-7 Sending emails for subscription-Non**Technical

App-24=>Toppr

- **Topic-1 Money payment required for doubt answer-Non**Technical
- **Topic-2 Payment issue-**Technical
- **Topic-3 No video solution available-Non**Technical
- **Topic-4,5,6,7 Good reviews-Positive**

App-25=>Udemy

- **Topic-1 Time to time work update-Non**Technical
- **Topic-2 Login issues-**Technical
- **Topic-3 Wanted offline video downloading option-Non**Technical
- **Topic-4 Video playing issues-**Technical
- **Topic-5 Working time issue-Non**Technical
- **Topic-6 Payment related issues-**Technical
- **Topic-7 Working time issue-Non**Technical

App-26=>Unacademy

- **Topic-1 Good Course, Live free class along with interaction-Positive**
- **Topic-2 Test Update issue-Technical**
- **Topic-3,4,5 ,6,7 Positive Reviews**

App-27=>Vedantu

- **Topic-1 Good live experience-Positive**
- **Topic-2 Good App but time issues-**Technical
- **Topic-3,4 Positive**
- **Topic-5 Content Issue-Non**Technical
- **Topic-6 Good class-Positive**
- **Topic-7 Messaging issue to the teacher-**Technical

App-28=>ePathshala

- **Topic-1 Positive**
- **Topic-2 Internet issues-**Technical

- **Topic-3 Book downloading available-Positive**
- **Topic-4 Time issues-Non**Technical
- **Topic-5 ,6,7 Positive reviews**

App-29=>edX

- **Topic-1 Browser unable to download-**Technical
- **Topic-2 Difficulty in exam-Non**Technical
- **Topic-3 Crashing-**Technical
- **Topic-4 Error in loading the course page-**Technical
- **Topic-5 Registration issues-**Technical
- **Topic-6 Good Experience-Positive**
- **Topic-7 Crashing device while watching video-**Technical

App-30=>uLesson

- **Topic-1 : Opening, downloading and subscription issues-**Technical
- **Topic-2 : Not working properly , update issues-**Technical
- **Topic-3: Time delay issues-**Technical
- **Topic-4: Positive reviews**
- **Topic-5: Good reviews**
- **Topic-6: Loading issues, time delay and download issues-**Technical
- **Topic-7: Working and update issues mostly regarding data and time-**Technical

CONCLUSION:

The Technical reviews that are mostly present according to the CWE view

Issues	CWE Errors	Code
Internet Connectivity Issues	Signal Errors	387
Login Issues	Authentication Errors	1211
Crashing	UncaughtException Permission issue	248
Password Issues	Cryptographic issues- Weak Encoding for passwords	261
Loading time issues	Complexity issues	1226
Fake/ Fraud Certificates	Improper Certificate validation	295
Password forgotten	Weak Password Recovery Mechanism for Forgotten Password	640
Payment related Issues	Business Logic Errors	840
Issue in notification receival	Behavioral Error	438
Video resolution issue	Insufficient Visual Distinction of Homoglyphs Presented to User	1007
Problem with interface	User interface Security issue	355

The colab files containing all the data scraping and also topic modeling done are below.

Scrapper-https://colab.research.google.com/drive/1sBu4b-jJQsP4sQG3LkkiH6iifDg9BGBW?usp=sharing

Exploratory Data

Analysis-https://colab.research.google.com/drive/1D_8pzB8EQYncNoRIxED1wgKjHnaKPRyz?usp=sharing

Topic Modeling-https://colab.research.google.com/drive/1FRY4GWubMf3hTBrxIPL2VT-UUTw8b1Eb?usp=sharing

REFERENCES

1. https://play.google.com/store/search?q=Education%20Apps&c=apps
2. https://colab.research.google.com/github/dipanjanS/nlp_workshop_odsc19/blob/master/Module05%20-%20NLP%20Applications/Project04%20-%20Topic%20Modeling.ipynb#scrollTo=W1WPQBOBW7pq
3. https://www.kaggle.com/code/prakharprasad/mobile-reviews-topic-modeling
4. https://cwe.mitre.org/

THE END