Data Science Fundamentals

(CS 890ES)

Instructor: Dr. Alireza Manashty
Department of Computer Science
University of Regina
Winter 2020

Executives User Guide
on
"Games Description Analysis & Keyword Suggestions"

Submitted by:

Name	**ID**	**Email**	**Role**
Sai Keerthi Mettu	200416252	smj102@uregina.ca	BI Analyst Data Engineer Data Scientist Communicator
Somi Deepthi Nalamalpu	200412879	sny899@uregina.ca	

Table of Contents

INTRODUCTION:

The present-day technology evolution has made many changes and provided everything at fingertips. This evolution has helped in the extraction of new fundamentals and terminologies effecting and operating the world today in every possible way. Websites and applications are some of those terminologies that have evolved and encouraged us to use them and procreate advanced technologies to help ourselves out. The change in generations and the way of thought processing and imaginations of mankind allowed users to just focus on making everything available in their hands itself. Games and video games have been upgraded from television connected version to online and application-oriented mediums. We now have a game available on many platforms and mediums at users' convenience.
Digital games that elongate to Xbox versions, PlayStation accessibility, Online versions, Mobile application versions and many more means of accessibility have been designed and developed which has features that correspond to its growth and sustainability in the market. The gaming industry is a huge platform for the developers and Google Play store is the biggest marketplace for the games. Statistics show that there are 10 million people playing games every minute and new developers count increase at a rate of 1 million every year. There is a need for a survey or analysis needed by the self-employed game developers to design a game. As of now, there is no automated application or website that shows the updated information about these data insights for the developers. In this project, we would like to apply our focus on games available on the mobile medium and the appropriate game description written for it. As we can observe we have different games each categorized into different genres like adventure, action, racing, puzzle, etc.

Each game constitutes having a special description that describes the basic features and functionality of the game. This description is considered unstructured data because it cannot be easily computed to derive any insights out of it. In order to make it useful for the new game developers, this data can be used. Text extraction techniques are identifying the keywords using the NLP methodologies would help in the resultant outcome derivation. So, the prediction of these keywords is possible when there is a combination of these NLP techniques on the text with the machine learning algorithm. Data science life cycle is an approach to organize this data and produce the results out of this text content from different games. All the steps within this life cycle are explained and detailed with the self-created games dataset for this project.

PROBLEM STATEMENT:

In our project, the main aim is to analyze the description available for an individual game and explore the text in ways, that can help in giving the users some inputs on how description can be amplified. The problem we are trying to analyze focus on suggesting certain keywords or phrases that help elongate the description of a game in a more read full way. To provide certain inputs for the description to be attraction and help the users to have a better viewpoint on the game. So, extracting and suggesting important phrases that can be embedded in the game description to help the developers with an intellectual and sophisticated game intro is the main concern.

SOLUTION OVERVIEW:

In a detailed study, it is identified that the game description provides in the home page of a particular app in the play store plays a key role in its success and download rate of the game. To attain these success results, it is necessary to study the data and know about different text analysis and classification methods, as the problem statement itself contributes to the factor that this project aims at text data. So, in achieving the above task or operating the problem statement, it is very much an essential factor that we have access to the data the gives information about unique games and its game description as well. To support this, we created a dataset by web scrapping the play store website to identify unique games and contributing factors to achieve the task. And once the dataset is generated, we applied different text analysis and visualization techniques to clean the dataset and make it operative enough to feed it to the model. TF-IDF (Text Frequency-Inverse document frequency) vectorizer is used to divide and explore the text and obtain the word matrix for each game description. The obtained unique features were passed to the machine learning model, for it to get trained and form clusters with appropriate words that can be suggested to the user when provided a game description. Providing more useful and content related phrases that can improve the intro or description of a game making it reachable to most of the users is the main concern. And in the process of achieving it, many other operations have been performed on the data available which will be clearly discussed further sections.

BUSINESS VALUE:

There is always a growing streak in the rate of game developers in any platform. Stats show that every year there is a rise of these developers count by 5 million in all platform resources altogether. The gaming industry generated almost $135 billion in 2018. The prediction is that in 2019 it will increase to generate 152.1 billion. In the entire world, the major proportion of working people belong to the USA, which comprises 2457 companies that support up to 220thousand jobs in its country alone [8]. Also, it is expected that by 2020 the revenue generation from the gaming industry would be approximated to 160+ billions in USD. Which means the growth rate of jobs and self-employed developers would rise to around 25 million. As their growth rate is constantly increasing both in terms of investment and development, the demand for skilful developers with new ideologies are highly welcomed. This project helps those kind of developers in achieving their initial survey and analysis about the current trends through keywords extractions and suggestions. Clearly from these figures it can be stated that the graph will keep a constantly rising curve of revenue generations that are yielded at regular intervals.

PROJECT BENEFITS:

There would always be a puzzling situation a new game developer. Sometimes they would not be clear about the idea and type of features, functionalities expected in a game. In order to provide some research insight, this is the project that guides them what to do. Apart from just game developers, there are also ways for knowledge for the people who are interesting in knowing the facts about the google play store trends. As this is solution developed based on text extraction combined with machine learning model implementations, this project could be an example for data engineers who wants to resolve a problem of similar kind. This kind of problematic approach could be said as one of its kind and it can help the one's looking to solve it. By rendering accurate results to the audience would increase billions of revenue each year to the gaming industry and can keep a constant streak of success for the vendors or the stakeholders. Out of all these benefits, the main objective of this project, i.e., to provide assistance for the amateur developers who are ready to design a game can be the most beneficial people.

TOOLS:

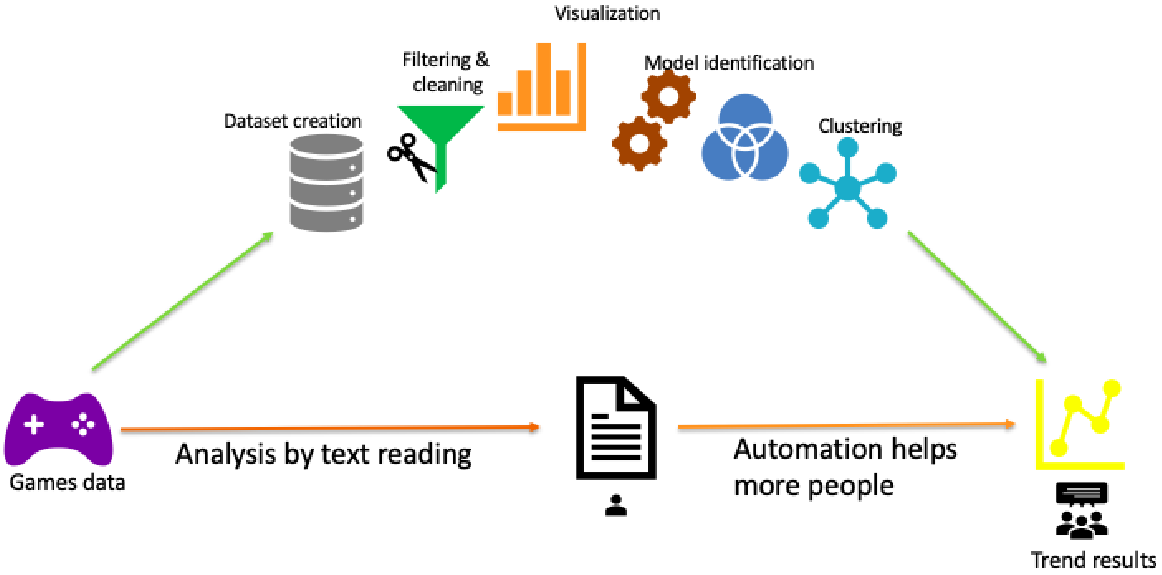
These are the tools utilized for the construction of this project:
Python
Google Colab
Google Drive
Google play scrapper, Beautiful Soup for scrapping
Numpy, pandas, string, re for computations
NLTK library for text extraction
SKLearn for vectorization and modelling
Matplot, Seaborn for visualization

ROLE OF EXECUTIVES IN DATA ANALYTICS LIFE CYCLE:

Data Analytics life cycle consists of six phases. Executives are the users that needs to use the end product. So executives are people who travels with the project team from the beginning till the end of a project development process. Executives are also the part of stakeholders group who provide the team with costs of development in each phase. These are the people responsible for giving business to achieve a product what they are looking for. As data analytics life cycle is an iterative process, all each phase of the development cycle, the executives are communicated about the changes, measures and implementations.

To be noted being on a same page as that of the expectations and development results, the project team would update the deliverables and the statuses of changes happening in the process. Likewise, in this project development, the instructor of the course was considered as our key executive and we constantly took updates for the project modifications. These periodical updates ensures that executives are receiving the product as expected.Thus, executives play a key role from the start of development till the final deliverable is hosted on to the web server.

PROJECT DEVELOPMENT:



The development of this project began from scratch like creation of a data set by web scrapping. 15k records of games data was web scrapped altogether from the google play store in different batch files. Since this project is all about the text analysis and keyword suggestions, we considered only the game description feature of all the game records in order to build the model. Post this consideration, a whole bunch of ETL operations like data cleaning with stop words, emojis, punctuations and URLs removal are performed on the text description feature. For better text processing and filtering, operations like tokenization, stemming and lemmatizing were also applied it. Till here the data cleansing and preparation are done. Moving on to the model development, the text feature was vectorized where all the words are counted and made into columns having the inverse ratio of their counts. This vectorized data is presented to the clustering model where data is grouped based on the similarity of the vectorized data under the influence of other main features like games genre. This way the model gets trained and grant us with its predictions. A consistent iteration process of this model training and key words cleaning before modelling could yield better accuracies and predictions whenever complicated descriptions are provided to the user.

EXECUTION REQUIREMENTS:

This project is all about rendering a research analysis of games from the Google play store. Executives are the people who are going to use the software at the end.

In order to see the project results, a software set up has to arranged and to do this and a series of instructions were documented in the setup guide for this project. On a higher level of understanding, an executive needs the following. They are:

Project code files

Google account to access google drive

Google colab

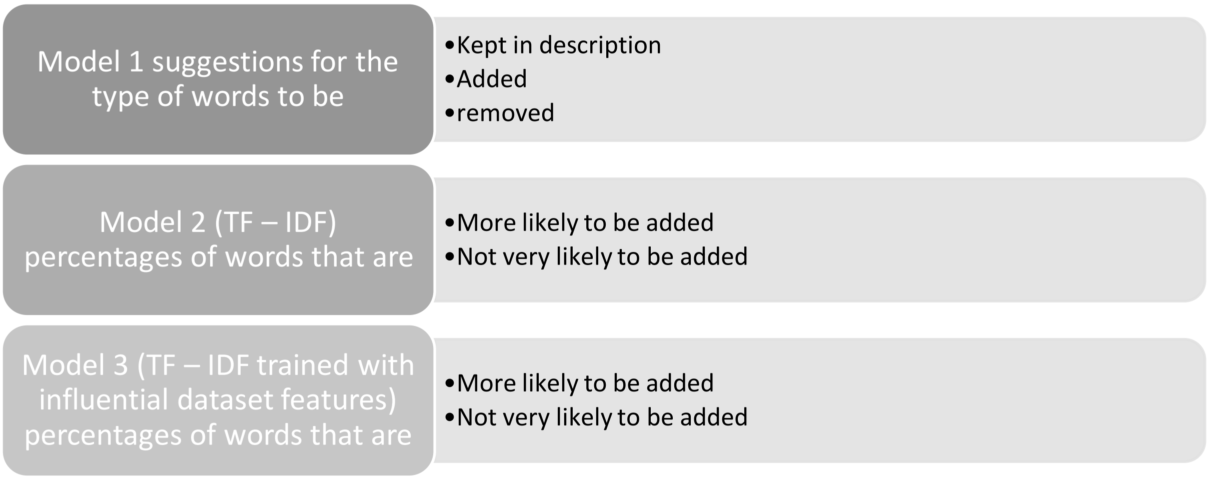
Access permission to the dataset

Web link to connect to the application

If one has a google account and permission to the games dataset available in google drive, they can open the google colab in the browser to see the results. Here the project file i.e., keywords_suggestions.py file is loaded, give a test description at the end of the file and results can be seen in the same file itself. In future this project can be hosted on to a web server to see in a better representation of the application.

OUTPUT DETAILS:

This is the image that shows the output that user wants to lookup to. Here is the detailed discussion on what the executive sees:



Model 1 suggestions for the type of words to be

- Kept in description
- Added
- removed

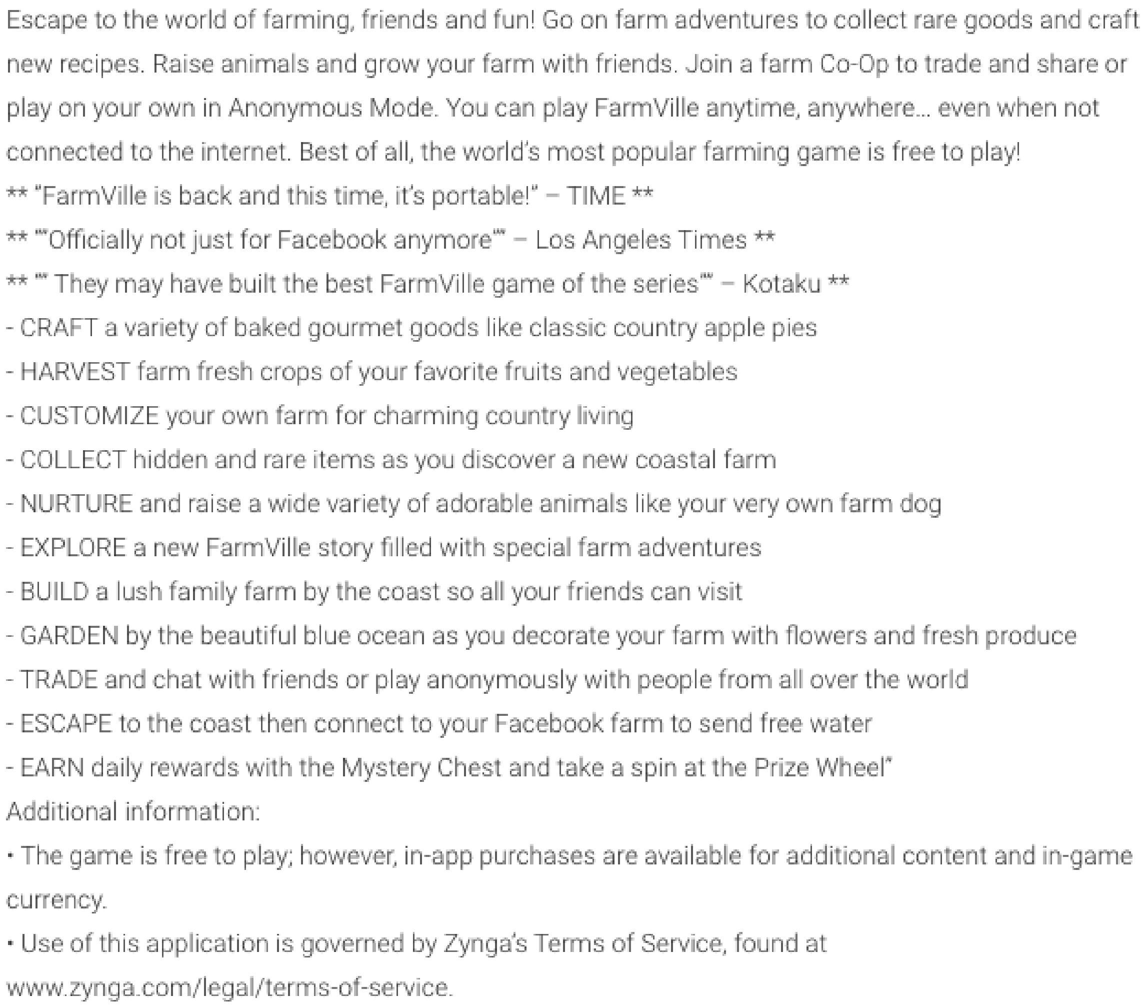
Model 2 (TF – IDF) percentages of words that are

- More likely to be added
- Not very likely to be added

Model 3 (TF – IDF trained with influential dataset features) percentages of words that are

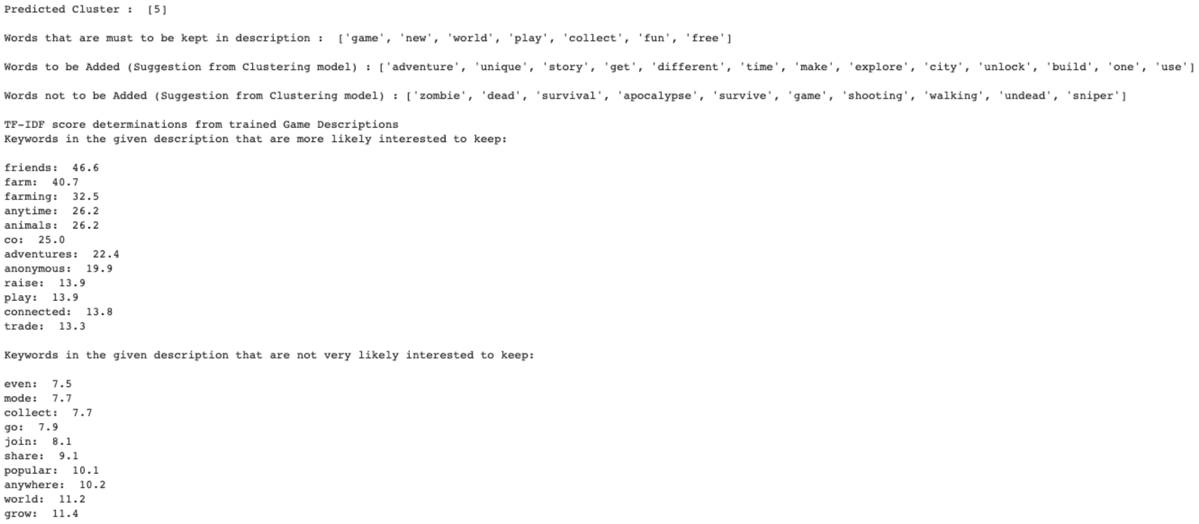
- More likely to be added
- Not very likely to be added

Now let’s see when a sample existing game description is give for testing, how this model provide its suggestions:



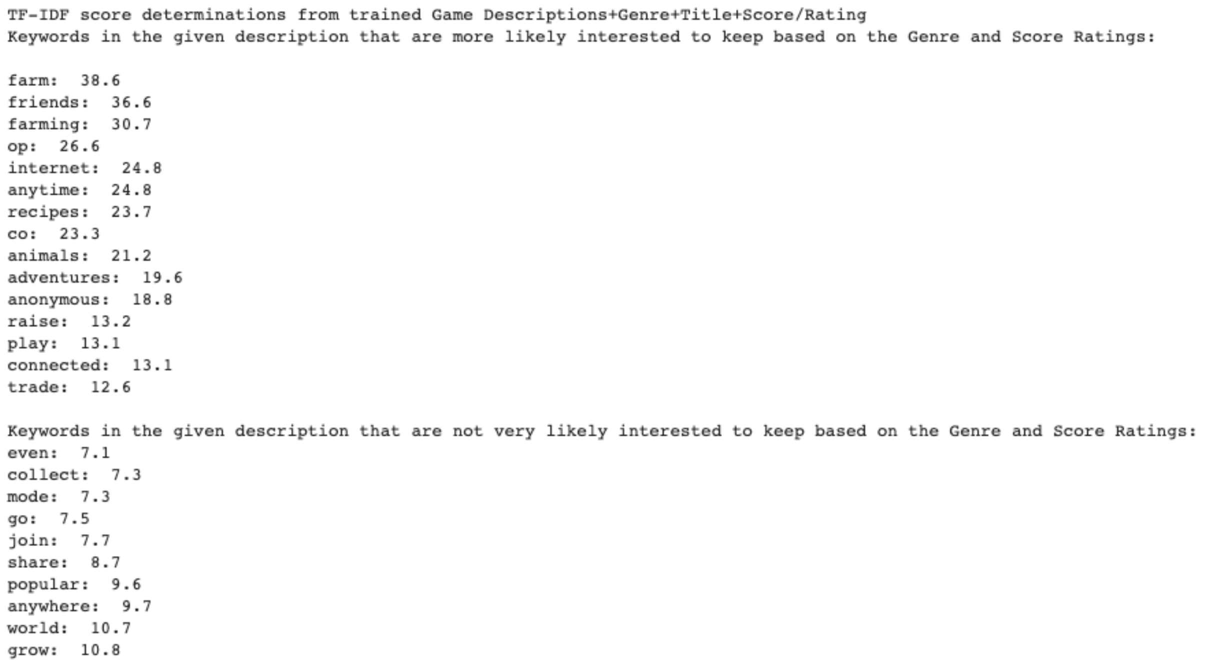
Escape to the world of farming, friends and fun! Go on farm adventures to collect rare goods and craft new recipes. Raise animals and grow your farm with friends. Join a farm Co-Op to trade and share or play on your own in Anonymous Mode. You can play FarmVille anytime, anywhere… even when not connected to the internet. Best of all, the world’s most popular farming game is free to play!
** “FarmVille is back and this time, it’s portable!” – TIME **
** ““Officially not just for Facebook anymore”” – Los Angeles Times **
** “” They may have built the best FarmVille game of the series”” – Kotaku **
- CRAFT a variety of baked gourmet goods like classic country apple pies
- HARVEST farm fresh crops of your favorite fruits and vegetables
- CUSTOMIZE your own farm for charming country living
- COLLECT hidden and rare items as you discover a new coastal farm
- NURTURE and raise a wide variety of adorable animals like your very own farm dog
- EXPLORE a new FarmVille story filled with special farm adventures
- BUILD a lush family farm by the coast so all your friends can visit
- GARDEN by the beautiful blue ocean as you decorate your farm with flowers and fresh produce
- TRADE and chat with friends or play anonymously with people from all over the world
- ESCAPE to the coast then connect to your Facebook farm to send free water
- EARN daily rewards with the Mystery Chest and take a spin at the Prize Wheel”
Additional information:
• The game is free to play; however, in-app purchases are available for additional content and in-game currency.
• Use of this application is governed by Zynga’s Terms of Service, found at www.zynga.com/legal/terms-of-service.

This is the description of an existing game in play store



```
Predicted Cluster :  [5]

Words that are must to be kept in description :  ['game', 'new', 'world', 'play', 'collect', 'fun', 'free']

Words to be Added (Suggestion from Clustering model) : ['adventure', 'unique', 'story', 'get', 'different', 'time', 'make', 'explore', 'city', 'unlock', 'build', 'one', 'use']

Words not to be Added (Suggestion from Clustering model) : ['zombie', 'dead', 'survival', 'apocalypse', 'survive', 'game', 'shooting', 'walking', 'undead', 'sniper']

TF-IDF score determinations from trained Game Descriptions
Keywords in the given description that are more likely interested to keep:

friends:  46.6
farm:  40.7
farming:  32.5
anytime:  26.2
animals:  26.2
co:  25.0
adventures:  22.4
anonymous:  19.9
raise:  13.9
play:  13.9
connected:  13.8
trade:  13.3

Keywords in the given description that are not very likely interested to keep:

even:  7.5
mode:  7.7
collect:  7.7
go:  7.9
join:  8.1
share:  9.1
popular:  10.1
anywhere:  10.2
world:  11.2
grow:  11.4
```

Keyword suggestions from Model and TF-IDF vectorized data

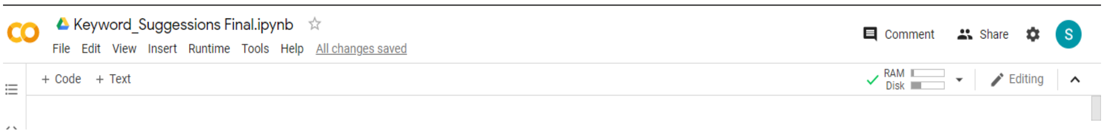


```
TF-IDF score determinations from trained Game Descriptions+Genre+Title+Score/Rating
Keywords in the given description that are more likely interested to keep based on the Genre and Score Ratings:

farm:  38.6
friends:  36.6
farming:  30.7
op:  26.6
internet:  24.8
anytime:  24.8
recipes:  23.7
co:  23.3
animals:  21.2
adventures:  19.6
anonymous:  18.8
raise:  13.2
play:  13.1
connected:  13.1
trade:  12.6

Keywords in the given description that are not very likely interested to keep based on the Genre and Score Ratings:
even:  7.1
collect:  7.3
mode:  7.3
go:  7.5
join:  7.7
share:  8.7
popular:  9.6
anywhere:  9.7
world:  10.7
grow:  10.8
```

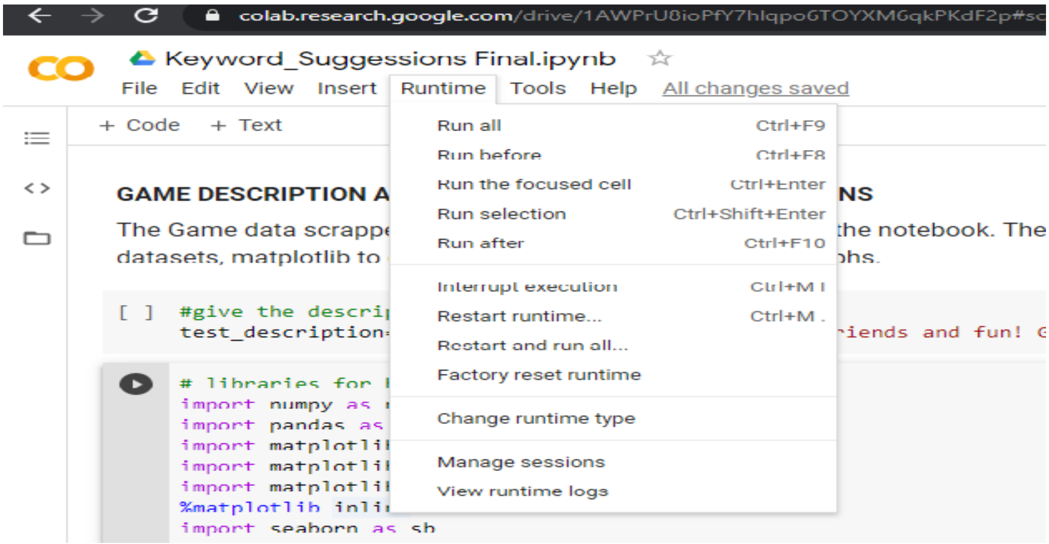
Keywords that are recommended by the TF-IDF vectorised object that is trained with game genre, game ratings score and the title

STEPS TO BE PERFORMED:

- User needs to download theGaming.csv and BatchFile1.csv and store it their Google Drive.
- The Keyword suggestions.ipynb available in the GitHub can be opened with Colab (Google Collaboratory).
- The user needs to sign in his Google account to get the runtime and working environment in Google colab or create an account if it doesn't exist and then connect to a local runtime.



- Once the Keyword suggestions.ipynb is imported on to your colab environment allot the runtime and check if the drive path is correct. The path specified should be the same as the one where the dataset is tore. Make sure it is accurate.
- There exist a text description variable where user can give their game description within the inverted commas. The variable takes the user input and provides it to the model as test data.
- Then the user can execute the code by opting for "Run ALL" at runtime to attain the suggestions ignoring the remaining methodology or can run each cell to view and analyze the output of each technique and operation performed on the data.


colab.research.google.com/drive/1AWPrU8ioPfY7hlqpo6TOYXM6qkPKdF2p#sc
Keyword_Suggestions Final.ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text
Run all Ctrl+F9
Run before Ctrl+F8
Run the focused cell Ctrl+Enter
Run selection Ctrl+Shift+Enter
Run after Ctrl+F10
Interrupt execution Ctrl+M I
Restart runtime... Ctrl+M .
Restart and run all...
Factory reset runtime
Change runtime type
Manage sessions
View runtime logs

GAME DESCRIPTION A NS

The Game data scrappe the notebook. The datasets, matplotlib to phs.

```
[ ]  #give the descri                         riends and fun! G
     test_description=

     # libraries for l
     import numpy as r
     import pandas as
     import matplotlil
     import matplotlil
     import matplotlil
     %matplotlib inlir
     import seaborn as sb
```

CONCLUSION:

Through a constantly iterating data science life cycle process, the unstructured data is represented to the users in a useful way. The problem of identifying trends in the gaming industry has been solved by achieving at the steps of the lifecycle using the dataset that we created. A new developer would get some insight from the suggestions provided by this structured model. These keyword recommendations are some of the first of its kind application developed on the play store data. This project can be taken to the next level by improving the model performance and design quality of the web application. Alongside, the same dataset can be expanded by adding more games from different platforms and help a larger crowd of the developers.

ACKNOWLEDGEMENT:

We are very thankful and highly indebted to Dr Alireza Manashty for his guidance and thought-provoking ideas to work with python programming, NLP machine learning implementation on text extraction & also for his continued support in completing the project. He has been like one of the executives and project guide from the start of this project. His continual assistance in every step of the project development has pushed us ahead in everything we are doing.

Git Hub Link to access the project files:
https://github.com/KeerthiMettu/Games-Analysis

https://github.com/somideepthi/Game-description-analysis-and-keyword-suggestion

Access links for our 2 datasets from Google drive:

1. BatchFile1.csv:
   https://drive.google.com/open?id=18cRi8kB54mGyih2ufc1pPr6hHvxPDbJT
2. originalgaming.csv:
   https://drive.google.com/file/d/10-E7sLMq1P2E8Ubv0IrlCugyXrsyfoHw/view