

**Report**

**For**

**Movie Data Analysis**

**(Using Data Science with GUI)**

**CSE 308 – Computing Project**

**SUBMITTED TO: Baljit Singh Saini**

**Department of Computer Science**

**SUBMITTED BY:**

**Prasanth G. 11601526**

**Nikhil JSK 11601789**

**Section K1632**

**DECLARATION**

We hereby declare that the project work entitled Movie Data Analysis is an authentic record of our own work carried out as requirements of Project of course CSE-308 for the award of B.Tech degree in Computer Sciences from Lovely Professional University, Phagwara, under the guidance of (Baljit Singh Saini), during January to April 2016. All the information furnished in this capstone project report is based on our own intensive work and is genuine.

Project Group

Name of Student 1: Nikhil JSK

Registration Number: 11601789

Name of Student 2: Prasanth Gelli

Registration Number: 11601526

(Signature of Student 1)

Date:

(Signature of Student 2)

Date:

**CERTIFICATE**

This is to certify that the declaration statement made by this group of students is correct to the best of my knowledge and belief. They have completed this Project under my guidance and supervision. The present work is the result of their original investigation, effort and study. No part of the work has ever been submitted for any other degree at any University. The Project is fit for the submission in Computing Project - II subject, Computer Science & Engineering from Lovely Professional University, Phagwara.

**Signature and Name of the Mentor**

**Designation**

**School of Computer Science and Engineering,**

Lovely Professional University,

Phagwara, Punjab.

Date :

**ACKNOWLEDGEMENT**

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to extend my sincere thanks to all of them.

I am highly indebted to Lovely Professional University for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project. We would like to express my gratitude towards member of Lovely Professional University for their kind co-operation and encouragement which help me in completion of this project.

I would like to express my special gratitude and thanks to industry persons for giving me such attention and time. Our thanks and appreciations also go to our QT mentor in developing the project and people who have willingly helped us out with their abilities.

(Signature of Student 1)

Date:

(Signature of Student 2)

Date:

**Table of Contents**

1. Introduction  
   1. Abstract
   2. Purpose
   3. Technologies used
2. Problem Survey  
   1. System Study
   2. Challenges
   3. Scope of project
3. Design  
   1. System Analysis and Design
      1. Problem Statement
      2. Feasibility Study
         1. Economic Feasibility
         2. Technical Feasibility
4. Product Descriptions  
   1. Modules and Code components
5. Software Requirements

5.1 General Description

5.2 Application Requirements

1. Proposed Methodology of Design  
    6.1 DFD  
   1. Use Case Diagram

6.3 Flow Chart

1. Test Cases
2. Reference
3. ***Introduction***
   1. ***Abstract***

The Project helps the user in understanding the dataset “Movie\_data”, the data all by itself if highly unorganized, and nearly impossible to draw inferences from merely visualization of raw data.

It is a GUI based interface (QT) using Python-3 which is capable of asking the query from user in text format or even in Image format and get the answer from the dataset. We’ve used Natural Language Processing and Image processing libraries to achieve this. The solution contains different types of graphs and text outputs to give a holistic comprehensibility to the end user.

The end-goal is to use this kind of system to any given dataset, making data science not just easy, but reaching out more people and spreading the usability of data science.

* 1. ***Description***

A data-science based system integrated with a user-friendly and compelling UI where the user can upload the data, communicate with the system in a high-level language (English), use different graphs to visualize the data points. And, all of that is done only through typing the text in the tool and not a single line of code, where real-time processing is done in the backend.

There are basically a few sessions of the system, available to the user.

* Welcome Screen
* Dashboard
  + Image Input
  + Text Input
* Result screen
  + Graphs
  + Text based ouput
  + Inferences in form of text
  1. ***Technologies used***

A lot of technologies and tools related to Data science and machine learning are used. Visualization tools and image processing tools really help the user to gain holistic insights of the dataset.

* GUI
  + PyQt5
* Image Processing
  + OpenCV
  + Pytesseract
* Natural Language Processing
  + Tensorflow
  + Tokenizer
* Data Science & Visualization
  + Matplotlib
  + Seaborn
  + Pandas
  + Numpy

***2. Problem Survey***To extremely stream-line the process of Data Analysis in the field of Data science. Helps every data-enthusiast to pre-process the data, visualize, gain insights of the data just while communicating with *DUX* in high-level language (English). This tool is highly influential to prove that Data Science is highly powerful, simultaneously demonstrating how easy it is to use Data to produce significant and critical results for any domain in a global perspective. This tool will prove that analysis of any given data can be done ***without*** a single line of code.

A user-friendly and compelling UI where the user can upload the data, communicate with the system in a high-level language (English), use different graphs to visualize the data points, gain insights about the data in-hand. And, all of that is done only through typing the text in the tool and not a single line of code, where real-time processing is done in the backend. These are the basic three steps in a conventional data science field, all streamlined to achieve the best productivity along with ease of analyzing the data.

***2.1 System Study***

The system study is regarding the study of the existing sytem which deals with a similar kind of problem. The existing systems usually deal with data in the for pre-processing and visualization with the use of notebooks with some code written in computer languages.

The idea is to eliminate the need of code just to do some basic exploration of the dataset. Here are the steps for study:

1. Gather resources
2. Study and explore the existing sytems
3. Re-analyze and explore alternatives
4. Analyze the feasibility of the solution
5. Scope for enhancements
   1. ***Challenges***

* Using Natural language processing to classify the higher level language into machine-understandable language
* Finding the key points in the datasets, to help the user visualize and draw inferences
* Organizing the data into a format to make it comprehensible
* Processing image input, to find the right text and match it with a movie in the database.
* Organizing all of the data in the GUI using QT application.
* Dealing with huge amounts of data, using the right type of graphs and plots
  1. ***Scope of project***
* **Highly Scalable for other datasets as well**
  + The solution provided here can be used for different kind of datasets to analyze the query of the user be it in text format or Image format.
  + The solution which processes the query and fetches the outcome from the dataset can not only be used by the developers to understand the dataset but can also be used by the end-users (if any) to understand the values in the database.
* **Predict the year given the values of emotion**
  + The values of emotion of each category (Angry, sad, neutral, disgust, surprise, fear, happy) can be used to show the release date of any movie.
* The prediction can also be used to know if the movie is relatable to the current trend of emotions

***3. Design***

***3.1 Problem Statement***

Given a dataset, the problem statement is to find a intuitive and innovative way to analyze the dataset and draw inferences from it. Also, the platform should provide a way to pre-process, modify and change the dataset which is already uploaded.

***3.2 Problem Statement***

***3.2.1 Economic Feasibility***

The platform or the entire system is highly feasible in terms of the economic perspective. At least for a system where it is used on a local system.

***3.2.2 Technical Feasibility***

The technicality involved in this project as been deeply analyzed. The technical details and all the different libraries and technologies used in this project has been already discussed above. The frontend demands QT to be installed and the backend requires python libraries mentioned in the technologies section.

1. ***Modules and Code Components***
2. **Data Pre-Processing**
   1. The dataset given, contains a lot of un-processed information. Using Python library **“Pandas”** we have converted the datasets into inferences.
   2. The inferences are then stored into another dataset including all the previous data.
3. **Image Input format**
   1. The query from the user is in the form of image. The image is then converted to four different formats like RGB and grayscale etc. Using this combination we detected the text in the image differentiating it with the background using edge-detection. **Open-cv** is used for the above.
   2. Now, using **Pyteserract** from the converted image, text is extracted and then an array of all required constraints and keyword parameters are created and sent for processing/visualization.
4. **Text Input format**
   1. The query from the user is in the form of text. The text is converted into vector assigning different values using **Tensorflow** (Tokenizer) simulating Natural Language Processing(**NLP**).
   2. The converted vectors are then matched with pre-defined keywords and constraints to extract the values as parameters for sending it to processing/visualization
5. **Data Visualization**
   1. The converted input which is generated using Image and Text processing is used for fetching the data from the datasets.
   2. In some cases, Output should be just a simple sentence. These cases have been taken into consideration.
      1. Ex
         1. Query: Who is the lead role in the movie 3 idiots.
         2. Answer: The lead actor is “Rancho”
   3. In some other cases, Output should be in the form of graphs. Consider the example “What is the variation of values of emotions over the years since 2007?”
   4. Complex queries are also handled, where multiple queries can also be asked by the user.
      1. **Ex.** What is the genre of the movie Shamitabh and who is the lead role in the movie 3 idiots and what is the plot of the movie black mail.

1. ***Software Requirement Analysis***
   1. ***Description***

* This system is designed to understand the dataset in-brief
* System can take input from user in the form of high-level language and then use it to fetch data from the dataset.
* Pre-processing is made easy with this flow model  
  1. ***Application Requirement***

|  |  |
| --- | --- |
| Front-End | QT |
| Back-End | ML Libraries |
| OS | Windows, Mac, Linux |

***6.1 Proposed Methodology***

Movie Analysis System

***6.2 Use-case Diagram***

Text Output

Inferences

Graph Output

Image Input

Text Input

* 1. ***Flow Chart***

Choose

Text Input

Image Input

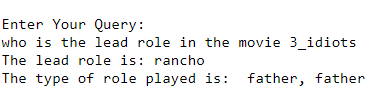
Insights, Graphs, Text Output

***7. Test-Cases***

**Test case – 1**

**Query :** Who is the lead role in the move 3 idiots

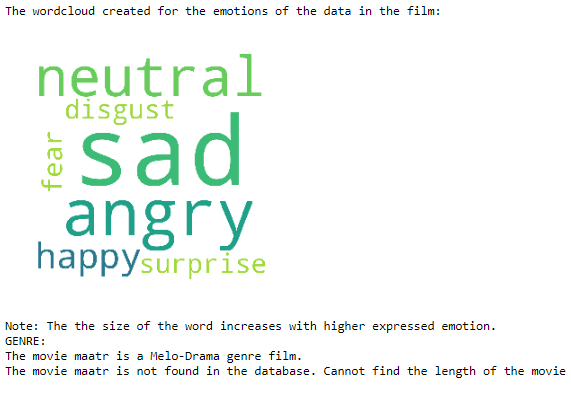
**Solution:**

****

**Test case – 2**

**Path of image:** <absolute\_path>

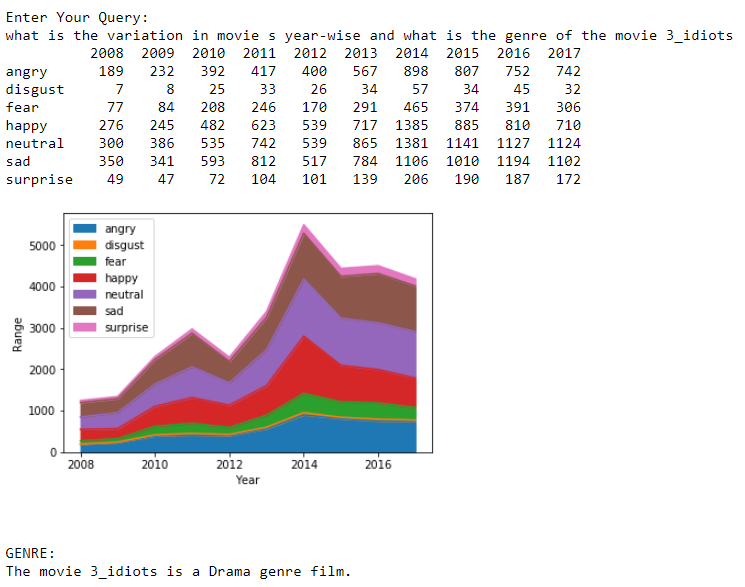
**Solution:**

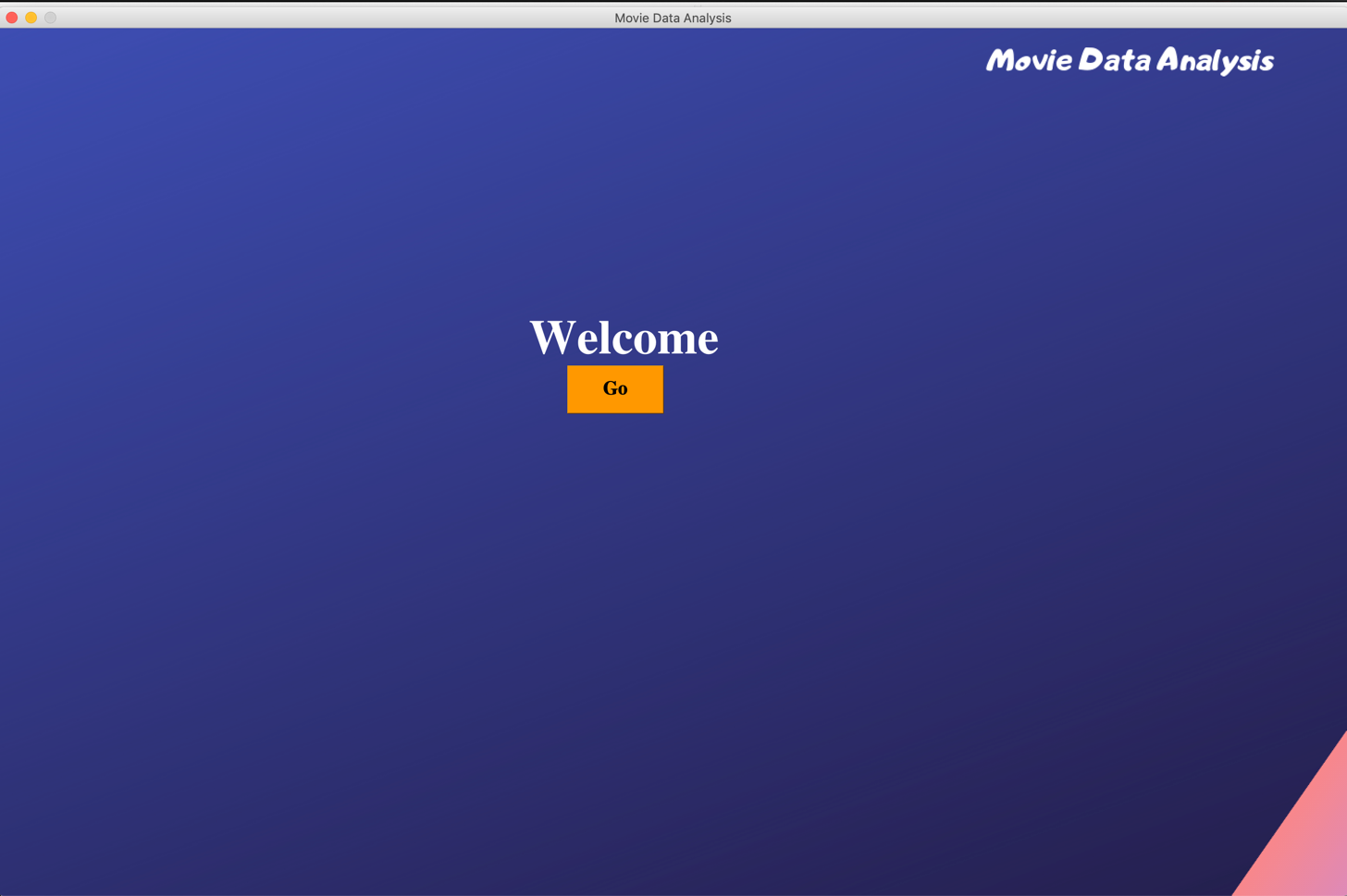


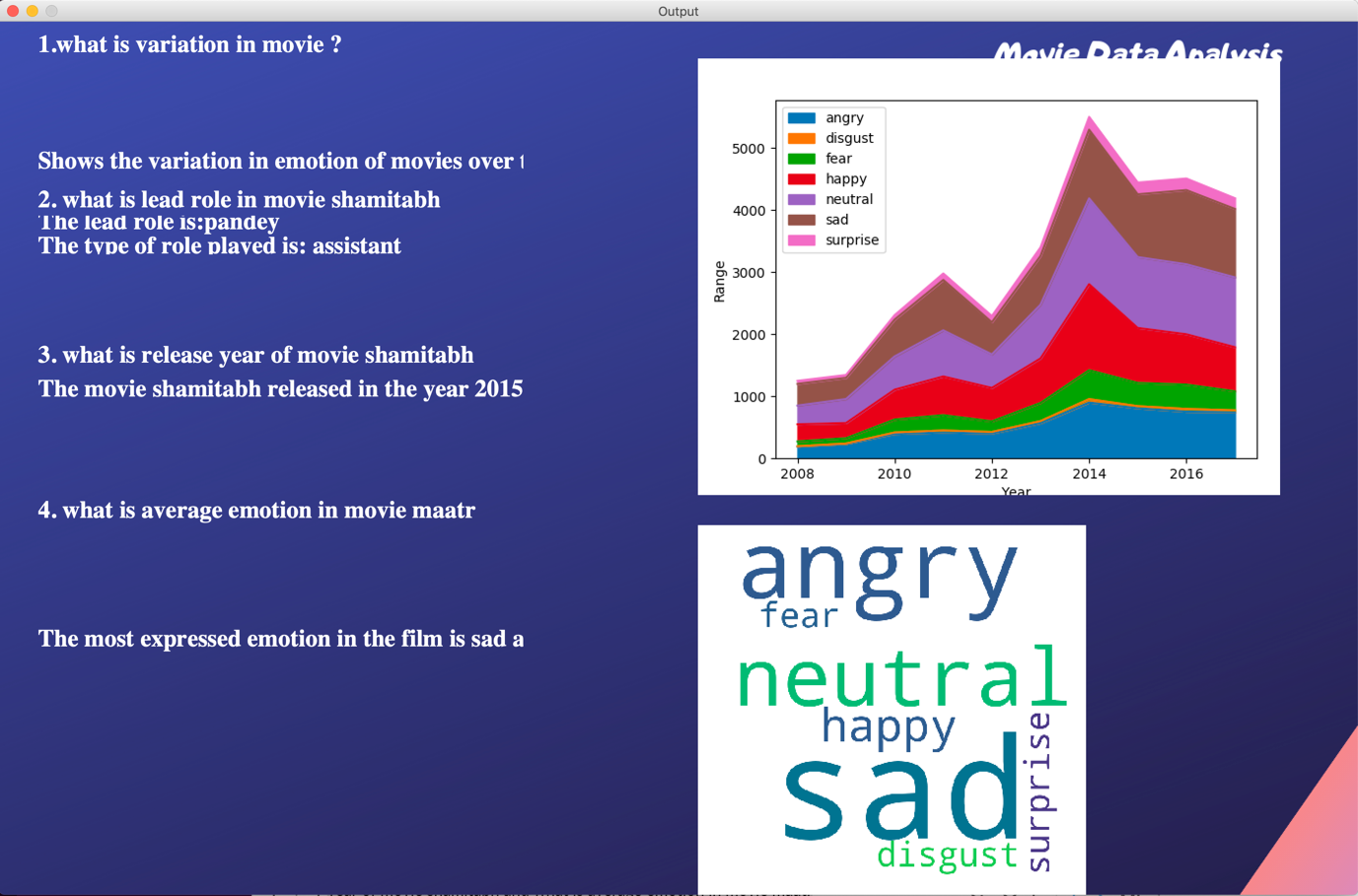
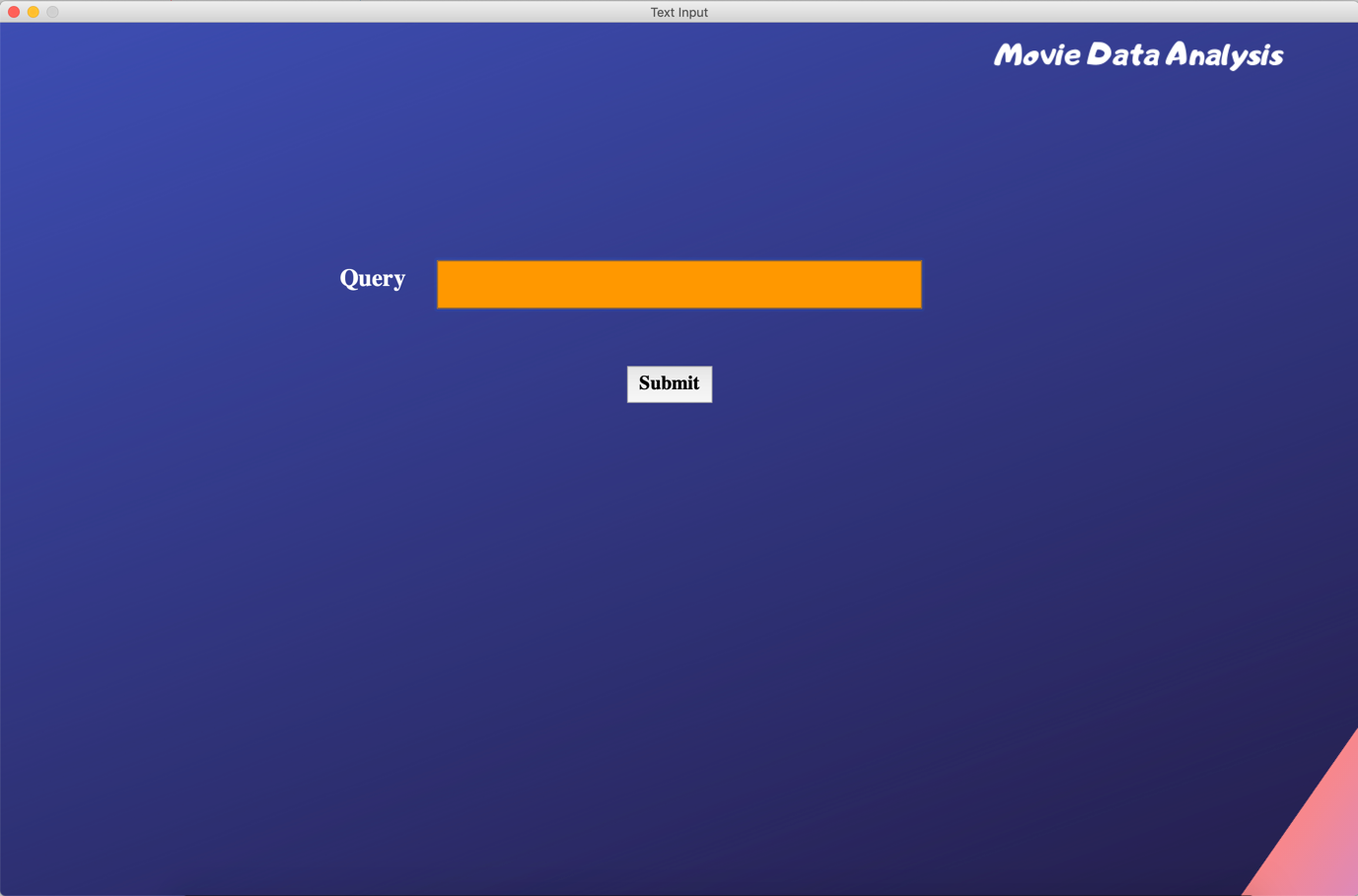
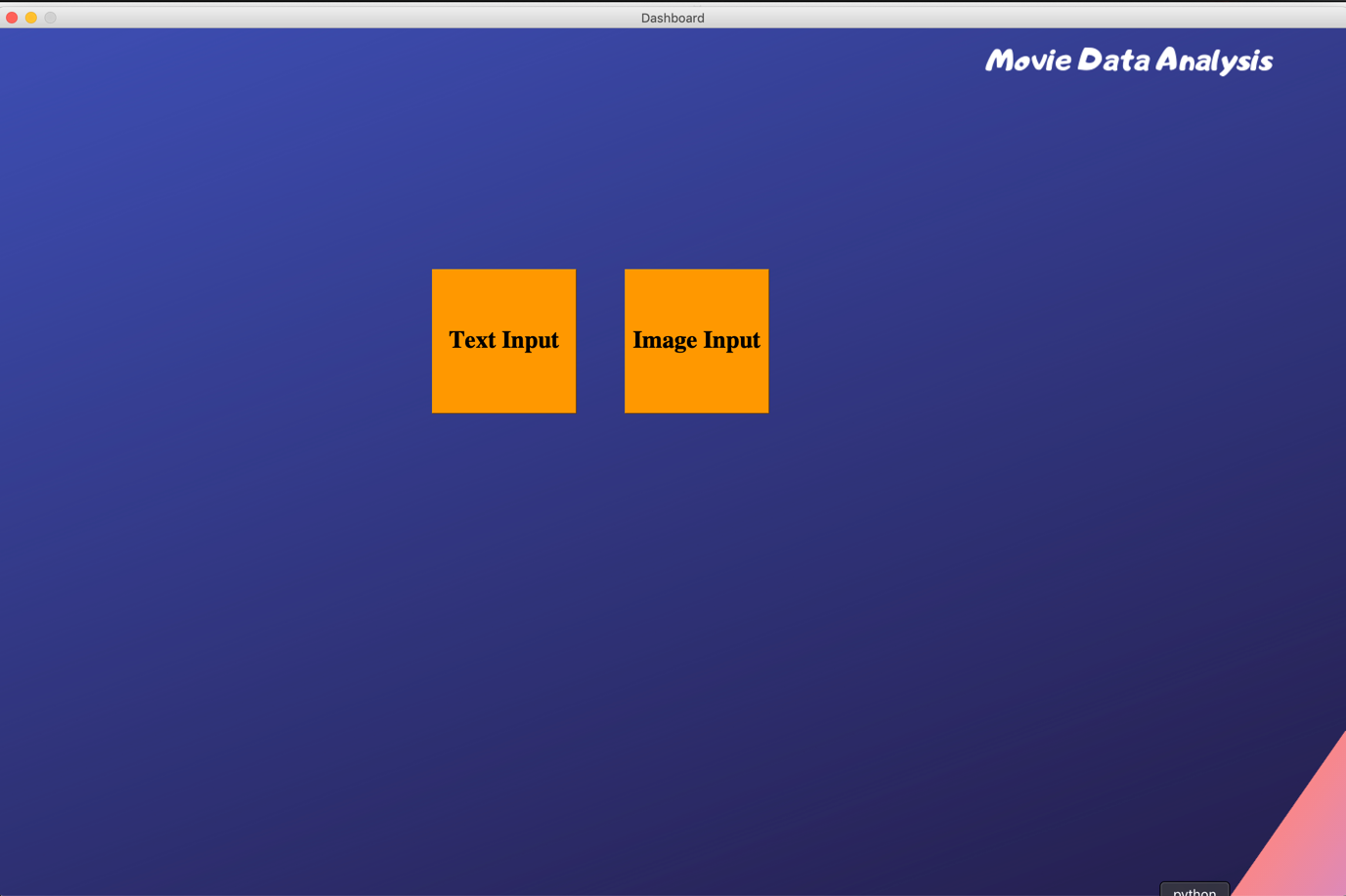
**Test case – 3**

**Query :** what is the variation in movie year-wise and what is the genre of the movie 3 idiots?

**Solution:**



1. ***UI Snapshots***



***9 References***

* [www.github.com/nikhiljsk/qtproject](http://www.github.com/nikhiljsk/qtproject)
* <https://www.britannica.com/technology/image-processing>
* <https://en.wikipedia.org/wiki/Data_science>
* <https://www.expertsystem.com/machine-learning-definition/>
* <https://www.programcreek.com/>