

**SOFTWARE REQUIREMENTS SEPCIFICATION**

**FOR**

**Movie Data Analysis**

**(Using Data Science with GUI)**

**CSE 308 – Computing Project**

**SUBMITTED TO: Baljeet Saini**

**Department of Computer Science**

**SUBMITTED BY:**

**Prasanth G. 11601526**

**Nikhil JSK 11601789**

**Section K1632**

**Table of Contents**

1. Abstract
2. Introduction
3. Technologies Used
4. Background Study
   1. QT
   2. Data Science
   3. Machine Learning
   4. Image Processing
5. Challenges
6. Code Components
7. Proposed Methodology
8. DFD
9. Use Case Diagram
10. Flow Chart
11. Scope of Project
12. Test Cases
13. Reference
14. ***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. We have also analyzed the emotions of the film in each year and then could predict the year given the rate of emotion. 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. ***Introduction***

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
  + Scikit-Learn

1. ***Background Study***

* QT
* Qt is a cross-platform application development framework for desktop, embedded and mobile. supported platforms include Linux, OS X, Windows, VxWorks, QNX, Android, iOS.
  + Qt is not a programming language on its own. It is a framework written in C++. A preprocessor, the MOC (Meta-Object Compiler), is used to extend the C++ language with features like signals and slots. Before the compilation step, the MOC parses the source files written in Qt-extended C++ and generates standard compliant C++ sources from them. Thus, the framework itself and applications/libraries using it can be compiled by any standard compliant C++ compiler like Clang, GCC, ICC, MinGW and MSVC.
* Data Science   
  + Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data
  + It encompasses all the ways in which information and knowledge are extracted from data. The term ‘Data Science’ is the study which deals with identification, extraction, and representation of meaningful information from raw data set to be used for business determinations.
* Machine Learning  
  + Machine learning is an application of artificial intelligence (AI) that provides systems the ability t Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.o automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.
* Image Processing  
  + Set of computational techniques for analyzing, [enhancing](https://www.merriam-webster.com/dictionary/enhancing), compressing, and reconstructing images.

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. ***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.
6. **Prediction using Machine Learning**
   1. Firstly, we created a graph from the emotions dataset which is used to show the variations in emotions over the years from 2007 to 2017
   2. Given the values of different categories in emotion (Angry, Sad, Happy, Disgust, Surprise, Neutral, Fear) by the user we can predict the year in which the movie released or if it is relatable to the audience in the present scenario.
   3. We have used “**SciKit-Learn**” to predict the value of year.

***7. Proposed Methodology***

Movie Analysis System

***Use-case Diagram***

Text Output

Inferences

Graph Output

Image Input

Text Input

***8. Flow Chart***

Choose

Text Input

Image Input

Insights, Graphs, Text Output

1. ***Scope of project***

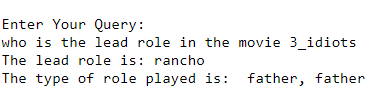
* **Highly Scalable for other datasets as well**
  + The solution provided here can be used for different kind of datasets to analyze the query of the user be it in text format or Image format.
  + The solution which processes the query and fetches the outcome from the dataset can not only be used by 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

***10.Test-Cases***

**Test case – 1**

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

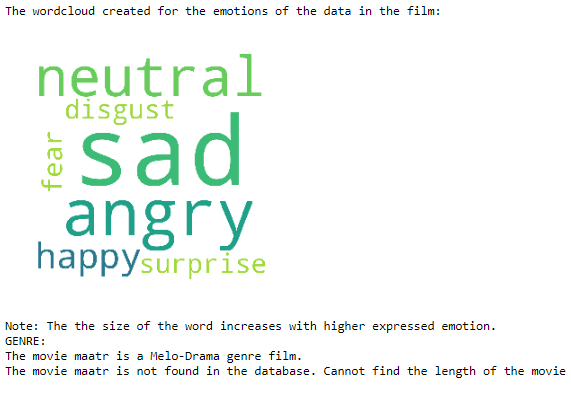
**Solution:**

****

**Test case – 2**

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

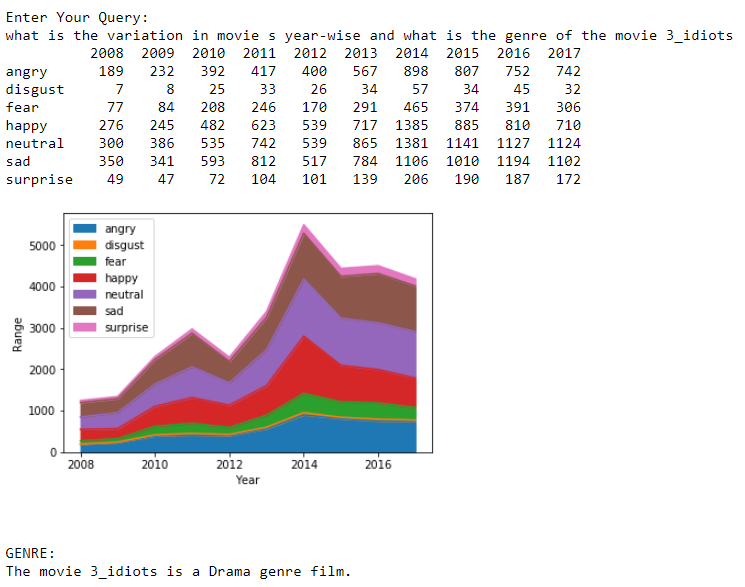
**Solution:**



**Test case – 3**

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

**Solution:**



***12.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/>