Project Iris:
Colour and Mood-Based Recommendations Models for Films

Prepared by:
Win K. Phyo
MS Applied Data Science Candidate ’22
Email: wkpwinphyo@gmail.com, win.phyo@mymona.uwi.edu

For:
COMP 6830 - Data Science Capstone Project II
Professor Ricardo Anderson
February 27, 2022

CONTENTS

EXECUTIVE SUMMARY 3
DOMAIN BACKGROUND 4
OBJECTIVES 5
DESCRIPTION OF DATA 6
DATA SCIENCE OBJECTIVES 7
KEY DELIVERABLES 8
REFERENCES 9

EXECUTIVE SUMMARY

Popular movie content streaming services offer limited options for how users can browse or tailor their recommendations. Traditionally, users have only been able to filter by genre, language, cast & crew, and maybe themes. **Project Iris**, a data-driven web application, aims to provide all fans of film - from casual cinema patrons to movie buffs - a carefully curated movie recommendations experience. In addition to the usual ways users have come to interact with movie content suggestions, **Iris** will focus on two fairly unexplored ways of categorizing films: by colour, and mood.

Users will register on our website, and self supply watch histories that will help drive our recommendations engine. By leveraging *Internet Movie Database (IMDb)* data, image data from the films themselves, and publicly available user and critical reviews and opinions published on *Twitter* and review aggregation sites such as *Metacritic* or *Rotten Tomatoes*, the **project** aims to build a deep repository of movie data and a machine learning model that is capable of cleverly identifying titles that are similar in visual and emotive feeling.

DOMAIN BACKGROUND

The modern obsession with streaming digital media has led to the interest to develop robust content recommendations engines that aim to effectively utilize data science techniques to identify similar media that a particular user might be interested in based on other variables such as personal watch history or demographic information. In addition, most mainstream contemporary movie/television streaming services offer limited options for how users can browse their libraries and suggestions (typical filters include genre, cast & crew, language, and country). Colour, and mood are two ways in which we think about our experiences with films, that have been largely unexplored in terms of content recommendation models.

Colour can not only evoke certain feelings, but can also be characteristic of a certain genre, or director, or even period in time. Colour data possesses large untapped potential that could be used to derive new insights about films. There are even particular viewers that seek out a specific colour palette when choosing what movie to watch. And there are certainly viewers that seek out a specific mood or emotion when choosing what to watch.

The aim of **Iris** is to develop a platform where users can explore films that cater to not only particular tastes in genre or themes, but in colour or mood. Though the app will be centred around these two main categories, the development of models relating to these variables will lead to the improvement of existing content recommendation models as well.

Some quick figures:

Global revenue from video streaming services was estimated to be $70.845B USD in 2021 (Statista). It is expected to grow from $82.431B in 2022 to $115.92B in 2026.

Netflix ($NFLX) has a market capitalization of $173.50B as of February 27, 2022.

Netflix currently has around 220B paid subscribers (Statista).

OBJECTIVES

The primary goal of this project is to provide a web application for all types of people interested in films to find similar content based on two parameters that mainstream video streaming platforms do not currently support: colour, and mood.

By analysing image data, we can visualise the use of colour in movies, and how hues change throughout the runtime of a film. This opens up a variety of potential applications for this data: perhaps we find that a particular director or cinematographer or even cast member is commonly associated with a specific colour palette. Then maybe, this colour palette is associated with a particular type of mood or genre. Maybe we find that colour tends to change in a specific way in certain types of films. There are a number of new insights that can be gained from the products of these analyses.

The analysis of mood, and the progression of moods in a film, similarly to colour data, will be used to offer users with suggestions on content to watch if they are looking for a certain type of emotive feeling.

The combination of these time-based data of colour, and mood, with the existing movie categorical data (i.e., genre, cast, runtime, etc.), and the massive volumes of rich, human-generated text data published on public internet forums, social media, and other websites creates brand new opportunities for new features and improvements to movie recommendation engines. This product will go beyond the kind of selection choices that typical streaming services offer, and the ultimate goal of the product is to create tastefully crafted recommendations that provide the watcher with a movie experience that appeals genuinely to their visual and emotive interests.

DESCRIPTION OF DATA

The public has for a long time had access to a public database of movie/tv information through the Internet Movie Database (IMDb), an online database that includes information related to plot, genre, cast, runtime, reviews, etc. of over eight million titles across multiple languages (imdb.com). This serves as a fantastic starting point for compiling our structured dataset. IMDb is owned by IMDB.com, Inc., a subsidiary of Amazon. IMDb, however, does not have a public API. Instead, we will rely on The Open Movie Database (OMDb) API, a RESTful web service, to supply this data (omdbapi.com).

The Colors of Motion (TCOM), a project launched by designer Charlie Clark finds the average hue of each frame in a movie, and generates a visual timeline of that title using colour (thecolorsofmotion.com). HappyCoding.io, an open source project provides similar datasets through their Movie Colors project (happycoding.io/gallery/movie-colors). There are also a number of public projects published by Reddit users on the r/dataisbeautiful subreddit that can also be utilized to fill out our image dataset (reddit.com/r/dataisbeautiful). For those titles for which we cannot retrieve image data on the open web, we can build our own processes to extract the relevant data.

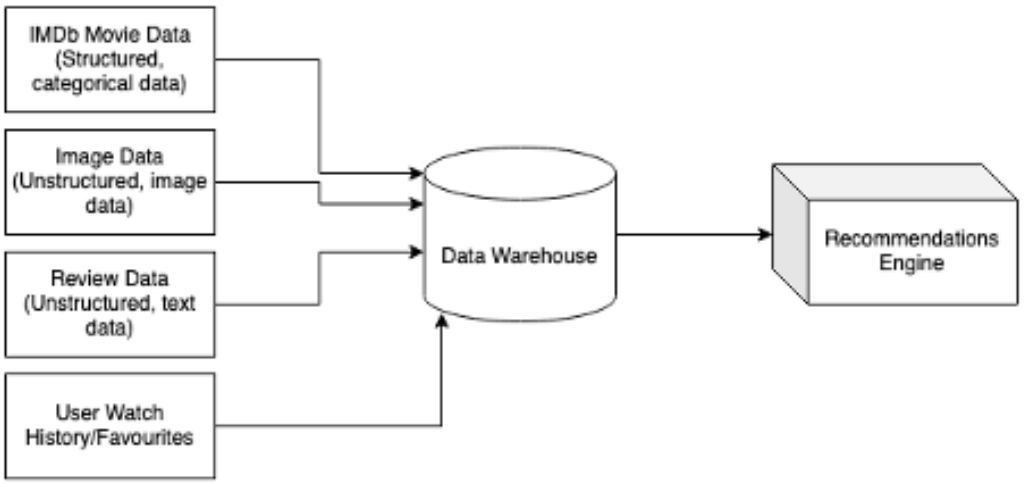
For sentiment analysis of movie review data, we will use public user and critic reviews from Metacritic, Rotten Tomatoes, IMDb, and Twitter. Twitter's API offers a free Academic Research access level that allows for the retrieval of up to ten million tweets per month. For the other sites, a solution will need to be developed to scrape the text data. There are potentially hundreds of thousands of user-submitted reviews available online, and this is likely the data source that would be updated most frequently.

Individual watch histories and favourite movies will be self-supplied by users of the app.

DATA SCIENCE OBJECTIVES

The primary feature of this product is the robust recommendation engine that will offer the user the functionality to browse suggestions based on colour, and mood.

The proposed data architecture is illustrated below:



To build out the data warehouse needed to drive this app, we rely on:

- IMDb data – title, genre, runtime, cast, language, year, etc. Supplied by OMDb API as JSON objects,
- Image data – average colour per frame, time data. Scraped from various publicly available sources on the internet, including TCOM,
- Review data – self-reported score (if supplied), sentiment data. Scraped from various movie review aggregation sites and Twitter,
- Watch history/favourites data – self-reported by registered users of our app. This can also be expanded to include self-reported biographical data.

The data provided by OMDb API will serve as the base dataset for our models. Image data will need to be analysed against other data describing the film to derive relationships between colours and other variables. The review data will be instrumental in the mood-based analysis of our movies, as the input here is actual text written by humans who have some opinion on that film. The overall goal of this stage of the project is to reliably identify these relationships and similarities, and group titles based on these common themes. The expectation is that the unstructured data can provide insights into why specific colours are utilized, and why certain films evoke certain feelings. These models will drive our recommendations engine, and will be ultimately deployed on our web app to generate personalised suggestions.

KEY DELIVERABLES

Deliverable	Description	Proposed Date
Build and deploy working pipeline to data warehouse	Build and maintain a pipeline to automatically create and update the necessary datasets in the data warehouse	June 1, 2022
Build colour model	Analyse image data and combine with other data to build model to find titles similar in colour information	July 31, 2022
Build mood model	Analyse data to build model to find titles similar in feeling or emotion	July 31, 2022
Deploy models as recommendations app	Release web app that offers users personalised movie suggestions based on colour, mood, etc.	August 10, 2022

REFERENCES

The Colors of Motion - thecolorsofmotion.com

HappyCoding.io – happycoding.io

Internet Movie Database (IMDb) – imdb.com

Metacritic – metacritic.com

The Open Movie Database (OMDb) API – omdbapi.com

Rotten Tomatoes – rottentomatoes.com

r/dataisbeautiful - reddit.com/r/dataisbeautiful

Statista - statistica.com