

Department of Computer Science

Individual Project – CS3IP16

# Predictive Analytics for Film User Ratings

|  |  |
| --- | --- |
| **Student:** | Kane Small |
| **Student Number:** | 23013043 |
| **Supervisor:** | Jonathan Boyle |
| **Submission Date:** | 20th April 2018 |
| **Word Count:** |  |

### Abstract

### Here you create a brief summary of the content of your report. Making clear the subject matter, the problem, solution and headline results. An abstract should not normally be longer than about 250 words, and in most cases should be finalised last when the main content of your report is complete.

# Acknowledgements

First and foremost, I would like to express my gratitude toward Jonathan Boyle for his continued support throughout my project’s lifetime and helping to guide the project in the right direction.

Secondly, I would also like to thank my Mother for her wisdom, support and strength throughout this entire year.

Lastly, I would like to thank my friends for assisting me with carrying out testing and providing valuable feedback toward improving my project along the way.

Table of Contents

[Predictive Analytics for Film User Ratings 0](#_Toc511664094)

[Acknowledgements i](#_Toc511664095)

[Glossary of Terms and Abbreviations iii](#_Toc511664096)

[1 Introduction 1](#_Toc511664097)

[2 Problem Articulation/Technical Specification 2](#_Toc511664098)

[3 Literature Review 3](#_Toc511664099)

[4 Solution Approach 4](#_Toc511664100)

[5 Design and Implementation 5](#_Toc511664101)

[6 Testing and Validation 8](#_Toc511664102)

[7 Discussion: Contribution and Reflection 9](#_Toc511664103)

[8 Social, Legal, Health, Ethical Issues 10](#_Toc511664104)

[9 Conclusion 11](#_Toc511664105)

[10 Future Improvements 12](#_Toc511664106)

[10.1 Improving prediction model accuracy 12](#_Toc511664107)

[10.2 TV show support 12](#_Toc511664108)

[10.3 More trend graphs 12](#_Toc511664109)

[10.4 External ratings used as model features (rotten tomatoes, metacritic, etc) 12](#_Toc511664110)

[10.5 Updater function, running on a schedule 12](#_Toc511664111)

[10.6 Making the web application multi-user accessible 12](#_Toc511664112)

[10.7 Securing the API endpoints 12](#_Toc511664113)

[11 References 13](#_Toc511664114)

[12 Appendices 14](#_Toc511664115)

[12.1 Appendix 1: Project Initiation Document (PID) 14](#_Toc511664116)

[12.2 PID Sign-Off 14](#_Toc511664117)

[13 SECTION 1 – General Information 15](#_Toc511664118)

[13.1 Project Identification 15](#_Toc511664119)

[13.2 Student Identification 15](#_Toc511664120)

[13.3 Supervisor Identification 15](#_Toc511664121)

[13.4 Company Partner (only complete if there is a company involved) 15](#_Toc511664122)

[14 SECTION 2 – Project Description 16](#_Toc511664123)

[15 SECTION 3 – Project Plan 19](#_Toc511664124)

[15.2 Appendix 2: Logbook 0](#_Toc511664125)

# Glossary of Terms and Abbreviations

API – Application Programming Interface

Clustering – A machine learning problem whereby the outcome is to create groups of values based on their similarities.

Classification – A machine learning problem whereby the outcome is to predict a discrete output, i.e. a class or category.

Data mining – Is the process of finding patterns within data.

DBMS – Database Management System

FTP – File Transfer Protocol

JSON – JavaScript Object Notation

Machine learning – Is the process of building models using data mining techniques, in order to predict future outcomes.

MAE – Mean Absolute Error

Model – In machine learning, a model is the output generated after training a machine learning algorithm with a training dataset. It takes in a predefined set of inputs and produces an output based on these variables.

One-hot encoding – Converts a set of categorical values into separate features, that are all represented by a binary value (0, 1).

ORM – Object-relational Mapping

PCA – Principal Component Analysis

PID – Project Initiation Document

RDP – Remote Desktop Protocol

Regression – A machine learning problem whereby the outcome is to predict a continuous value.

SSH – Secure Shell

Supervised learning – A machine learning category that relies on humans to act as the teacher whereby, we feed in a training dataset that includes the features and target values and from this data an algorithm can learn the relationships between the features and the target values in order to predict target values for previously unseen data.

UoR – University of Reading

UPS – Uninterruptable Power Supply

VFX – Visual Effects

VPN – Virtual Private Network

# Introduction

Machine learning is still in a very primitive state at this point in time but is being utilised a lot more frequently across different industries, from entertainment, to finance, to transportation services and many more (Sas.com, 2018). The process for machine learning models is to formulate patterns and understand the inter-linking features within the dataset being analysed, that can then be used against previously unseen datasets in order to predict values to a certain degree of accuracy. But ultimately, the end goal for using this new technology is to make better decisions without the need for human intervention.

This is where this project comes into play. The application uses machine learning regression algorithm to train a model that can successfully predict rating values for films that have yet to be released, based on the most important key attributes (the features) that influence viewer’s decisions. The main purpose behind building this application is to essentially make it easier for anyone to decide whether to see a film, based solely on a single value. This singular rating percentage is the result of, finding the inter-linking patterns and relationships of the most important film attributes and is ultimately as unbiased as possible by using only factual information.

The project commences by identifying the problem that is trying to be solved, listing the stakeholders that will form the main target audience of the application and more importantly, the main objectives of the project. In order to accomplish these objectives, it is necessary for the project to be split into five separate parts, data acquisition, data cleansing/pre-processing, regression algorithm implementation, API routing and front-end design/user experience. The first phase focuses on acquiring data from a viable source that meets the requirements (i.e. relevant metadata), the next phase transforms the acquired data into a correct format for the input to the algorithm and stores it into the database, the thirds phase implements the regression algorithm and train a model; the fourth phase revolves around building an API that will interface with the database and display data to the front-end and the last phase will concentrate on the user-experience and ease-of-use of the application.

Furthermore, the motivation and reasoning behind undertaking this project is brought to light, followed by the constraints that illustrate potential challenges for the project.

The application must then look next at any existing applications that carry out either the same or similar tasks and see how it can be improved upon, from both a machine learning accuracy and a user-experience perspective. This is accomplished by reviewing existing literature and from the findings, justify an informed solution approach on how best to meet the objectives outlined. The solution approach itself will also aim to cover any tools and technologies utilised throughout the development process in brief.

The design and implementation of the application will be explored in detail, illustrating exactly how the technologies and any libraries aided in the development of the product. The implementation will build all the way to the result of the model, which will be examined in detail as part of the discussion. This will also address the success of the solution, as well as any limitations that have been identified along the way. A reflective piece will lead on from here, bringing to light what the challenges were, what has been carried out well and how the project would be approached differently if attempted a second time.

Any potential social, legal, ethical and health and safety risks of the project are reviewed in short at this point, which will support the continued development and improvement of the application in the future.

Lastly, the report comes to an with a conclusion that quickly restates the main project objectives and discusses whether the solution that has been implemented has in fact met them. Nevertheless, this outcome leads into a list of future improvements that could potentially elevate both the project’s effectiveness and use.

# Problem Articulation and Technical Specification

## Problem statement

The number of variables that determine whether a film will be a high user-rated success has increased dramatically over the years, in-part due to the advancements made within the film industry allowing production companies to offer a wider variety of films to the viewers, specifically in the Visual Effects (VFX) category. So, with this comes the pressure of leaving it to the viewer to determine whether a film is going to be enjoyable based on these attributes. But, a viewer can only use what they already know about an upcoming film such as the genre of the film, or the actors cast, or the trailer and so on and these attributes will simply *influence* their decisions; but will not tell them whether or not this film will result in a decent or meagre film-viewing experience.

This project aims to alleviate that pressure on viewers by taking all of the factual data already present and using it to make accurate predictions on the overall rating of unreleased films, but by taking into account each of the most important film attributes and even finding relationships between them.

## Stakeholders

A total of three stakeholders were identified for this particular project, revealing how they would be affected by or provide assistance in achieving the outcome of the project.

### Developer – Kane Small

The developer, Kane Small, is responsible for advancing the project through all phases of the development lifecycle. This involves making sure that the problem statement outlined in section *3.1* is resolved and this will be accomplished by fulfilling the objectives outlined in section *3.3*. Due to time constraints and other important engagements throughout the year, the developer will also be responsible for maintaining an effective pace through time management and organisation, to ensure that project goals are met on time.

### Project supervisor – Jonathan Boyle

The project supervisor, Jonathan Boyle, is the person who will provide continued support and assistance toward guiding the project in the right direction at all times. This will be accomplished by scheduling weekly and/or bi-weekly meetings whereby a detailed discussion will comprise of the work that has been carried out and any problems encountered since the last meeting ending with a summary of what will be tackled by the next meeting.

### User – Film Viewers

The user of the finished product will be, any film viewer that is uncertain on whether or not to see an upcoming film. They will expect to use a system that is easy-to-view and understand, where the content has been organised in a concise and logical manner but most importantly that the prediction result being displayed is as accurate as possible; as they will be relying on this factor to make their decisions for them.

## Technical specification

The application to be developed will reside on a server that will deliver the content to the internet. There will be two pages, the *main* application page that will display all of the film metadata including the prediction result and a *trends* page that will provide users with updated trend graphs for the data being held.

Utilising the previously created PID, the project must satisfy the objectives outlined. However, the objectives that are listed below have been altered in order to conform to realistic time constraints and developer experience:

**Data acquisition**

* Implement a viable API data source
* Utilise the YouTube API for film trailer acquisition
* Design, build and configure the database to store the film data
* Implement a viable Python framework
* Build the scripts that will perform the data acquisition
* Setup and configure a server to host the scripts
* Create a questionnaire to aid with feature selection

**Data cleansing/pre-processing**

* Use imputation to format NULL data values
* Use one-hot encoding to format categorial features
* Use Principal Component Analysis (PCA) to reduce dimensions

**Machine learning implementation**

* Select an appropriate machine learning algorithm
* Fit the formatted data to the algorithm
* Train the algorithm to produce a model that can output a set of predictions
* Test the accuracy of the model using different regression metrics

**API routing**

* Create API routes to serve data to the front-end
* Implement basic API route authentication

**Front-end design and user experience**

* Create a functional front-end interface for the user to interact with
* Implement the typeahead and bloodhound autocomplete and suggestion engine
* Allow the user to query for films
* Implement charting/graphing capabilities to display data trends

### Project motivation

The main motivation behind carrying out this project is, the developer’s passion for film. Having had the idea to implement some form of film aggregation and search platform in the past, but not knowing exactly how to make the application stand-out, combining a field of interest such as machine learning has enabled the project to reach a state that is worthy of development.

The application itself, also has a lot of possibility for future implementations such as, including TV show support, or allowing custom search parameters to be user-inputted to test different film recipes – which would be extremely useful for production companies.

### Project constraints

The biggest constraints to consider are scheduling and scope. Ensuring that, within the given time it will be possible to research, learn and implement a successful machine learning algorithm that will be able to output a set of predictions. But more importantly, guaranteeing that the project is not over-scoped and that the objectives outlined are focused on first and foremost, where additional features and functionality can be implemented if there is additional time at the end of the development lifecycle.

# Literature Review

#### The literature review is an essential component of your project report. You should discuss the existing literature that is relevant to your project with full and proper referencing. You should aim to refer to a range of material including academic papers, text books, articles and existing product descriptions. It should be clear to the reader why the literature you identify is relevant and how you have incorporated the learnings from your review into your project. For example, you may have made a number of project Page 4 of 5 decisions based on your review of the literature and these decisions should be described. The literature review should also lead to the creation of a number of possible solutions to your problem articulation and technical specification.

# Solution Approach

#### Here you should identify and evaluate the solution options against your problem statement / technical specification and make a reasoned choice of your chosen solution approach. Why did you do what you did? Conclude with a succinct definition of your solution approach and criteria by which the solution would be accepted as adequately verified. It is very likely that you will add further reference material in this section.

# Design and Implementation

#### Here you describe the detailed design of your solution and the details of the actual implementation. It may be appropriate to discuss aspects of design or implementation that were particularly problematic and/or novel. This section may well be one of the largest in your report and the exact contents will be unique to your project and so there are no general guidelines. Use of several sub-sections here is appropriate.

## Data acquisition

### Libraries utilised

**Acquisition:**

* requests
  + Handling POST/GET HTTP requests
* pymysql
  + Connecting Python to the MySQL database
* progressbar
  + Visual aid for time-consuming operations
* time
  + Sleeping the code to wait out the rate limiting

**Error handling:**

* csv
  + Printing errors to a csv
  + Prevents any operations from halting
* traceback
  + Allows for the printing of stack traces

### Functions created

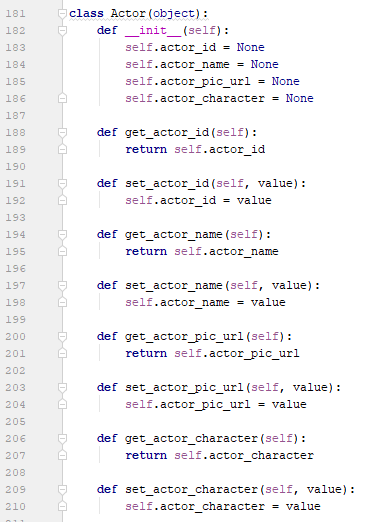
* getLatestFilmID()
  + Checking latest film ID via API
* insertData(sqlQuery, params)
  + Inserting data into the database
* getTrailerData(film\_id)
  + Acquiring trailer data
* matchActors(film\_id)
  + Matches actors to film\_id (separate API routes)

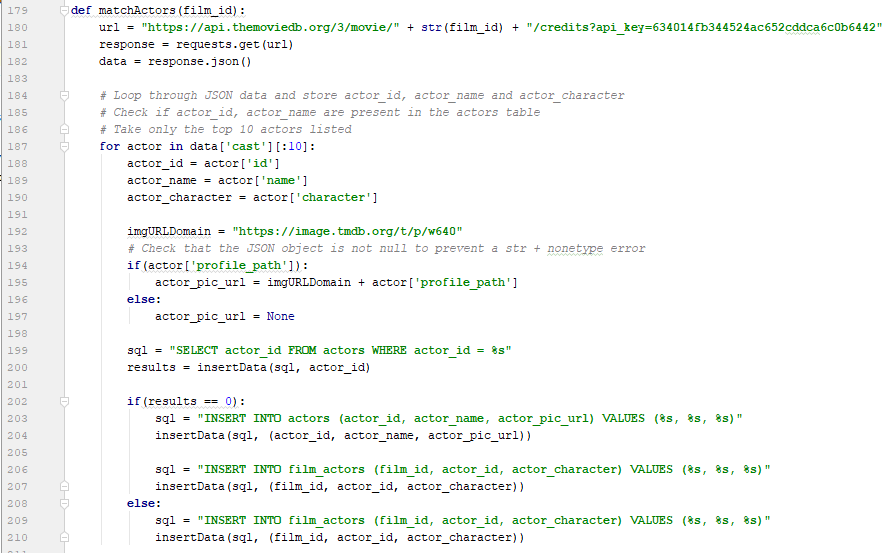
## Data cleansing/pre-processing

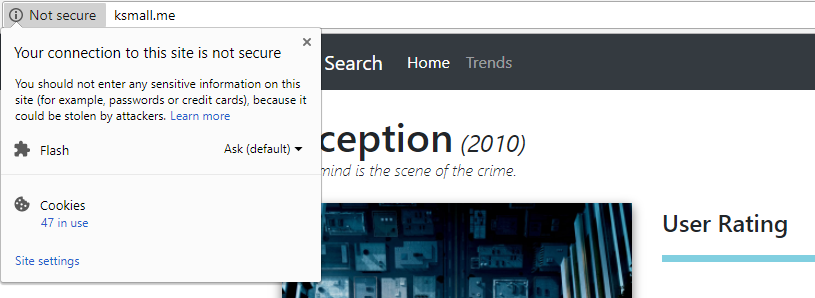
## Regression algorithm and predictions

## API routing

## Front-end design and user experience





* 1 class per database table
* 5 separate SQL queries
* Map results to an object
* Return jsonpickled object



# Testing and Validation

#### Here you explain your approach to testing and show your results. Testing should be a directed process and so there should be some discussion of why you have done the tests you have and why they are appropriate to the validation of your problem solution. You should also consider the limits to your presented verification and validation.

## Unit testing

**Back-end code**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test ID | Description | Expected Outcome | Actual Outcome | Actions Taken |
| 1 |  |  |  |  |
| 2 |  |  |  |  |
| 3 |  |  |  |  |
| 4 |  |  |  |  |

Table 1 - Back-end code unit test results

**Front-end code**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test ID | Description | Expected Outcome | Actual Outcome | Actions Taken |
| 1 |  |  |  |  |
| 2 |  |  |  |  |
| 3 |  |  |  |  |
| 4 |  |  |  |  |

Table 2 - Front-end code unit test results

**Compatibility**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test ID | Device | Operating System | Web Browser | Pass/Fail |
| 1 | Desktop | Windows 10 | Google Chrome | Pass |
| 2 | Desktop | Windows 10 | Mozilla Firefox Quantum | Pass |
| 3 | Desktop | Windows 10 | Microsoft Edge | Pass |
| 4 | Desktop | Windows 10 | Microsoft IE 11 | Fail |
| 5 | Laptop | macOS High Sierra | Safari | Pass |
| 6 | Laptop | macOS High Sierra | Google Chrome | Pass |
| 7 | Laptop | macOS High Sierra | Mozilla Firefox | Pass |
| 8 | Mobile | Android 8.0 “Oreo” | Google Chrome | Pass |
| 9 | Mobile | Android 8.0 “Oreo” | Mozilla Firefox Focus | Pass |
| 10 | Mobile | iOS 11 | Safari | Pass |
| 11 | Mobile | iOS 11 | Google Chrome | Pass |

Table 3 - Application compatibility test results

## Testing limitations

# Discussion: Contribution and Reflection

## Discussion

#### The discussion section follows on naturally from the results presented in the previous section. The discussion is usually substantial and is where you discuss your test results in detail, and their implications, and also potentially make links to relevant previous work. Analyse the success of your solution. Discuss limitations of this project, which will help when you introduce your Future Improvements in the Conclusion section. Reflect also your learning experience, e.g. what you will do differently if you would do this project again.

## Reflection

As a final year student completing his/her degree, you are required to undertake a final year project that must be completed throughout the last year and alongside other pieces of coursework. I will be looking at how I personally felt at each point during the project’s lifetime and ultimately, what I learned from this experience and how I could improve if given a second chance.

My initial feelings for the project were ambivalent, mainly due to the sheer size of the undertaking and because we had to continually work on the project throughout the year alongside multiple pieces of coursework being set by each module lecturer. However, once I had settled on a project idea and started working on it, because I was working on something that I find of interest to me I think this helped in viewing the project as an enjoyable set of tasks as opposed to mandatory *work*.

One of the worst parts of the project for me was before the year had even started. A few months before beginning our final year, we were tasked with either selecting a pre-defined project from a list that was provided or coming up with our own idea and submitting a PID for it. For my placement I worked at a cyber security company and so for the best part of that year I had been thinking of cyber security related projects that could have been interesting to complete. However, I eventually decided that cyber security was no longer a field I wished to pursue, as I wasn’t enjoying it as much as I was in the beginning. Therefore, completing a project related to that field no longer made sense. So, this left me with a few weeks of having to think of an entirely new project.

The stress and anxiety that the final year project caused me was severe and at times, very overwhelming. Having to juggle separate pieces of coursework for different modules I found challenging enough in the first two years of university, but having this additional project weighing down on you and also knowing that it’s worth a third of the entire year in terms of marks was again, at times too much to handle.

That being said, I have learned with the help of my incredible Mother, certain coping techniques and strategies to alleviate some of the pressures of work similar to this. So, there were definitely good experiences that I had whilst completing my project. I thoroughly enjoyed improving my programming ability with every piece of code I had to write. I decided to build my application in Python, because I believe that I have the most experience with this language over any other that I have used in the past, yet after having completed this project I feel as if I’ve learned so much more about the language then I could have imagined.

Learning how to implement machine learning has been a very interesting journey. I didn’t quite anticipate the level of difficulty that using these tools would provide but learning how to implement them as well as *why* to implement a specific algorithm for example has been really enjoyable. I have only just touched the surface when it comes to machine learning but, with so many industries beginning to adopt this technology it could be a very interesting field to break into after university.

In conclusion, from my experience I now know that it is vital to carry out research first and at the earliest point in development as possible. Before you can tackle any task, you need to have all of the tools, including the knowledge of how to implement *with* those tools.

Looking at my personal experiences and my unique style of working I would say that, industry experts spend months and months, sometimes even years, attempting to perfect machine learning models. I can’t expect to have achieved outstanding results on my first attempt and that shouldn’t hinder me from trying to get better results, trying different models and comparing accuracy scores. Just because this is the end of my project from an academic-perspective doesn’t mean it *has* to be the end of my journey of continuing to expand my knowledge in this vast field.

I think, if I were to undertake this or a similar project again I would try to focus on asking more questions. Making sure that I came prepared to each supervisor meeting with detailed questions, where we could then have discussions about. I would also ensure to utilise strategies to help me, such as recording feedback my supervisor gives me so that I don’t forget what has been said by the time I get home. I need to focus on my worst qualities first next time, such as time management and organisation. If I create a plan of action of *how* I am going to tackle these obstructions, it will be easier when it comes to starting.

# Social, Legal, Health and Safety and Ethical Issues

## Social, legal and ethical issues

When constructing the PID initially no social, legal or ethical issues were identified. However, upon reflection there are a couple of social and ethical issues that could in fact relate to this project.

Firstly, one of the social issues concerning the usage of the application itself relates to users who speak another native language other than English and would therefore not be able to understand the content that is being displayed to them. A simple solution to this problem would be to implement some form of translation feature, possibly with the help of an existing library that supports hundreds of languages and give the user the option to change the language manually. However, a more streamlined approach would be to automatically detect the user’s location via their browser session and to convert the language for them. But, having the manual option present would still be required for users that are visiting the application whilst using a VPN.

Secondly, an additional social issue which again, concerns the usage of the application itself, would be the fact that the application does not cater toward users with disabilities such as visual impairments. There are a variety of solutions of for this type of issue including but not limited to, ensuring that all *<img>* tags are making proper use of the ‘alt’ tag that describes what an image (often used for screen readers), ensuring that the web application is fully-operable via *­just* the keyboard, providing users with the control over text sizing for all elements and having an option for users to be able to alter the contrast levels of the application or include specific colour-blindness modes.

Finally, an ethical implication of this project could be that if the prediction accuracy of the ratings was high enough some could argue that there would no longer be a need for film critics, thus putting a lot of people out of work. A second ethical issue, that follows on closely from the last relates to the emotional side of films. A lot of people believe that computers are not capable of replicating human emotion and therefore, they will never be as accurate as a human critic. There are often films that do draw heavily on a viewer’s emotion, regardless of the known variables such as the genres, budget, trailer view count, etc and therefore it is hard to envisage a scenario in which a computer would consider these parameters when formulating predictions. Such film’s that are considered *cult-classics* would also be extremely difficult to predict accurately for, since a lot of the time these films are liked by viewers regardless of critic ratings/reviews, or factual metadata such as the actors in the film.

## Health and safety risks

There were a few risks that were identified during the creation of the PID relating to the work area and these were eye strain, hardware failure and repetitive strain injury. Eye strain was reduced by taking regular breaks and by utilising an application call *f.lux* (f.lux, 2018), which reduces blue light exposure. To reduce the risk of hardware failure, all documents and code were backed up to a private GitHub repository as well as Google Drive. Lastly, the risk of repetitive strain injury was reduced by taking regular breaks, using an ergonomic chair with lumbar support and height adjustments and by simply adjusting the workstation to appropriately conform to the user’s requirements.

# Conclusion

#### Briefly restate the project objectives and then make straightforward conclusions about your project work and results. What have been the key outcomes? You can suggest future work that logically stems from your work and refers to the limitations of your own work in the Discussion section.

# Future Improvements

A lot of time and thought has gone into how this project could be improved if more time had been allocated, or simply as the project continues to grow past a university project.

## Improving prediction model accuracy

One of the most important considerations is improving the accuracy of the prediction model. This is the core feature of the application and being able to specify to your users that this application is capable of predicting user ratings for unreleased films, to certain high-level degree of accuracy, would encourage a lot more people to use the platform.

In machine learning it is common practice to implement multiple algorithms to see which is best suited toward your specific dataset. So, with more time it would be possible to train different models and compare the accuracy scores (explained variance, MAE) of each to see which is the most precise at predicting user ratings.

## TV show support

Implementing TV show support as well as film, would allow for the application to have a lot more range and to appeal to a wider audience. The Movie Database API (The Movie Database, 2018) already allows developers to retrieve metadata for TV shows, so by utilising the same scripts that have already been created for the film data acquisition and altering them slightly all TV show metadata could be acquired.

## Additional trend graphs

It would be advantageous to provide the user with additional trend graphs via the *trends* page, but also allow for more user interactivity by having some form of filtering option available on the existing and any future graphs. An added interactive feature could include the option to search for a specific production company for example and to show a variety of trend analytics specifically for that search query.

## External ratings

Requesting ratings from other external sources such as Rotten Tomatoes and/or Metacritic would not only allow users to see all ratings for a film in a singular location but could also help influence prediction models by providing these specific ratings as additional features. Rotten Tomatoes for example, has a *want-to see list*, which indicates how many users have added an unreleased film to their want-to see list. So, utilising this feature could help increase the accuracy of the model by revealing the percentage of users that are already interested in seeing the film.

A similar idea could be applied to both Twitter and Facebook pages whereby, users can follow or like pages that have been created for upcoming films. But, this could be taken a step further and tweets and/or Facebook comments could be analysed using sentiment analysis to determine whether the content is positive or negative and again used as additional features for training the model.

## Updater function

This is an important function to have for an application like this one, whereby on a set schedule (once a week, or once a month) a function would trigger that would be responsible for looking at all existing data in the database and comparing it against The Movie Database API to see if any new films have been added, but more importantly to check if any existing films have been updated or altered in anyway. If a film’s metadata has changed, then the function will retrieve the new data and replace the old data in the database.

Once updated or additional data has been acquired, the function will then be responsible for re-training the model to account for the new data. This will ensure that as new data enters the system, all prediction rating values are also updated to better reflect any changes in the relationships that the model had previously been trained on.

## Targeting businesses

By adding additional functionality to the application whereby allowing users to configure their own custom test cases for unreleased films, this application could be targeted toward production companies. These industry experts are always trying to find the *winning* formula for a film and one that will generate the highest amount of revenue. If they had the ability to compute different combinations of variables such as, which actors, or which director would best suit an upcoming film based on the predicted rating output, then it would be extremely beneficial and lucrative for them.

## Oscar nominees and winners

A very interesting model feature was considered toward the end of the project and as such, has been constructed into a future improvement. The idea centred around utilising Oscar nomination, winner and potentially other film award or accolade data, as an additional feature when training the model; which could have a huge advantage on the outcome of the prediction values. It would work so that, whenever a film is either nominated or wins, a unique category would then be populated for each attribute of the film. Taking this year’s Academy Awards for example, Blade Runner 2049 won an Oscar in best visual effects category (Donnelly, 2018). So, each film attribute, the genres, the actors, the director, the production company and so on; would also receive numerical value appended onto this unique category mentioned above. Each value could also be weighted differently depending on the category in which it has originated from and depending on whether the end result was simply a nomination or an actual win.

# References

Donnelly, J., 2018. *Oscar Winners 2018: See the Full list.* [Online]   
Available at: http://oscar.go.com/news/2018/oscar-winners-2018-see-the-full-list

f.lux, 2018. *justgetflux.* [Online]   
Available at: https://justgetflux.com/

Sas.com, 2018. *Machine Learning - What it is and why it matters.* [Online]   
Available at: https://www.sas.com/en\_gb/insights/analytics/machine-learning.html

The Movie Database, 2018. *Get TV Details API Documentation.* [Online]   
Available at: https://developers.themoviedb.org/3/tv/get-tv-details

# Appendices

## Appendix 1: Project Initiation Document (PID)

Individual Project (CS3IP16)

Department of Computer Science

University of Reading

Project Initiation Document

## PID Sign-Off

|  |  |
| --- | --- |
| **Student No.** | **23013043** |
| **Student Name** | **Kane Small** |
| **Email** | **k.small@student.reading.ac.uk** |
| **Degree programme** (BSc CS/BSc IT) | **BSc CS** |
|  |  |
| **Supervisor Name** | **Jonathan Boyle** |
| **Supervisor Signature** | **JONATHAN BOYLE** |
| **Date** | **04/10/2017** |

# SECTION 1 – General Information

## Project Identification

|  |  |
| --- | --- |
| **1.1** | **Project ID**  (as in handbook) |
|  | Own project |
| **1.2** | **Project Title** |
|  | Data mining film data for trend analysis |
| **1.3** | **Briefly describe the main purpose of the project in no more than 25 words** |
|  | The main purpose of this project is to develop a web application that collects film data and provides trend analysis via automated graphing techniques. |

## Student Identification

|  |  |
| --- | --- |
| **1.4** | **Student Name(s), Course, Email address(s)**  e.g. Anne Other, BSc CS, a.other@student.reading.ac.uk |
|  | Kane Small, BSc CS, k.small@student.reading.ac.uk |

## Supervisor Identification

|  |  |
| --- | --- |
| **1.5** | **Primary Supervisor Name, Email address**  e.g. Prof Anne Other, a.other@reading.ac.uk |
|  | Jonathan Boyle, j.n.boyle@reading.ac.uk |
| **1.6** | **Secondary Supervisor Name, Email address**  Only fill in this section if a secondary supervisor has been assigned to your project |
|  |  |

## Company Partner (only complete if there is a company involved)

|  |  |
| --- | --- |
| **1.7** | **Company Name** |
|  |  |
| **1.8** | **Company Address** |
|  |  |
| **1.9** | **Name, email and phone number of Company Supervisor or Primary Contact** |
|  |  |

# SECTION 2 – Project Description

|  |  |
| --- | --- |
| **2.1** | **Summarise the background research for the project in about 400 words. You must include references in this section but don’t count them in the word count.** |
|  | The background research for the project will begin by comparing the existing film aggregation and review platforms available today and gathering statistics on each site such as the total number of films stored, the number of unique attributes listed, as well as information about the users who visit on a daily basis. From this initial information, it will give a better understanding of the trends of film accumulation and the popularity of the film industry as a whole.  Research into the top open source film database API’s (Application Programming Interface) and how much data they provide access to will be conducted, providing an overview on which film trends to analyse. As it stands, the most prominent candidate is ‘The Movie Database’, which has a vast collection of films at over 400,000 and provides access to information such as cast, crew, plot keywords, release information, reviews and more. The disadvantage to utilising this API is that it does not allow of the entire film database to be dumped/extracted and would therefore have to be queried incrementally across the film ids, which would take a long time due to the requests cooldown period.  Furthermore, there are multiple options for handling the data requested via the API. All of the required data could be pulled into a singular local database, which would have the advantage of reducing the number of requests having to be made however, the disadvantage would be that the data would not be displayed in real time. Another option would be a tool such as ‘Elasticsearch’, which not only handles large volumes of data well and queries datasets almost instantly but is highly scalable; so would be able to handle the ever-increasing additions of new films to the database. Though, as this is a small-scale project as opposed to a business application implementing Elasticsearch may not be required and/or manageable in the given timeframe.  Finally, research will need to be carried out with regards to the different data mining techniques available. Currently, the four main techniques that will be compared are association, classification, clustering and prediction. Prediction is the most versatile method and would allow for additional features to be added to the project at a later date. Once a more detailed overview of each technique has been obtained, a selection will be made on which technique(s) to employ, primarily based on the advantages and disadvantages of each. From the selected technique(s), a method to present the data in a meaningful format will be considered. This will include textual data alongside such mediums as graphs and charts.  IMDb: <http://www.imdb.com>  The Move Database API: <https://www.themoviedb.org/documentation/api>  Data mining techniques: <https://www.ibm.com/developerworks/library/ba-data-mining-techniques/> |

|  |  |
| --- | --- |
| **2.2** | **Summarise the project objectives and outputs in about 400 words.** These objectives and outputs should appear as tasks, milestones and deliverables in your project plan. In general, an objective is something you can do and an output is something you produce – one leads to the other. |
|  | **Objectives:**   * Compare existing film aggregation and review platforms. * Learn how to use The Movie Database API, or an equivalent service. * Learn how to implement Amazon Web Services Lambda Functions. * Compare and select the required data mining technique(s). * Identify and analyse the limitations of the data mining technique(s) selected. * Identify how these systems can be exploited and ultimately mitigated against, for example preventing too many requests within a certain time period. * Design, build and configure the database to store the film data. * Design and build the front-end web interface. * Test the system’s functionality. * Review all tests and make any required improvements/fixes. * Compile a report of the entire process from planning to completion.   **Outputs:**   * Produce a document detailing the comparison between the film aggregation and review platforms. * Produce a web application that illustrates the current trends among film attributes. * Produce a testing plan. * Produce a final report detailing each stage of the development process of this program. * Output a reflection analysing how the project has been executed and what has been successful/unsuccessful. |

|  |  |
| --- | --- |
| **2.3** | **Initial project specification - list key features and functions of your finished project.** Remember that a specification should not usually propose the solution. For example, your project may require open source datasets so add that to the specification but don’t state how that data-link will be achieved – that comes later. |
|  | The following key features and functions will be present in this project:   * The system will deliver trend analysis on film data queried via an open source API. * A web front-end interface will be created to showcase the trend analytics. * Data will be presented to the user in both a graphical and textual format. * The analytics graphs will update automatically whenever new data is received from the API. * Users will have the option to filter the analytics based on certain information such as genre, budget and user rating. * The system will function across all popular web browsers, from Google Chrome to Safari. |
| **2.4** | **Describe the social, legal and ethical issues that apply to your project. Does your project require ethical approval?** |
|  | The project will not require any ethical approval, as all of the film data will be acquired via an open source API. There will also be no social or legal issues that apply to this project, due to the nature of the source of the data that will be acquired and the fact that said data will only be used for analytic purposes and not for any monetary gains. |
| **2.5** | **Identify and lists the items you expect to need to purchase for your project. Specify the cost (include VAT and shipping if known) of each item as well as the supplier.** e.g. item 1 name, supplier, cost |
|  | All of the content required for this project is open source and can be found online. |
| **2.6** | **State whether you need access to specific resources within the department or the University e.g. special devices and workshop** |
|  | All of the content required for this project is open source and can be found online, therefore there is no need for any specific departmental resources. |

# SECTION 3 – Project Plan

|  |  |  |  |
| --- | --- | --- | --- |
| **3.1** | **Project Plan**  Split your project work into sections/categories/phases and add tasks for each of these sections. It is likely that the high-level objectives you identified in section 2.2 become sections here. The outputs from section 2.2 should appear in the Outputs column here. Remember to include tasks for your project presentation, project demos, producing your poster, and writing up your report. | | |
|  | | | |
| **Task No.** | **Task description** | **Effort**  **(weeks)** | **Outputs** |
| **1** | **Background Research** | **2.6** |  |
| 1.1 | Research existing film aggregation and review platforms | 2 days | Research document |
| 1.2 | Research and compare data mining techniques | 2 days | Comparison document |
| 1.3 | Research database design | 1 | Research document |
| 1.4 | Research front-end frameworks | 1 | Research document |
| 1.5 | Research limitations to current data mining techniques | 2 days | Research document |
| **2** | **Analysis and design** | **2.3** |  |
| 2.1 | Context analysis | 3 days | Case model diagram |
| 2.2 | Use cases | 2 days | Use case diagram |
| 2.3 | Functional workflow | 2 days | Flow diagram |
| 2.4 | Back-end database design | 5 days | Database design |
| 2.5 | Trend algorithm design | 5 days | Flowchart/pseudocode |
| **3** | **Develop prototype** | **8** |  |
| 3.1 | Develop the database | 2 | A database to store the film data. |
| 3.2 | Develop the front-end interface | 3 | A web interface. |
| 3.3 | Develop the logic for trend analysis | 3 | The logic for the trend analysis in code form. |
| **4** | **Testing, evaluation/validation** | **4.2** |  |
| 4.1 | Unit testing | 1 | Testing of each individual component. |
| 4.2 | System testing | 1 | Testing carried out for the entire system, from end-to-end. |
| 4.3 | Functional testing | 1 | Testing of all requirements. |
| 4.4 | Usability testing | 1 | Testing of the user experience. |
| 4.5 | Live-testing with users | 2 days | Testing the application with actual end-users. |
| **5** | **Assessments** | **5.3** |  |
| 5.1 | Write-up project report | 4 | Project Report |
| 5.2 | Produce poster | 3 days | Poster |
| 5.3 | Prepare presentation/demonstration | 1 | Presentation/demo |
| **TOTAL** | **Sum of total effort in weeks** | **23** |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **SECTION 4 - Time Plan for the proposed Project work** | | | | | | | | | | | | | | | | | | |
| For each task identified in 3.1, please *shade* the weeks when you’ll be working on that task. You should also mark target milestones, outputs and key decision points. To shade a cell in MS Word, move the mouse to the top left of cell until the curser becomes an arrow pointing up, left click to select the cell and then right click and select ‘borders and shading’. Under the shading tab pick an appropriate grey colour and click ok. | | | | | | | | | | | | | | | | | | |
| **Project stage** | **START DATE: /10/2017****Project Weeks** | | | | | | | | | | | | | | | | | |
| 0-3 | | 3-6 | | 6-9 | 9-12 | 12-15 | | 15-18 | | 18-21 | 21-24 | | 24-27 | 27-30 | 30-33 | 33-36 | 36-39 |
| 1 Background Research |  |  |  | |  |  |  | |  | |  |  | |  |  |  |  |  |
|  |  |  |  | |  |  |  | |  | |  |  | |  |  |  |  |  |
|  |  |  |  | |  |  |  | |  | |  |  | |  |  |  |  |  |
| 2 Analysis/Design |  |  |  |  |  |  |  | |  | |  |  | |  |  |  |  |  |
|  |  |  |  |  |  |  |  | |  | |  |  | |  |  |  |  |  |
|  |  |  |  |  |  |  |  | |  | |  |  | |  |  |  |  |  |
| 3 Develop prototype. |  | |  |  |  |  |  |  |  | |  |  | |  |  |  |  |  |
|  |  | |  |  |  |  |  |  |  | |  |  | |  |  |  |  |  |
|  |  | |  |  |  |  |  |  |  | |  |  | |  |  |  |  |  |
| 4 Testing, evaluation/validation |  | |  | |  |  |  |  |  |  |  |  | |  |  |  |  |  |
|  |  | |  | |  |  |  |  |  |  |  |  | |  |  |  |  |  |
|  |  | |  | |  |  |  |  |  |  |  |  | |  |  |  |  |  |
| 5 Assessments |  | |  | |  |  |  | |  |  |  |  |  |  |  |  |  |  |
|  |  | |  | |  |  |  | |  |  |  |  |  |  |  |  |  |  |
|  |  | |  | |  |  |  | |  |  |  |  |  |  |  |  |  |  |
|  |  | |  | |  |  |  | |  |  |  |  |  |  |  |  |  |  |

**Risk Assessment Form**

|  |  |  |  |
| --- | --- | --- | --- |
| **Assessment Reference No.** |  | **Area or activity assessed:** |  |
| **Assessment date** |  |
| **Persons who may be affected by the activity (i.e. are at risk)** |  |

**SECTION 1: Identify Hazards -** *Consider the activity or work area and identify if any of the hazards listed below are significant (tick the boxes that apply).*

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Fall of person (from work at height) |  |  | Lighting levels |  |  | Use of portable tools / equipment |  |  | Vehicles / driving at work |  |  | Hazardous fumes,  chemicals, dust |  |  | Occupational stress |  |
|  | Fall of objects |  |  | Heating & ventilation |  |  | Fixed machinery or lifting equipment |  |  | Outdoor work / extreme weather |  |  | Hazardous biological agent |  |  | Violence to staff / verbal assault |  |
|  | Slips, Trips & Housekeeping |  |  | Layout, storage, space, obstructions |  |  | Pressure vessels |  |  | Fieldtrips / field work |  |  | Confined space / asphyxiation risk |  |  | Work with animals |  |
|  | Manual handling operations |  |  | Welfare facilities |  |  | Noise or Vibration |  |  | Radiation sources |  |  | Condition of Buildings & glazing |  |  | Lone working / work out of hours |  |
| 1. **55** | Display screen equipment | **✓** |  | Electrical Equipment |  |  | Fire hazards & flammable material |  |  | Work with lasers |  |  | Food preparation |  |  | Other(s) - specify | **✓** |

**SECTION 2: Risk Controls** *- For each hazard identified in Section 1, complete Section 2.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Hazard No**. | Hazard Description | Existing controls to reduce risk | **Risk Level** (tick one) | | | Further action needed to reduce risks |
|  | High | Med | Low | *(provide timescales and initials of person responsible)* |
| 5 | Extended periods of time looking at the computer screen | Regular breaks, computer glasses that prevent glare and/or eye strain. Potentially look into using a program such as *f*.*lux*, to reduce blue light exposure. |  | X |  |  |
| 30 | Hardware failure | Back up device on a regular basis. |  | X |  | Backup all project work on cloud services as well as external HDD’s to minimise the risk of irretrievable data loss as much as possible. |
| 30 | Repetitive strain injury | Take regular breaks, use good posture, ensure that your workstation is ergonomically designed/setup and make sure not to over-exert yourself whilst using the computer, i.e. stretching for hard-to-reach keys. |  | X |  |  |
| **Name of Assessor(s)** | |  | **SIGNED** | | | |
| **Review date** | |  |

## Appendix 2: Logbook