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** | |  |