FINAL REPORT

COMP5703 IT Capstone Project

Supervised by Prof. Seokhee Hong

Scalable Visual Analytics For Movie Data

By

Marnijati Torkel (200251384)

University of Sydney

November 2017

Acknowledgements

I would like to express my appreciation and thank my supervisor, Prof. Seokhee Hong, for the guidance, support and great advice she has offered throughout my time as her student. I would also like to thank Dr. Quan Nguyen for his feedback on my project.

ABSTRACT

This project presents *design, implementation* and *evaluation* of Scalable Visual Analytics framework for movie data. The *design* of the project consists of analysis, visualisation and interaction modules. The Analysis module provides centrality, clustering, time slice, collaboration and temporal dynamic. The Visualisation module generates node-link diagram, Wordcloud layout, Streamgraph, Sunburst diagram, Histogram, Line graph and stacked bar plot. The Interaction module allows zoom, filter, selection, interactive slider, hover, drag and click. The *implementation* of this project is the development of an interactive web application using R Shiny. There are two main tasks in the *evaluation*, to determine the most influential actors and to identify movie genre trends. To determine the most influential actors, interactive temporal dynamic networks and collaboration analysis are applied. Temporal dynamic two-mode networks are used to identify movie genre trends. This interactive web application developed for this project was able to provide fast and flexible solution for the tasks.

CONTENTS

1. INTRODUCTION

1.1. Background

In recent years, the advancement of technology in Multimedia, Social Media, Internet of Things and Cloud Computing has led to a massive increase in the amount of data generated and stored. To extract meaningful data quickly is often challenging. The ability to analyse huge and complex data is becoming increasingly important so that decision can be made efficiently [5].

To address this challenge, scalable visual analytics are used to extract knowledge from complex data and communicate effectively with the people. Good interactive visualization can provide rapid insight into complex data [5].

IMDB (Internet Movie Database) is a popular online database of information related to movies from around the world. It has over 4 millions titles, 8 millions casts and 250 million unique monthly users [1, 2]. IMDB is a large and complex temporal multivariate database [5]. The Numbers is a website which provides detailed movie financial analysis [6].

For this project, the IMDB5000 dataset from Kaggle is used. Kaggle is a data science and machine learning platform, in which quality datasets are posted. This dataset is a subset of the IMDB (Internet Movie Data Base) based on movies in The Numbers website [6]. It consists of 5043 records and 28 attributes. They are movies from the years 1927 to 2016. The attributes used for this project are movie titles, director names, actor names, genres, gross, years and ratings.

1.2. Motivation

The Scalable Visual Analytics website of the University of Konstanz states “The objective of scalable visual data analysis is to represent data graphically, so that structural connections and relevant characteristics of the data can be easily understood” [8].

The motivation is to design, implement and evaluate an interactive web application of scalable visual analytics for movie data.

For this project, movie data from Kaggle, the IMDB5000 is used. This dataset is interesting and challenging. This dataset has large and complex temporal multivariate features. We use this dataset to gain insight into the structure and relationships of the dataset. We introduce different analytics and visualization techniques to perform scalable visual analytics for this dataset.

1.3. Aims and Contribution

The aim of this project is to perform different tasks so that we can demonstrate different techniques in complex data analysis and visualisation. The tasks to perform for this dataset are to determine the most influential actors and identify movie genre trends.

The contributions for this project in *design, implementation and evaluation* of Scalable Visual Analytics framework for movie data are as follows:

- The *design* of the scalable visual analytics for movie data consists of three main modules including Analysis, Visualisation and Interaction. The Analysis module covers centrality analysis, clustering analysis, time slice analysis, collaboration analysis and temporal dynamic analysis. The Visualisation module generates node-link diagrams, Wordcloud layouts, Streamgraph layouts,

stacked bar charts, histogram, line graphs and Sunburst diagrams. The Interaction module allows zooming, dragging, selecting, hovering and clicking.

- The *implementation* of the interactive web application has a menu from which three different sections can be selected. They are Overview, Actors and Movies. The Overview section is for simple analysis of the IMDB5000 dataset. It uses histogram, line graph and Wordcloud for visualisation. The Actors section is to analyse and visualise co-starring network and ego network. It uses node-link diagrams, a stacked bar plot, time slice and a Sunburst diagram for visualisation. The Movies section is for identifying genre trends of movies. It uses a Streamgraph layout, a two-mode network diagram and a Wordcloud layout for visualisation.
- In *evaluation*, there are two main tasks for this project. The first task is to find out the most influential actors. This task can be performed using the Actors section of the application. The second task is to identify genre trends of movies. This task can be performed using the Movies section of the application.

1.4. Roadmap

The remainder of the report is organised as follows. **Section 2** provides an overview of *related work* in the movie data. **Section 3** describes details of scalable visual analytics for movie data *design* approach. **Section 4** presents an *implementation* of its system architecture and graphic user interface. **Section 5** *evaluates* the application on scalable visual analytics for movie data. Finally, **Section 6** concludes the report with summary, limitation and future work.

2. RELATED WORK

There have been many earlier studies on Visual Analytics using data from IMDB. These studies used a range of visualization technologies such as Microsoft Excel, Google Fusion Table, Many Eyes, Gephi, Pajek, Tableau and GEOMI to perform network analysis.

2.1. Temporal Dynamic Network

In 2007, 'Visualisation and analysis of the internet movie database' performed visualisation on large network analysis, network analysis and visualisation on Kevin Bacon and temporal dynamic network analysis and visualisation [5]. Figure 1 shows co-starring actors visualisation in 2000s [5].

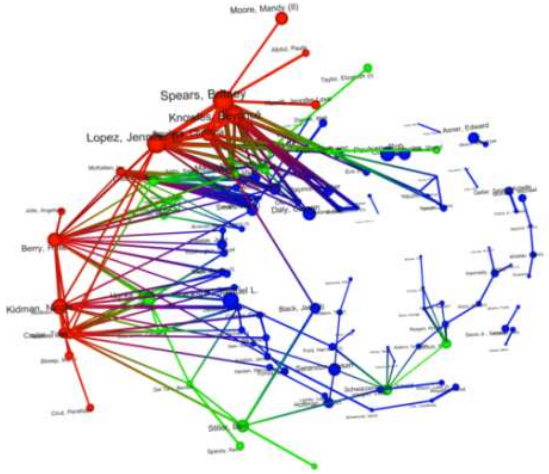

Figure 1. Co-starring actors visualisation (2000s) [5]

2.2. Community Detection

In 2016, Joshua Johnson's website 'Key Player & Community detection using Graph Theory on IMDB Movie Ratings' used Gephi to gain insights into actors, directors and movies relationships using one, two and three-mode network analysis and visualisation [4]. Figure 2 use modularity to cluster actors in different colours.

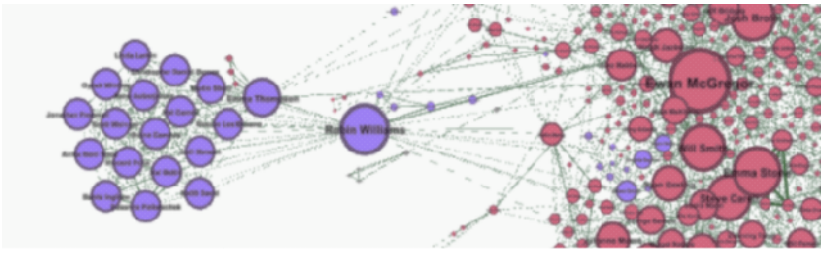

Figure 2. Modularity runs in Gephi [4]

2.3. Genre Analysis

In 2014, 'Information Analysis of Movie Genres' looked at relationship between movies and genres over the years using temporal dynamic network analysis, stacked bar plot and Wordcloud layout [6]. Figure 3 displays the relationships between movie and genres in the year 2000.

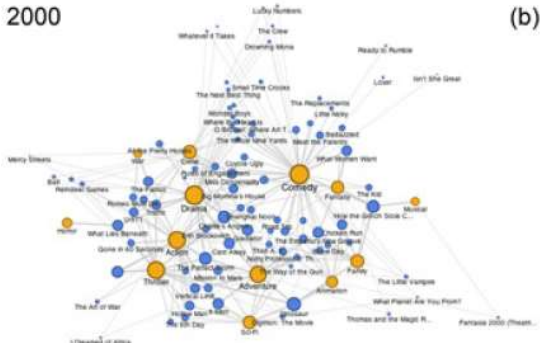

Figure 3. Movie genres network (2000) [6]

2.4. Movie Analysis

In 2016 and 2017, Kaggle users have posted a variety of analytics using Python and R for this dataset [3]. Figure 4 shows genres represented in the movie data. Figure 6 shows percentage of movies in decade

Figure 4. Genre representation in movie data

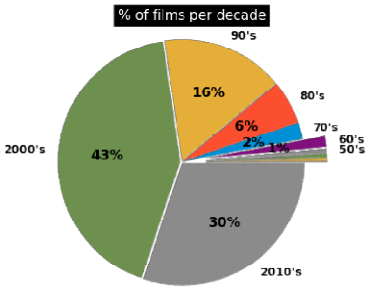

Figure 5. % of movies per decade

These scalable visual analytics studies on movie data inspired me to design, implement and evaluate our visual analytics system using movie data.

3. DESIGN

This section describes the design of our visual analytics system. This design integrates analysis techniques with interactive visualisation to gain insight of the movie data (see Figure 6).

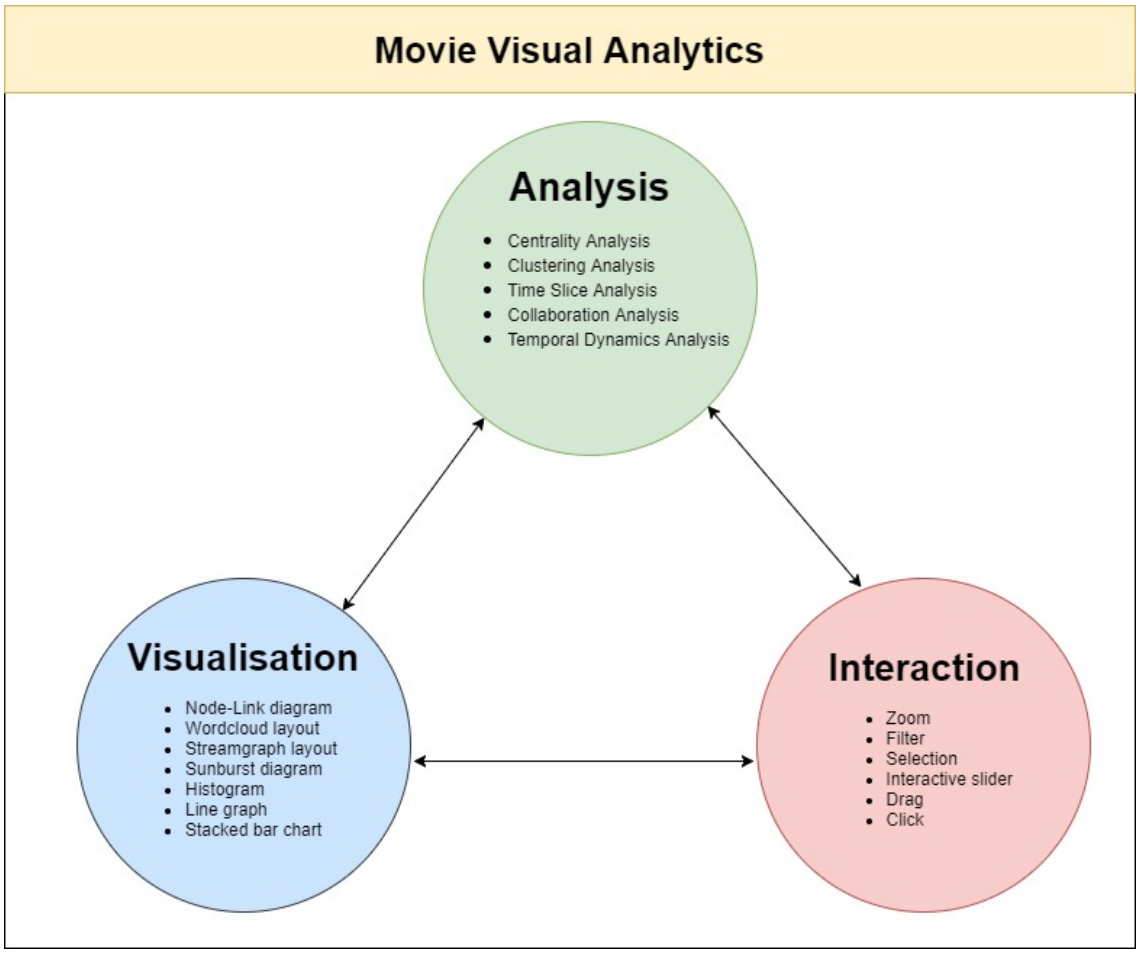

Figure 6. Design Framework Diagram

A detailed description of each module of the design scalable visual analytics in movie data is listed below:

3.1. Analysis

In the analysis module, network analysis methods with *centrality* are used to analyse co-starring relationships and movie genres relationships. Network *clustering* algorithms are used for groups' actors together in the same clusters. Time slice analysis [5] shows in which periods the actors are starring in the movies. Collaboration analysis shows actor and director relationships. Temporal dynamic analysis is applied separately to actors and genres.

An introduction to network graphs: Network graphs consist of nodes and edges. The edges are either directed or undirected. A directed graph is a graph with directed edges (with direction) and an undirected graph is a graph with undirected edges.

In network graph, centrality is a measure of how important a node is compare to other nodes based on its links with other nodes.

The following explains **centrality** and **clustering** analysis.

3.1.1. Centrality analysis

There are several centrality measures. This project uses degree [13], closeness [14] and betweenness [14] centralities.

- **Degree Centrality:**

The simplest centrality is degree centrality which measures how many links one node has. For directed graph, there are two different measures, in-degree and out-degree. In-degree measures how many incoming links and out-degree measures how many outgoing links [13].

- **Betweenness Centrality**

If one were to find the shortest paths between all pairs of nodes, betweenness centrality for a node is a measure of the number of such paths the node is on [14].

- **Closeness Centrality**

Closeness centrality measures the average length of the shortest path between a node and all other nodes [14].

3.1.2. Clustering analysis

Clustering analysis is the task to simplify the graph by detecting communities in the network. Modularity is used to measure the connection strength between nodes in the network communities [17].

In this project, four community detection algorithms from the "igraph" package are used. The algorithms are *Fast Greedy*, *Walktrap*, *Leading Eigenvector* and *Edge Betweenness* [17].

- **Fast Greedy**

The Fast Greedy algorithm was proposed by Clauset et al [18]. This algorithm is based on bottom up hierarchical clustering that optimises the modularity score. Individual nodes are merged to form communities repeatedly until the modularity reaches maximum score [17].

- **Walktrap**

The Walktrap algorithm was proposed by Pon & Latapy [19]. The algorithm is based on random walks and bottom up hierarchical clustering. This approach performs short random walks to generate communities. These communities are merged to form communities repeatedly [17].

- **Edge Betweenness**

The Edge Betweenness algorithm was proposed by Girvan & Newman [20].This algorithm is based on top down hierarchical clustering. Edges with high edge betweenness scores are remove iteratively [17].

- **Leading Eigenvector**

The Leading Eigenvector algorithm was proposed by Newman [22]. This algorithm is based on top down hierarchical clustering that optimises the modularity score. First, the graph is split into two parts based on the value of leading eigenvector of the modularity matrix. It stops when the modularity score is negative [17].

3.2. Visualisation

Visualisation is a technique of presenting data graphically in order to gain insights into the data [24].

In the visualisation module, there are different visualisation methods to represent data. Node-Link diagram is a common method to represent network data. Data frequencies can be viewed with *Wordcloud* layouts or line graphs. Temporal data can be visualised with *Streamgraph* or stacked bar

plots. Hierarchical data can be visualised with *Sunburst* diagrams [11]. Simple data distribution can be viewed with histograms.

Below are the detailed explanation of *a* Node-Link diagram, a Wordcloud layout, a Streamgraph layout and a Sunburst diagram [11].

3.2.1. Node-Link diagram

Node-Link diagrams are a type of visualisation that shows the connections of nodes and links (see Figure 7). A node is represented by a circle, a square, a text or an image and a link is represented by a line or an arrow [11].

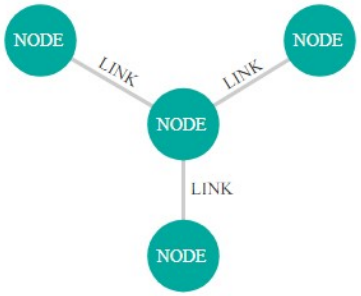

Figure 7. Node link diagram [11]

3.2.2. Wordcloud layout

Wordcloud layouts are a type of visualization that has relevant words in an image. The size of each word is based on its frequency (see Figure 8) [11].

Figure 8. Wordcloud layout [11]

3.2.3. Streamgraph layout

Streamgraph layouts are a type of visualization based on stacked area graphs which show the value of some measure over time. Each category is represented by a colour. The height of each category at any point in time corresponds to the value of the measure at that point in time. The graph is centered on a horizontal axis (see Figure 9) [11].

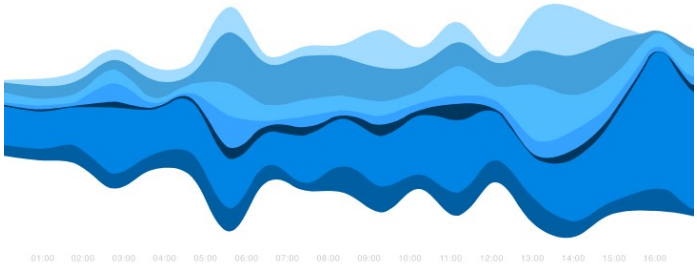

Figure 9. Streamgraph [11]

3.2.4. Sunburst diagram

Sunburst diagrams are a type of visualisation that displays hierarchical data. Each level of the hierarchy is represented by a ring. The top of the hierarchy is represented by the innermost ring and the hierarchy moves outwards. Each section of each ring is given a size representing some numerical value. Different colours can be used to distinguish the different rings and sections (see Figure 10) [11].

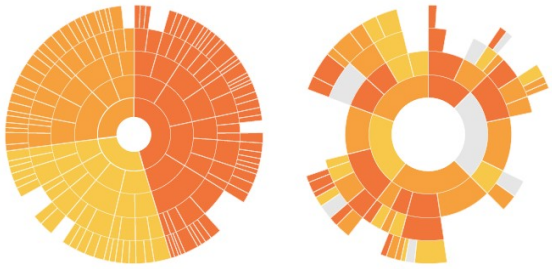

Figure 10. Sunburst Diagram [11]

3.3. Interaction

User Interaction in a visual analytics system is to enable users to perform visual data exploration. Interaction encourages users to participate in the process of analysing and understanding data [12].

In the interaction module, zooming, dragging, filtering, selecting, hovering and clicking are used.

- Zoom is to let the user zoom in or out of the visualisation.
- Drag is to move location of nodes on the visualisation.
- Filter is restricting the visualised data to a subset of the available data based on user-selected criteria.
- Select is to let user select a part of the graph.
- Hover is to move the mouse over a node.
- Click is to let user perform click action on the application.

4. IMPLEMENTATION

Scalable visual analytics has to cope with large and complex data. We implemented an interactive web application that can perform the evaluation tasks effectively. For this implementation the system architecture and graphical user interface of this web application are as follows:

4.1. System Architecture

The flow of data for this application can be seen in Figure 11. The IMDB5000 dataset is pre-processed at the back-end of the application. This application provides users with interactive control to reveal important features of the data. The application is built with R Shiny Dashboard, a web application framework for R.

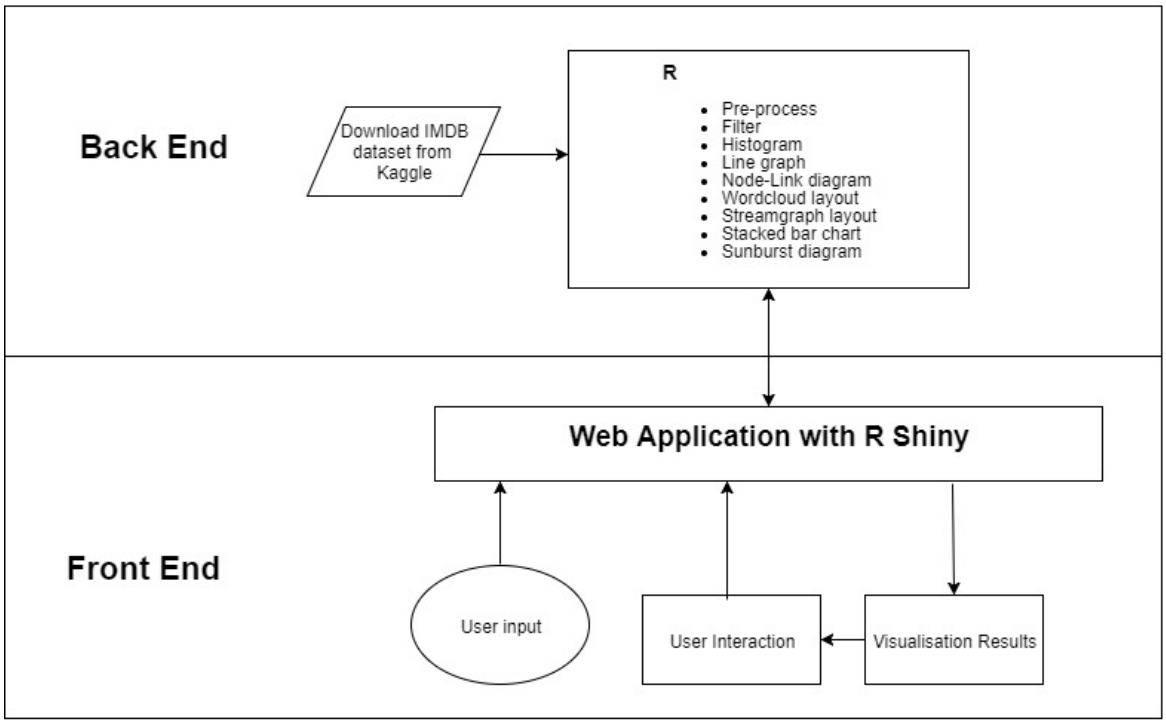

Figure 11. Flowchart of the application

4.2. System Requirements

For this project, the visual analytics system is a web application, built with R and R Shiny. R Shiny is a library that has web application functionality. Table 1 shows the system requirements for this application.

Hardware	Macbook Pro 2.4 GHz Processor 16 GB RAM
Operating System	MacOS Sierra
IDE	RStudio version 1.0.153
Language	R version 3.4.1
Libraries	devtools, dplyr, reshape2, tidyr, data.table, scales, randomcoloR, stringr, fastmatch, tm, NLP, rsconnect, jjalaire/sigma, rgexf, gplots2, wordcloud, SunburstR, igraph, networkD3, ggraph, ggforce, ggthemes, ggTimeSeries, gganimate, ggnet, GGally, intergraph, sna, animation, shiny, shinydashboard, ShinyBS, shinyJS, AlalytixWare/Shinysky
Web Browsers	Safari, Chrome

Table 1. Development environment for the application

4.3. Graphical user Interface

Graphical User Interface allows users to observe, control, extract important features and improve scalability [15]. The GUI of this application has three pages. The *Overview* page shows simple analysis of IMDB5000, the *Actors* page evaluates the most influential actors and the *Movies* page identifies movie genre trends.

4.3.1. Overview Page

Users can view simple analyses of IMDB5000 dataset attributes in histograms, line graphs or Wordcloud layouts. The attributes used for this application are directors, actors, gross, movie titles, years and genres (see Figure 12a). The GUI for simple analysis of IMDB5000 shows a Wordcloud layout of the top 20 grossing directors (see Figure 12c), as well as the average revenue figures of these directors (see Figure 12b).

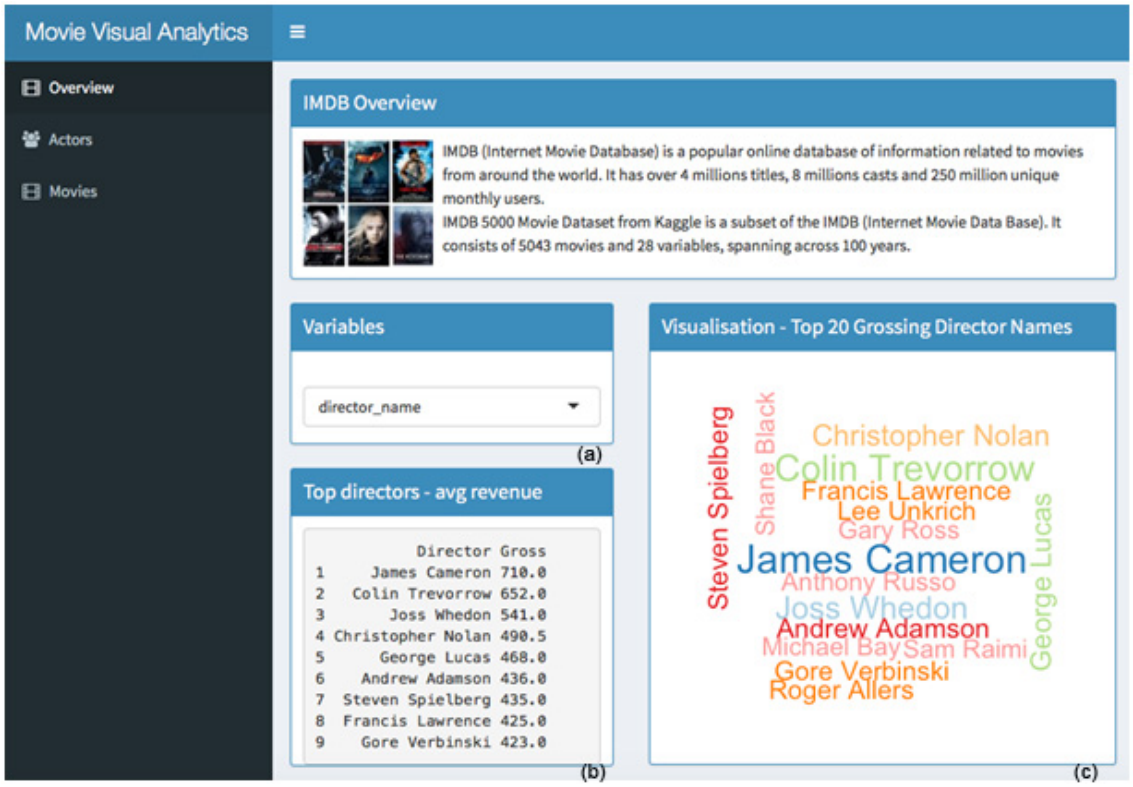

Figure 12. Main GUI. (a) Variable selection. (b) Statistical figures. (c) Graphic visualisation.

4.3.2. Actor Page

Users can determine the most influential actors based on filter selection (see Figure 13c). Visualisation for the co-starring network uses a force directed layout to generate an interaction node-link diagram. The top five highly connected actors are listed in decreasing order below the diagram (Figure 13b). Actor page filters IMDB Ratings for over 7.5. The network graph for the top 100 nodes, based on centrality is shown (see Figure 13a), as well as the names of the top five highly connected actors (see Figure 13b).

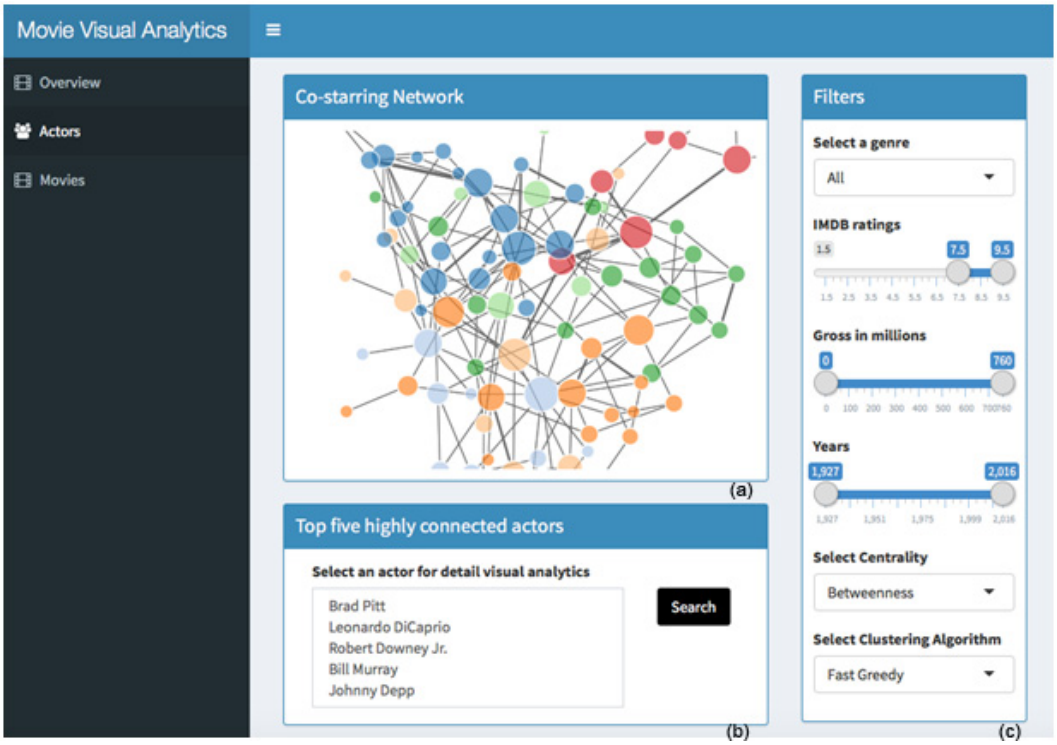

Figure 13. Actors Visualisation. (a) Co-starring Network. (b) Top five highly connected actors. (c) Filters.

- **Filters**

The filters are genres, IMDB ratings, gross and years. The increment of IMDB rating is 0.5, gross is 50 million and years is 4. The genres are Action, Adventure, Animation, Biography, Comedy, Crime, Documentary, Drama, Family, Fantasy, Film-Noir, History, Horror, Music, Musical, Mystery, Romance, Sci-Fi, Sport, Thriller, War and Western (see Figure 13).

There are different types of centralities and clustering algorithms for visualization. The centrality options are Degree, Closeness and Betweenness. The clustering algorithm options are Fast Greedy, Walktrap and Edge Betweenness (see Figure 13).

- **Co-starring Network**

Actors are represented by nodes and movies are represented by links. The relationships are actors starring in the same movies. The size of the nodes is determined by selected centrality and the thickness of the links is determined by the number of times the two actors starred in the same movie. The colour of the nodes is based on the selected clustering algorithms. For scalable purposes, only the top 100 nodes based on selected centrality are generated for visualization.

- **Top five highly connected actors**

Users can select any of the top five highly connected actors and click search button to obtain more visualisation on actors. The additional visualisations on highly connected actors are collaboration pattern between actors and directors with Sunburst diagrams, actor genre trends over the years with stack bar plots, actor networks with ego networks and time slice analysis to show in which periods the actors are starring in the movies.

4.3.3. Movies Page

Why are movie genres important? Genres are used for movie classification, decision making by producers and as a first reference point by movie-goers.

Users can identify movie genre trends based on filtering selection (see Figure 14b). Visualisation for the Genre Network Visualisation is a two-mode network diagram between movies and genres (see Figure 14a). Genre Temporal Trend shows genre frequencies over the years with Streamgraph layout (see Figure 14c). Genre Frequency displays the genre frequency with Wordcloud layout (see Figure 14d).

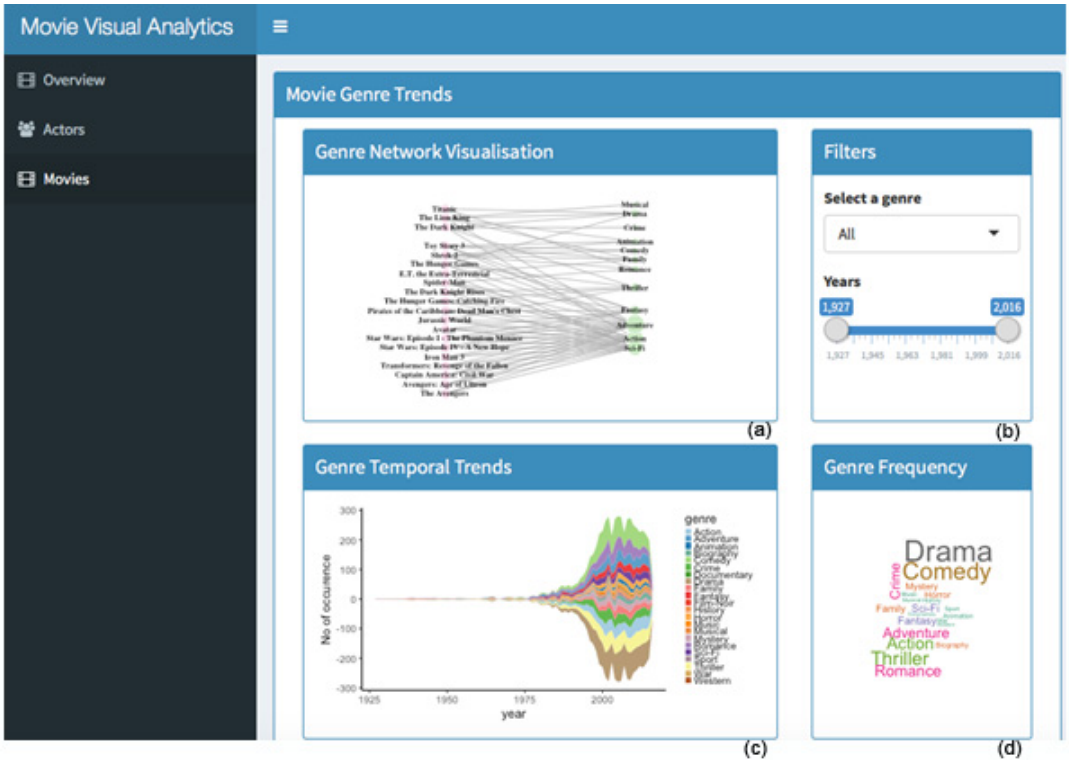

Figure 14. Movie Visualisation. (a) Genre Network Visualisation. (b) Filters. (c) Genre Temporal Trends. (d) Genre Frequency.

The filters are genres and years. The increment of years is 1. The genres are Action, Adventure, Animation, Biography, Comedy, Crime, Documentary, Drama, Family, Fantasy, Film-Noir, History, Horror, Music, Musical, Mystery, Romance, Sci-Fi, Sport, Thriller, War and Western (see Figure 15b).

- **Genre Frequency**

Popularity of the genres is based on their frequency. The larger the genre text, the more movies were made of that genre (see Figure 14d).

- **Genre Temporal Trends**

Colour represents genres. We can compare which genres occur more frequently than others based on the thickness of the graph (see Figure 14c).

- **Genre Network Visualisation**

Genres are represented by green nodes and movies are represented by pink nodes. Node sizes for genres are based on grossing. The layout of this graph is Bipartite Layout. For scalability, only the top 20 grossing movies are on display (see Figure 14a).

5. EVALUATION

To assess the effectiveness of visual analytics in the IMDB5000 dataset, two main tasks are performed. The tasks are *to determine the most influential actors* and *to identify movie genre trends*. The details of the task are as follows:

5.1. TASK 1 – Determine the most influential actors

This task is to determine the most influential actors. The first step of the task is to find the top five highly connected actors based on filtering criteria. Users can select an actor from the top five highly connected actors to perform detailed visual analytics. The second step of this task is to perform subtasks by comparing actors from the top five highly connected actors list.

The four subtasks to compare actors are

- Actor collaboration pattern with directors, movies and years
- Actor temporal genre trends
- Actor network
- Actor time slice

The result of the comparison is shown at the end of this task.

This task uses centrality analysis and clustering analysis to determine genre node size and colour respectively in Co-starring Networks. Node-Link diagrams are used for visualisation in Co-starring Networks. The interactions for the co-starring network visualisation are zoom, drag, and hover. The hover interaction shows the actor's name and the links to the co-stars, while fading out all unrelated nodes and links (see Figure 15).

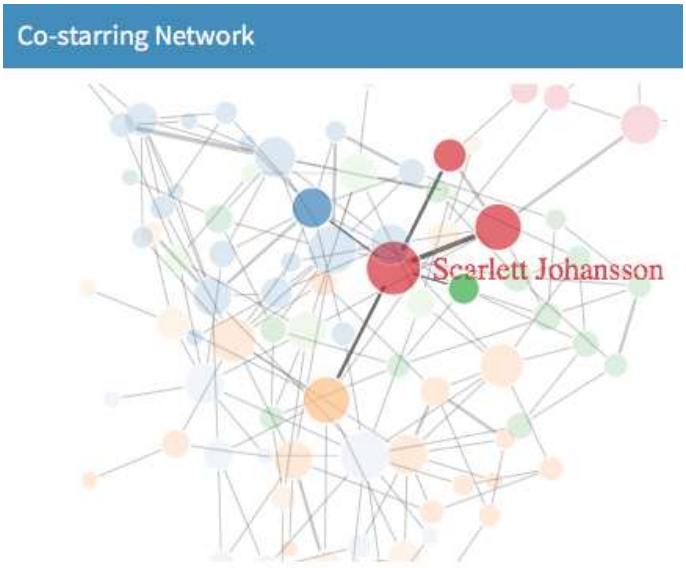

Figure 15. Hover interaction on Scarlett Johansson node

There are two experiments for the first step of this task to find the most highly connected actors. The first experiment is to filter with IMDB rating over 7.5. The second experiment is to filter with grossing over $400 million from 1991. The top five highly connected actors' names are shown in table 1 below.

Filters				Top five most highly connected actors
Years range	IMDB ratings	Box Office	Genres	
1927-2016	7.5-9.5	All	All	Brad Pitt (*) Robert De Niro Bill Murray Tom Hanks Robert Downey Jr
1991-2016	All	Over $400 million	All	Robert Downey Jr Christian Bale Jennifer Lawrence (**) Josh Hutcherson Scarlett Johansson

Table 2. Comparison of the top five highly connected actors based on different filtering

The top five highly connected actors with IMDB rating of 7.5 to 9.5 are Brad Pitt, Robert De Niro, Bill Murray, Tom Hanks and Robert Downey Jr. The top five highly connected actors with grossing over 400 million and year after 1991 are Robert Downey Jr, Christian Bale, Jennifer Lawrence, Josh Hutcherson and Scarlett Johansson.

This application allows users to present detail visual analytics on any actors from the top five highly connected actors list. To demonstrate this part of the application is to compare two diverse actors. Brad Pitt (*) an older male actor from the list with IMDB score over 7.5 and Jennifer Lawrence (**) a younger female actor from the list with grossing over $400 million (see Table 2).

5.1.1. Sub-task 1.1. Compare actor collaboration with directors, movies and years.

Collaboration analysis and Sunburst diagram visualisation is used for this subtask. Sunburst diagram uses to display actor collaboration with directors, movies and years. The innermost ring corresponds to an actor, then directors, movies and years as it moves outwards. The width of the ring corresponds to grossing of movies (see Figure 16).

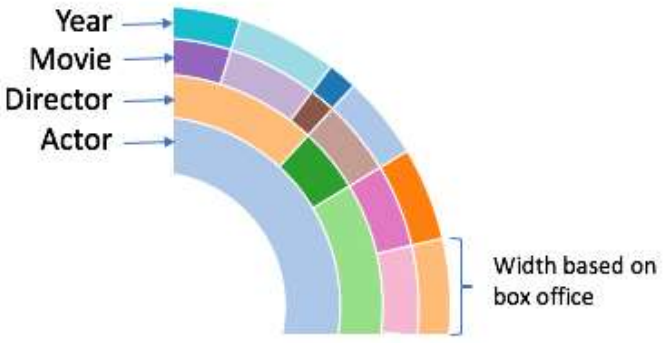

Figure 16 Sunburst diagram for actor

Figure 17 shows Brad Pitt collaboration with directors, movies and years. Brad Pitt collaborated with many directors for only one movie and the same directors for three movies twice. On hover, the name of actor, director, movie and year are shown. Figure 18 shows how hovering on a slice of the diagram reveals that Brad Pitt collaborated with Doug Liman (director) in Mr. & Mrs. Smith movie in 2005.

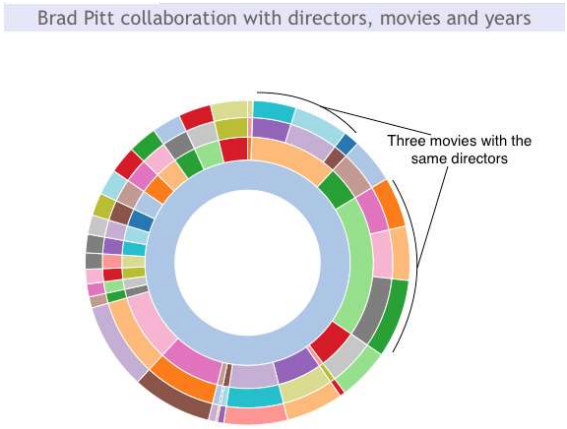

Figure 17. Brad Pitt's collaboration pattern

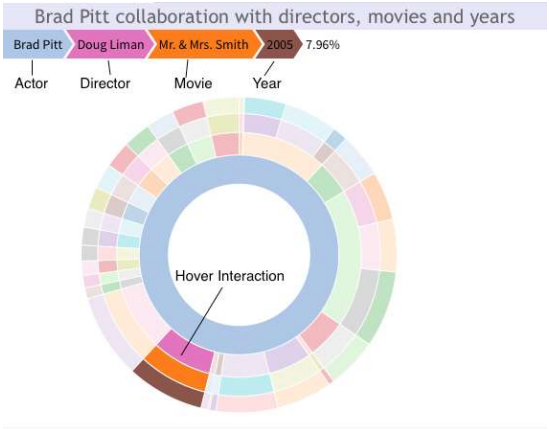

Figure 18. Hover interaction on Brad Pitt's collaboration pattern

Figure 19 shows Jennifer Lawrence collaboration with directors, movies and years. Jennifer Lawrence mainly collaborated with the same directors for two or three movies. Figure 20 shows how hovering on a slice of the diagram reveals that Jennifer Lawrence collaborated with Gary Ross (director) in The Hunger Games in 2012.

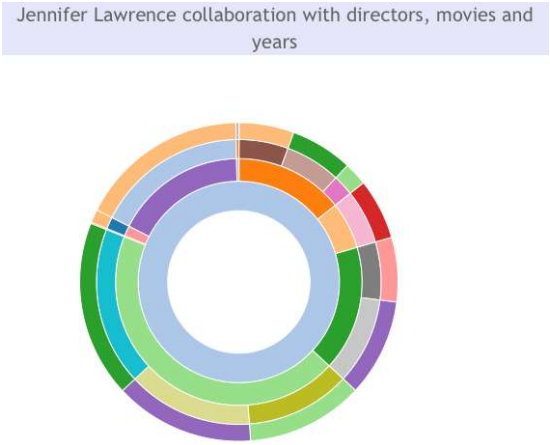

Figure 19. Jennifer Lawrence's collaboration pattern

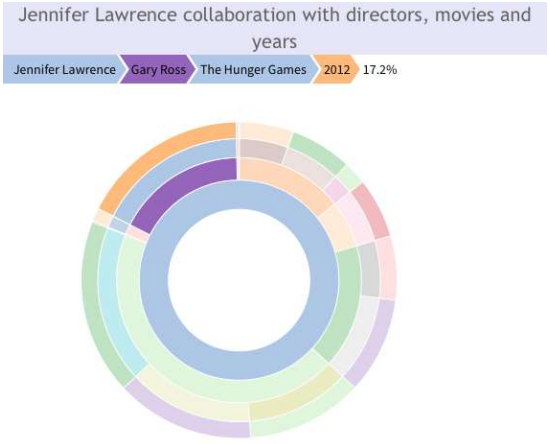

Figure 20. Hover interaction on Jennifer Lawrence's collaboration pattern

5.1.2. Sub-task 1.2. Compare actor temporal genre trends

Temporal dynamic analysis and stacked bar chart visualisation is used for this subtask. Stacked bar chart is used to display actor temporal genre trends. The chart displays how the actor starred in the movie genres changes over the years.

In the stacked bar chart, the x-axis represents the actor's active years; the y-axis represents the number of movies that the actor starred in for each year. The colour represents the movie genres for

the actor. A movie typically covers multiple genres, with the size of the rectangle for each genre representing an assessment as to how much it covers the genre (see Figure 21).

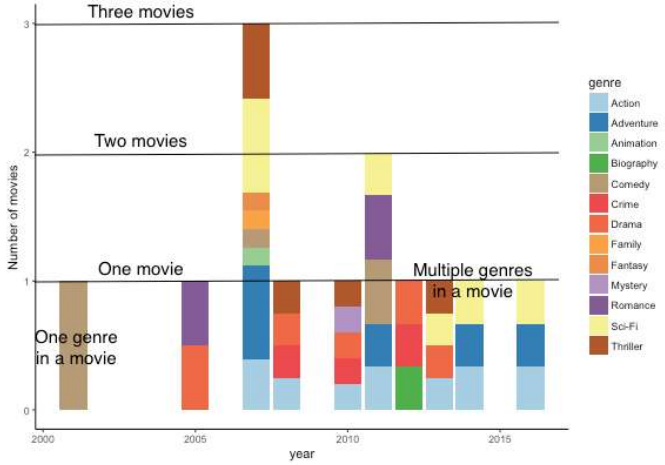

Figure 21. Actor Genre Trends

Figure 22 shows that the stacked bar chart is thinner and denser because Brad Pitt's has a long acting career and the years span from 1993 to now. Brad Pitt starred in a wide range of genres over the years but consistently starred in drama genres. Brad Pitt's movies cover up to six genres in a single movie.

Figure 23 shows that the stacked bar chart is wider because Jennifer Lawrence started her acting career in 2010. Lawrence's earlier movie genres are based on drama and romance. Later in her career, she acted in more sci-fi and adventure movies.

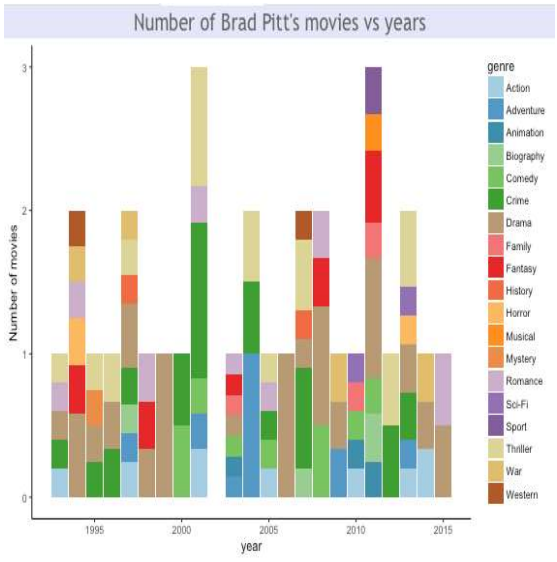

Figure 22. Brad Pitt Genre Trends

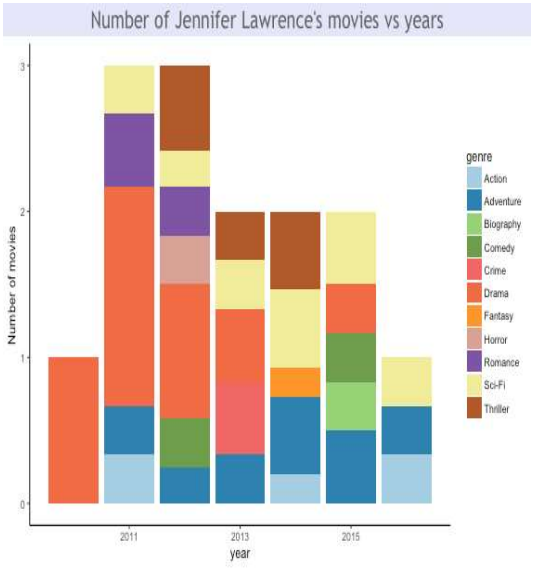

Figure 23. Jennifer Lawrence Genre Trends

5.1.3. Sub-task 1.3. Compare actor networks

Centrality and clustering analysis and Node-Link diagram visualisation are used in this subtask. Actor network diagram show actors' co-starring networks. In Node-Link diagrams, node size is based on

betweenness centrality, colour is based on leading eigenvector clustering algorithm, link width based on the number of movies in which the actors starred together. Users can hover on a node to show the name of the actor for that node. Click on a node to show the movie titles and years the actors starred in together.

Figure 25 shows Brad Pitt has worked with a lot of different actors in his career. Thus he has a lot of nodes in his network diagram. The thickness of the links indicates that Brad Pitt did not often collaborate with the same actors repeatedly. The hover interaction shows one of Brad Pitt co-stars is Angeline Jolie Pitt. The click interaction shows Brad Pitt and Angeline Jolie Pitt starred in "Mr. & Mrs. Smith" in 2005 and "By the sea" in 2015 (see Figure 26).

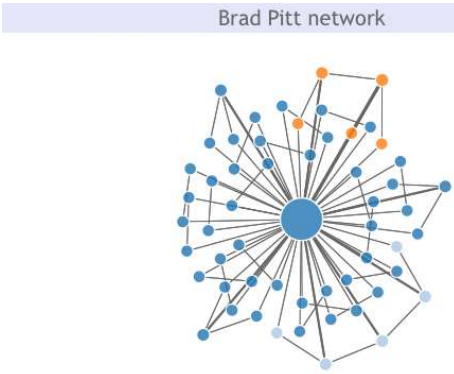

Figure 25. Brad Pitt network

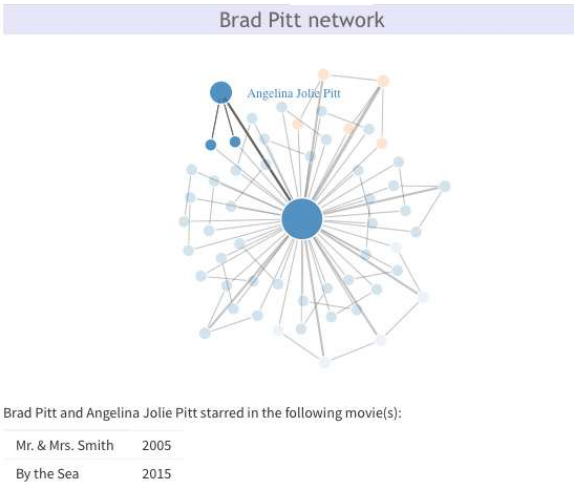

Figure 26. Click and hover interaction of Brad Pitt network

Figure 27 shows Jennifer Lawrence has a small network due the number of nodes present and she collaborated with the same actors repeatedly as shown in the thickness of the links. The hover interaction shows one of her co-stars is Josh Hutcherson. The click interaction shows Jennifer Lawrence and Josh Hutcherson co-starred in "The Hunger Games" in 2012-2015.

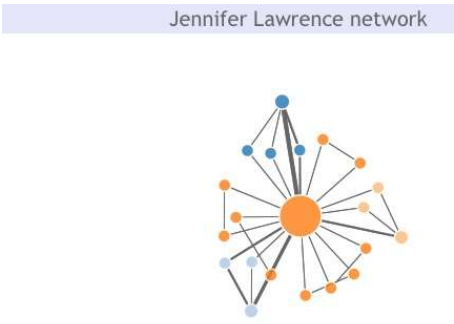

Figure 27. Jennifer Lawrence network

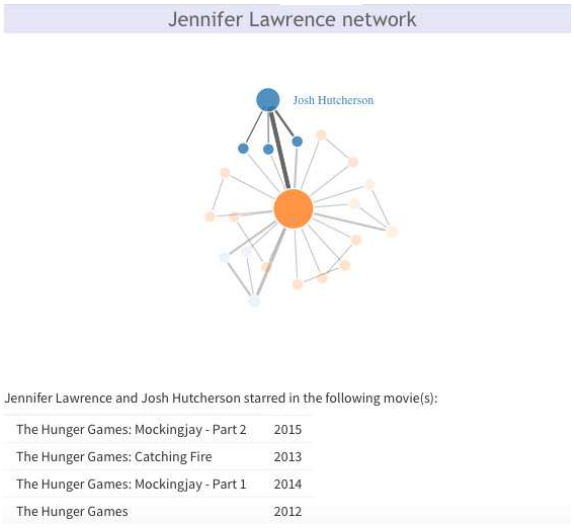

Figure 28. Click and hover interaction of Jennifer Lawrence network

5.1.4. Sub-task 1.4. Compare actor time slice

Centrality, time slice and temporal dynamic analysis and Node-Link diagram visualisation are used for this subtask. Time slice analysis is used to extract a sub-graph from the complete network [5] and to show in which periods the actor was starring in movies. There are up to four periods for the actor time slice, pre-1989, 1990-1999, 2000-2009 and 2010-2016. Colour on the links shows if the actors starred in the same movie in the same periods. Node size is based on degree centrality. The nodes positions are unchanged.

Brad Pitt's acting career spans 23 years since 1993 (see Figure 22) and there are diagrams for three periods. The biggest node is for Brad Pitt himself. The other nodes are for all the co-stars in his career, with the size representing the number of movies in which they co-starred with Brad Pitt. All of his co-stars are in every diagram in the same position. The links in each diagram indicate which actors he starred with in the corresponding period (see Figure 29).

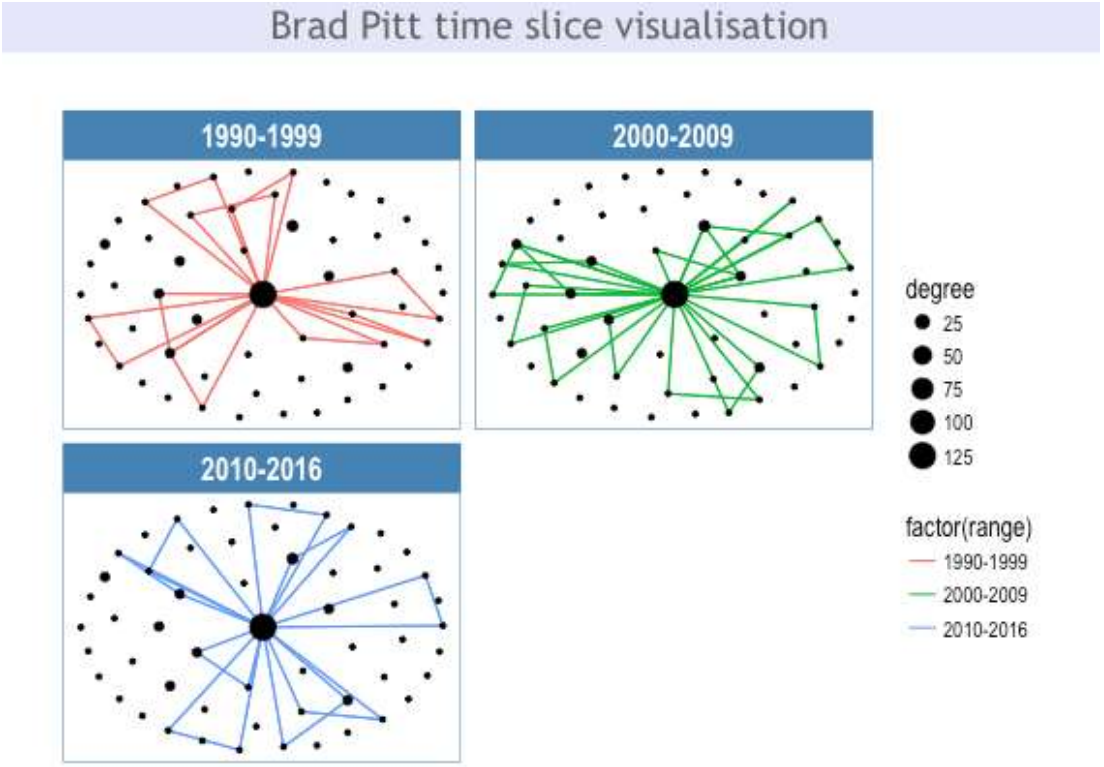

Figure 29. Brad Pitt time slice

Jennifer Lawrence's acting career spans for 7 years since 2010 (see Figure 23) and there is a diagram for one period. The biggest node is for Jennifer Lawrence herself. The other nodes are for all the co-stars in her career, with the size representing the number of movies in which they co-starred with Jennifer Lawrence. All of her co-stars are in every diagram in the same position. The links in each diagram indicate which actors she starred with in the corresponding period (see Figure 30).

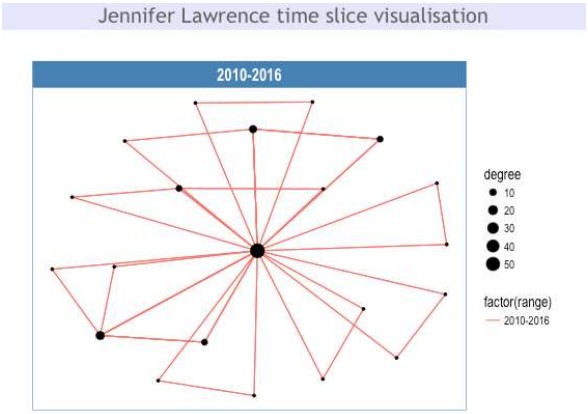

Figure 30. Jennifer Lawrence time slice

5.1.5. Result summary of the most influential actors comparison

- Brad Pitt has an acting career of 23 years (see Figure 22). Jennifer Lawrence has an acting career of 7 years (see Figure 23).
- Brad Pitt collaborated with many different directors and actors (see Figure 17 and 25), Jennifer Lawrence mostly collaborated with the same actors and directors repeatedly (see Figure 19 and 27).
- Brad Pitt starred in a wide range of genres over the years (see Figure 22). Jennifer Lawrence starred in mostly drama/sci-fi/adventure genres (see Figure 23).
- Brad Pitt has a higher IMDB score than Jennifer Lawrence but Jennifer Lawrence has a higher top grossing movies that Brad Pitt (see Table 2).

5.2. TASK 2 – Identify movie genre trends

This task is to identify movie popularity and top grossing by genre over time. Some genres become more popular while others are less popular. The reasons for the genre shift can be due to a combination of factors including world events [6].

This task uses centrality analysis to determine genre node size and Node-Link diagram for Genre Network Visualisation. Temporal dynamic analysis was used to generate a Streamgraph layout to show Genre Temporal Trends. A Word Cloud layout shows Genre Frequency. The interaction in Filters is selection and interactive slider.

There are two subtasks for this task. First subtask is to compare movie genre trends for different periods. Second subtask is to compare different genres for the same period (post-1980). The details of the comparisons and their summary results are shown below.

5.2.1. Subtask 2.1. Compare movie genre trends for different periods

The first subtask is to filter movie genres for different periods. The selected periods for this subtask are pre-1980 and post-1980.

The pre-1980 genre network shows the largest node and the most grossing genre is *Drama* due to the multi award winning movies “The Sound of Music” and “Gone with the Wind”. There are 18 genres in the top 20 grossing movies (see Figure 26).

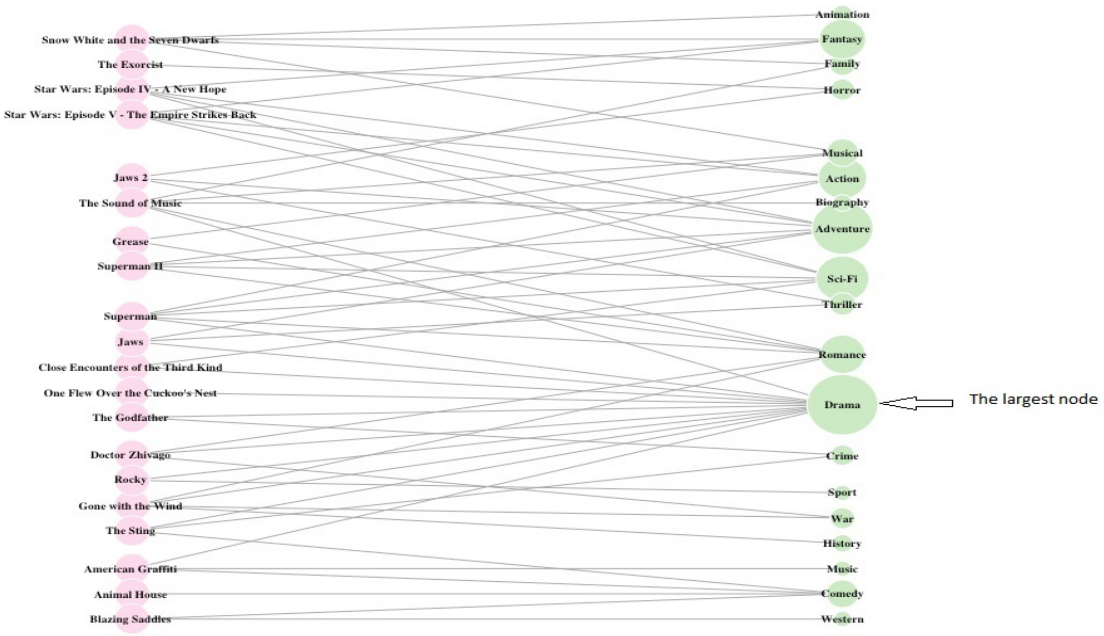

Figure 26. Pre-1980 Genre network

Genre frequency shows the largest size of text and the most popular genre is Drama (see Figure 27).

Figure 27. Pre-1980 Genre Frequency

Genre temporal trends shows that Thriller started to gain popularity in 1960. Sci-fi started to gain popularity in the mid-1970. More genres were introduced in the late 1970s (see Figure 28).

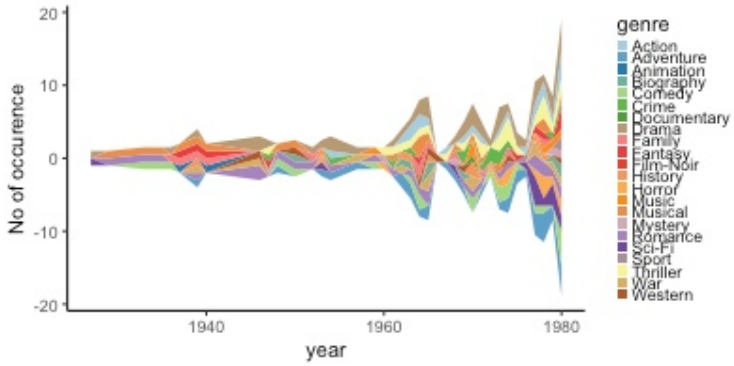

Figure 28. Pre-1980 Genre Temporal Trends

The post-1980 genre network shows the largest node and the most grossing genre is *Adventure* as superheroes movies with special effects gaining popularity such as “The Dark Knight”, “Iron Man” and “Spider Man”. There are 12 genres in the top 20 grossing movies (see Figure 29).

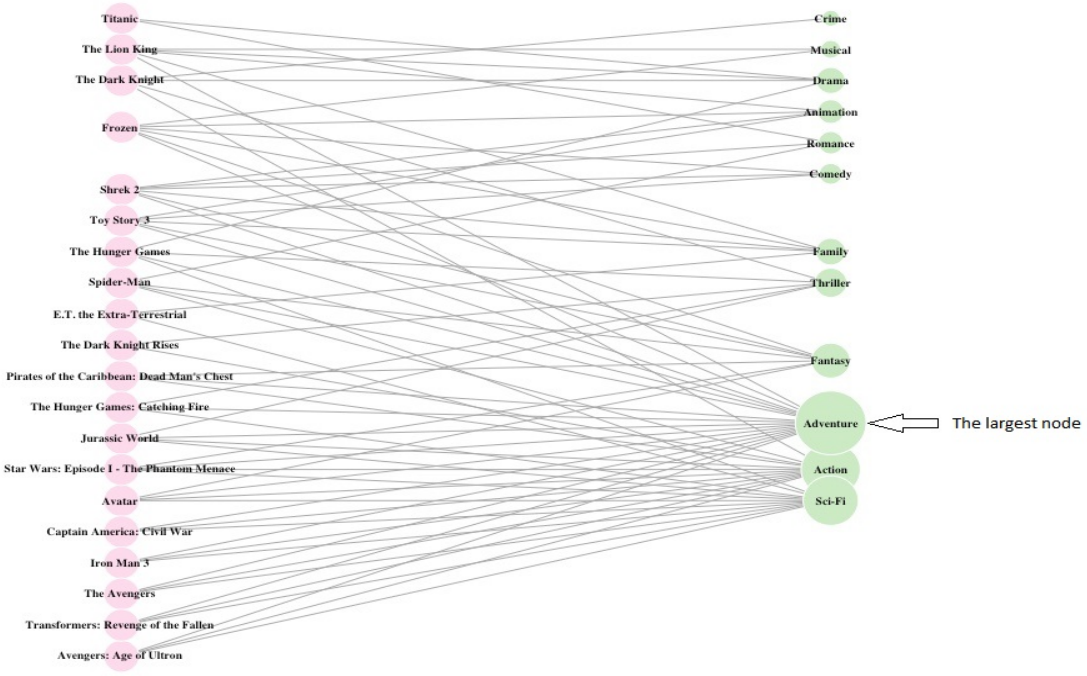

Figure 29. Post-1980 Genre network

Genre frequency shows the largest size of text and the most popular genre is Drama (see Figure 30).

Figure 30. Post-1980 Genre Frequency

The genre temporal trends diagram shows that Drama, Comedy, Thriller, Action and Adventure started to gain popularity in 1990 and stabilise around early 2000 (see Figure 31).

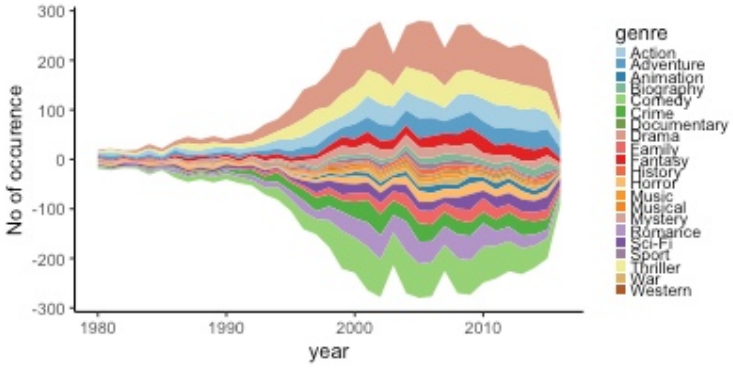

Figure 31. Post-1980 Genre Temporal Trends

5.2.2. Subtask 2.2. Compare different genres for the same period (post-1980)

In this subtask, the focus of the comparison is based on individual genre trends in the post-1980 period. The genres selected for the comparison are Drama, the most popular genre of all time and Adventure, the top grossing genre post-1980.

Figure 32 shows Drama trends. After the successful movie "Titanic" in 1997, there was a sharp upward trend immediately after the movie and it remains popular.

Figure 33 shows Adventure trends. After the biggest success of movie Avatar in 2009, there was a continuing upward trend.

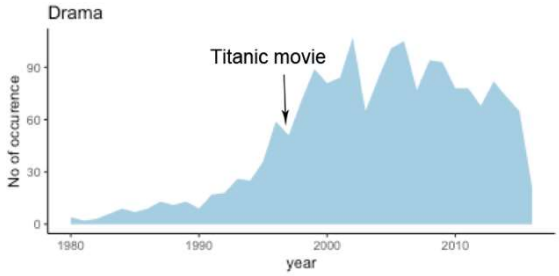

Figure 32. Drama trends

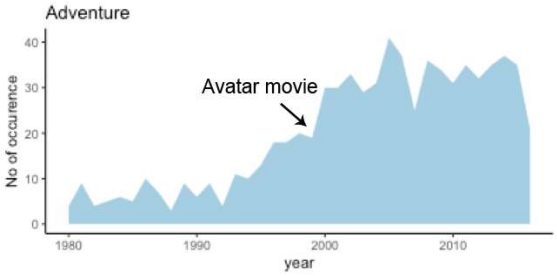

Figure 33. Adventure trends

5.2.3. Result summary of the movie genre trends comparison

The top grossing movie genre pre-1980 is Drama (see Figure 27). The top grossing movie genres post-1980 is Adventure (see Figure 28). However the most popular genre of all time is still Drama. Drama is still consistently popular (see Figure 14). Adventure's popularity has increased over the years (see Figure 30).

6. CONCLUSION AND FUTURE WORK

The massive amount of data today will only increase and effective analysis of the complex data to gain valuable insights is often challenging.

6.1. Summary

In summary, the IMDB5000 dataset from Kaggle is used for this project. In this project, an interactive application on scalable visual analytics for movie data was developed to address this challenge. This application has three development phases: design, implementation and evaluation. The design phase of this application integrates analysis methods with interactive and visualisation techniques to gain insights into movie data. This application was implemented with the R Shiny interactive web application to support scalable visual analytics. We performed two major evaluation tasks using this application. The tasks are to determine the most influential actors and to identify movie genre trends. The methods that our application provides to solve complex, multivariate, dynamic temporal movie data turned out to be fast and flexible.

6.2. Limitation

There are many data visualisation tools available; however, there is no tool that can perform all the tasks for the data visualisation effectively. In the online version of this web application, the Streamgraph layout is not displaying due to compatibility issues. Users' guidance on how use the application is required.

6.3. Future Work

There are many interesting future challenges that I would like to pursue. I am interested in combining advanced data analytics techniques such as machine learning and deep learning with different visualisation tools. I will continue to improve the online version of this application (https://visim/shinnyapps.io/siva/) to make it more interactive and user-friendly and to resolve compatibility issues.

7. REFERENCES

[1] Wikipedia on IMDB <https://en.wikipedia.org/wiki/IMDb>

[2] IMDB Press Room <http://www.imdb.com/pressroom/>

[3] Kaggle IMDB5000 Movie Dataset <https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset/kernels?language=RA>

[4]Joshua Johnson, Key Player & Community detection using Graph Theory on IMDB Movie Ratings, 2016 <https://www.linkedin.com/pulse/key-player-community-detection-using-graph-theory-imdb-joshua-johnson>

[5] S.-H. Hong, A. Ahmed, D. Merrick, A. Mrvar, V. Batagelj, X. Fu, Visualisation and analysis of the internet movie database (2007)

[6] J Dauenhauer, J Hockett, J Mammarelli, M Yarem, Information Analysis of Movie Genres (2015)

[7] The numbers for movies financial analysis <http://www.the-numbers.com/>

[8] SPP – Scalable Visual Analytics <http://www.visualanalytics.de/node/2>

[9] Interactive and Dynamic Network Visualisation in R <http://curleylab.psych.columbia.edu/netviz/netviz1.html>

[10] Katya Ognyanova – R tutorial <http://kateto.net/network-visualization>

[11] Data Visualisation Catalogue<https://datavizcatalogue.com/index.html>

[12] Alex Endert, Semantic Interaction for Visual Analytics: Inferring Analytical Reasoning for Model Steering (2016)

[13] Louisiana State University, Network: Basic Concepts <https://www.lsu.edu/faculty/bratton/networks/closeness.ppt>

[14] University of Cambridge, Social and Technological Network Analysis, Lecture 3: Centrality Measures <https://www.cl.cam.ac.uk/teaching/1314/L109/stna-lecture3.pdf>

[15] Wikipedia - Graphical user Interface <https://en.wikipedia.org/wiki/Graphical_user_interface>

[16] Methos in Ecology and Evolution <http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2012.00236.x/full>

[17] Zhao Yang, Rene Algesheimer and Claudio J. Tessone, A comparative Analysis of Community Detection Algorithms on Artificial Networks (2016)

[18] Clauset, A., Newman, M. E. & Moore, C. Finding community structure in very large networks. *Physical Review E***70**, 066111 (2004)

[19] Pons, P. & Latapy, M. Computing communities in large networks using random walks. In *Computer and Information Sciences-ISCIS 2005*, 284–293 (Springer, 2005)

[20] Girvan, M. & Newman, M. E. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences***99**, 7821–7826 (2002)

[21] Freeman, L. C. Centrality in social networks conceptual clarification. *Social Networks***1**, 215–239 (1979)

[22] Newman, M. E. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E* **74**, 036104 (2006)

[23] Xie, J. & Szymanski, B. K. Community detection using a neighborhood strength driven label propagation algorithm. In *Network Science Workshop* 188–195 (IEEE, 2011)

[24] What is visualisation? <http://libguides.library.curtin.edu.au/c.php?g=388681&p=2688784>

8. APPENDIX

8.1. Appendix A: How to run the application

- Files structure for the R Shiny applicationis as follows:
  1. ui.R
  2. server.R
  3. Global.R
  4. data (folder)
     a. movie_metadata.csv (IMDB5000 dataset)
  5. www (folder)
     a. dataset-thumbnail.png
     b. imdb.png
- In RStudio, open ui.R and server.R files, install all the libraries and click on Run App to run the application.

8.2. Appendix B: How to use the application

The application has three menu options, Overview, Actors and Movies.

- **Overview** is the basic analysis of IMDB5000 dataset.
  - Interaction:
    Select one of the attributes from the drop down list.
  - The visualisations are histograms, a line graph and Wordcloud layouts.
- **Actors** is for determining the most influential actors in task 1
  - Interaction:
    - Filter selections are genres, IMDB Ratings, gross, years, centrality and clustering algorithms. Genre, centrality and clustering algorithm options are in a drop-down list and IMDB ratings, gross and years range are in a slider control.
    - In co-starring network diagram, user can zoom, drag and hover over nodes.
    - In actor and director collaboration, user can hover over the ring layers to reveal the name of actor, director, movie and year.
    - In actor network, user can zoom, drag, hover and click on nodes. Hover on a co-actor to show the name, click on a co-actor to show the movies which they starredtogether.
  - The visualisations are network diagrams, a Sunburst diagram and a stacked bar plot.
- **Movies** is for identifying movie genre trends in task 2
  - Interaction:
    Filter selections are years and genres. Genre options are in a drop down list and years range is in a slider control.
  - The visualisations are two-mode network diagram, a Streamgraph layout and a Wordcloud layout.