Problem Chosen
D

2021
ShuWei Cup
Summary Sheet

Team Control Number
202109256115

Box office analysis model for the Chinese market based on time forecasting

Summary

With the demand of people's cultural consumption, China's film industry continues to show prosperity. At the same time, there are many external factors affecting the development of China's film industry. This paper analyses these factors.

In response to question 1: The K-means clustering algorithm was used to classify the data of the set samples. Firstly, given the number of clusters and the data set as the input quantity, then k initial clustering centers were selected. By analyzing the concentrated data, this paper finally selected a clustering analysis model with $k=3$ to classify the movie data into three major categories. The results show that: the first category is IMAX romance and IMAX science fiction films released by ordinary Chinese production companies in the summer season, with low movie box office; the second category is 2D drama and 2D comedy films released by small Chinese production companies in the ordinary season, with average movie box office; the third category is IMAX comedy films and IMAX action films released by famous Chinese production companies in the ordinary season, with high movie box office. In order to verify the accuracy of the model, the APriori algorithm was used for correlation analysis as well as variable correlation analysis, and the results showed that the classification model was highly accurate and practical.

In response to question 2: Our group developed a time smoothed series forecasting model. The data was first integrated with the results already analysed in Topic 1, then substituted into the time series forecasting model for forecasting analysis, and the data was tested against the different category models for comparison with realistic analysis. The results show that the projected box office forecasts for each category show a year-on-year increase in relation to the overall box office forecasts.

In response to question 3: This paper uses a Python crawler data mining algorithm. Firstly, the specified content is found in the developer tools, and the content of the json file of the source URL is crawled and converted into a CSV file based on third-party libraries such as request and regular expressions in Python. A mining algorithm for microblogging data was established to analyse the correlation between online public opinion and box office, and a Bi-LTSM model was built to further analyse valuable in-depth data for the specific method of identifying film ratings by online water forces. The results show that public opinion largely influences whether or not a film viewer watches a film, and therefore public opinion also influences the box office of a film..

In response to question 4: In this paper, we propose a time series forecasting method that takes full account of the duration of the epidemic, the scale of the impact and the extent of the impact, and we develop a predictive analysis model based on a time-smoothed series. We have followed the model in Problem 3 for this problem, data crawled a large amount of data and finally processed it by a linear weighting method to analyse the realistic impact of different attendance requirements on movie box office after the epidemic has stabilised and the impact of future trends. The results show that at 30% theatre attendance greatly affects movie box office; at 50% or 75% theatre attendance also affects movie box office but not to a significant extent.

Key word: K-means clustering ; APriori algorithm ; Time series forecasting models ; Python crawler ; Bi-LTSM model

Content

1. Introduction

1.1 Background

The increase in people's cultural consumption demand has also driven the development of China's film industry. At the same time, box office not only directly reflects the economic value a film creates for the investing company, but also reflects the artistic quality and business strategy of the film from the side. It is an important indicator of a film's success. It naturally reflects the degree of market demand and investment appeal of a film production. If one can predict in advance the degree of acceptance and profitability of a film product in the market, it will have a huge impact on the decisions made by various links in the film industry chain. Accurate forecasting of film box office is therefore particularly important for risk control and decision-making at this time.

However, there are many factors that affect the box office of a film, such as the quality of the film itself, its running time, advertising, the social environment, the number of cinemas in which the film is shown, and even the weather at the time of the screening. Among these factors, the evaluation of films by online public opinion has a great impact on box office, and its participants are diverse and complex. Therefore, identifying and managing the influence of online water forces on film ratings is a pressing issue in the development of China's film industry. At the same time, unexpected events can also have a huge impact on box office, such as the sudden emergence of a new coronavirus in 2020 that almost destroyed the openly assembled film market. It is therefore clearly important to use models to analyse the impact of various unexpected events on film box office.

1.2 Work

（1). The core of early prediction of movie box office is to select effective predictive features. The factors that affect the box office of a movie are complex and vary in measurement methods. Features include: movie duration, actors, director, movie type, movie format (2D, 3D, IMAX), whether the movie is a sequel, release date, production company, and distribution Company and so on. According to the characteristics of movie classification, consider the characteristics of movie classification, movie type, director and other classification characteristics, director rating and other classification characteristics, cluster and classify the movies in the provided data set, and verify the effectiveness of the classification.

(2). Common box office prediction models include multiple regression, neural network, etc., and some scholars predict the box office by studying audience word-of-mouth communication, network information communication and network search in social networks (see references), and establish a positive influence on movie box office Movie box office prediction model, for the box office of the movie market, a classification model (result of title 1) is given based on the provided data, and the estimated box office forecast and the overall box office forecast for each category are given in advance.

(3). Collect online public opinion evaluation data about movies from platforms such as Douban and Maoyan, and establish an algorithm to identify the positive and negative scores of online public opinion (standardized to [-1,1]); establish a model to extract topic words, topic classification or Other important indicators; establish a model to analyze the correlation between online public opinion and the box office and the degree of influence on the box office; design ideas and specific methods to identify the movie scoring network navy based on the problem and the current situation. The method needs to be logically self-consistent and feasible.

(4). In response to the sudden outbreak of the new corona-virus, the state's guidelines for the prevention and control of epidemics in cinemas opening up: The attendance rate of each venue shall not exceed 30%, 50%, 75%, etc., considering the impact of the epidemic on the model, and analyzing its impact Realistic influence and future prediction of movie box office. Using the data provided, the model analyzes the impact of different attendance requirements (30%, 50%, 75%) on movie box office forecasts after the epidemic has stabilized.

2. Problem analysis

2.1 Analysis of question one

With the demand of people's cultural consumption, China's movie industry keeps showing prosperous scenes. In this paper, we mainly analyze by K-means clustering algorithm, establish time smoothed series prediction model and Bi-LTSM model to solve the related movie industry problems.

2.2 Analysis of question two

Problem two requires us to model the results of problem one and give a per-category box office forecast and an overall box office forecast. For this problem, the box office highs and lows can be processed first, then the box office is organized based on the three categories derived from Problem 1 for recent years, and then the data is predicted for future box office using a time series prediction model.

2.3 Analysis of question Three

Question 3 asks us to build a model to analyze the correlation between online public opinion and box office and the degree of influence, and to design a method to identify online water forces. The data of a platform can be crawled first, and then a suitable model can be built to observe the relationship.

2.4 Analysis of question Four

Question 4 asks us to consider the extent to which the epidemic has hit the movie industry based on the attendance rates issued by the state. This question can be discussed with respect to the movie box office in recent years and the movie box office data after the epidemic outbreak, while a time series prediction model can be used to make box office forecasts for the future, from which the impact of the epidemic on reality can be observed. After that, we can forecast and analyze the attendance requirements for different periods, and finally make a comprehensive judgment on the impact of different attendance rates on movie box office.

3. Symbol and Assumptions

3. 1 Symbol Description

No.	Symbol	Symbol meaning
1	x	Coordinates of a point
2	y	Coordinates of the other point
3	x_i	i variable value of point x
4	y_i	i variable value of point y
5	c	Level of impact
6	a	2020 Forecast Box Office
7	b	Actual box office in 2020

3.2 Fundamental assumptions

Hypothesis 1: The data of some media or netizen comments are captured to be true and reliable.

Hypothesis 2: The development of public opinion is sometimes also influenced by fan back-up such as controlled comments, which is not considered in this paper.

Hypothesis 3: It is assumed that each comment can truly reflect the inner thoughts of netizens and there is no case that the comment is inconsistent with the inner thoughts.

There is no case of inconsistency between comments and inner thoughts.

Hypothesis 4: Assume that the epidemic situation is uniform throughout the country. When opening 30%, 50% and 75%, the national yard is uniform.

4. Problem solving

4.1 Solution to question one

4.1.1 Pre-processing of data

In this paper, the movie box office data from 2013 to 2019 was selected from the website of Yi en Entertainment Numbers (https://ys.endata.cn/DataMarket/BoxOffice), which has more authentic and reliable data, provides more detailed movie-related information and ensures the integrity of the data, so the data from this website was selected for analysis. The target variable for prediction in this paper is the movie box office, and the dependent variable for prediction is the six influencing factors of the movie box office.

The data obtained from the crawler was imported into EXCEL, the rows with blank cells were deleted, and the remaining data were rearranged.

Among the important factors affecting the box office of a film are the following Figure 1.

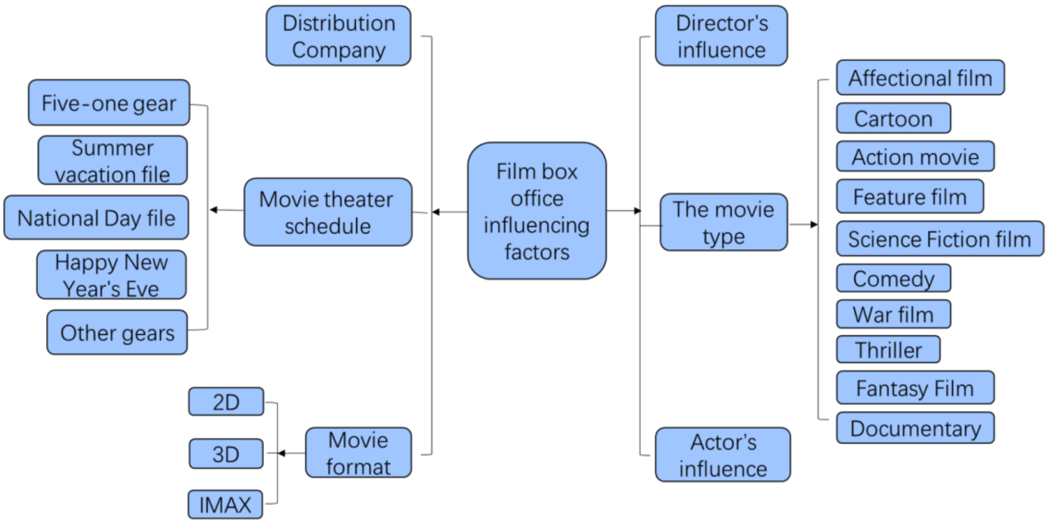

Figure 1 Factors affecting the box office of a film

(1) Distribution companies.

The film and television investment industry is an industry with a high perception threshold, low transparency, low controllability and high risk, so the top-ranked film distribution companies must have professional production teams, advanced technology and strong capital accumulation in order to gain a foothold in this industry. Some of China's best known film companies include China Film Group Corporation, Wanda Film and Light Media. The Chinese film industry is on the rise, so there are many film companies, and in addition to the more well-known ones, there are also many small production companies. Due to their lower profile and professionalism, their films will, in general, also have a lower box office.

(2) Film genres.

Different types of films have different audience groups, and different groups have different consumption levels. For example, animated films and science fiction films are niche favourites, while the audience for animated films are mostly children with low spending power, and therefore have a significant impact on the box office. This article discusses the ten categories of film genres, namely romance, animation, action, drama, science fiction, comedy, war, thriller, fantasy and documentary.

(3) Time of year.

People's demand for cinema varies at different times of the year, for example, there are far more people watching films on holidays than on non-holidays, and far more people watching films during the summer holidays than during regular periods. Here we use the four quarters of the year as a basis for classification. The first quarter is March-May, the second quarter is June-August, the third quarter is September-November and the fourth quarter is December-February. The first quarter includes the May Day slot, the second quarter includes the summer holiday slot, the third quarter includes the National Day slot and the fourth quarter includes the New Year's Eve slot.

(4) Film format.

Films with stereoscopic animation effects and giant screens can be more immersive than ordinary films and will appeal to a wider audience due to their good viewing experience. In this paper, film formats are divided into 2D, 3D and IMAX.

(5) Director influence.

If a director is a well-known director, then his or her work will also be highly anticipated by the public and will have a positive effect on the box office of the film. If a director is not known to the public, his or her attention will be greatly reduced unless his or her films are produced extremely well before being accepted by the masses. In this paper, the director's influence is divided into five categories: low, low, average, high and high.

(6) Lead Acting Influence. The influence of the lead actor is very similar to that of the director. The lead actor is an integral part of a film's performance and box office appeal. If the lead actor has a high level of performance power and popularity, his or her film is highly anticipated by audiences. In this paper, the top three lead actors of a film are selected for analysis, and their three previous films are summed to rank the results in terms of box office sales, and certain criteria are selected to find the level of influence of the lead actor. In this paper, the influence of the lead actor is divided into four categories: poor, moderate, good and excellent.

By compiling data related to movies released from 2013 to 2019, the total box office amount of movies released in each quarter from 2013 to 2019 is obtained in Figure 2 below. The analysis revealed that the third quarter of each year is the peak period for film revenue, and it was also found that the third quarter contains the summer holiday season, when many blockbuster films choose to be released in order to gain higher attention and box office revenue. At the same time, the first quarter of the year has fewer holiday slots and lower movie box office. The box office fluctuates significantly from quarter to quarter, indicating that people have different needs for cultural consumption at different times of the year[1].

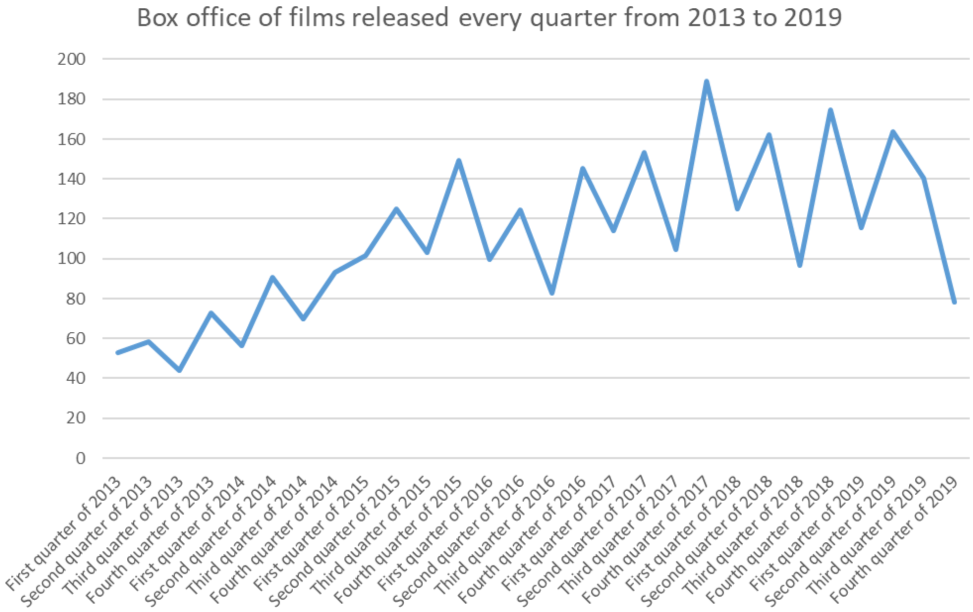

Figure 2 Total box office receipts of films released by quarter from 2013 to 2019

Also due to the different preferences of people for different film formats, the data was analysed to produce Figure 3 below. This indicates that two extremes are coexisting in China: firstly, people with high incomes are more willing to maintain their movie-going experience and spend more money in order to have a better movie-going experience; secondly, people with low incomes view movies in 2D, and the general rule of thumb for this mode of viewing is that the price is low and the price is high.

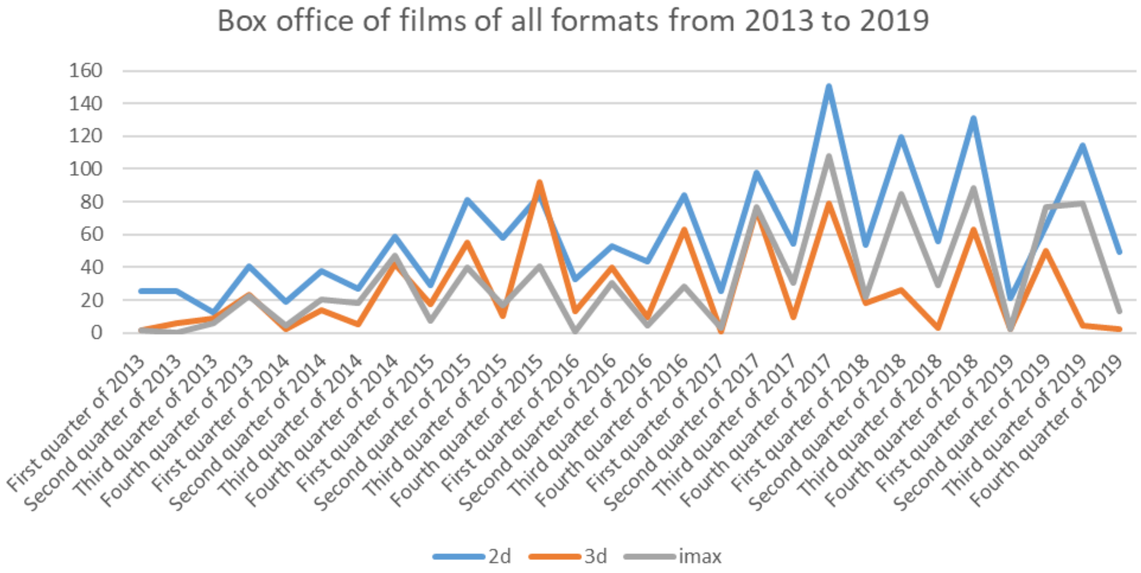

Figure 3 Total box office receipts of films released in different formats by quarter from 2013 to 2019

Finally, a further statistical analysis of the total box office generated by the film genre by quarter in each year gives the following Figure 4.

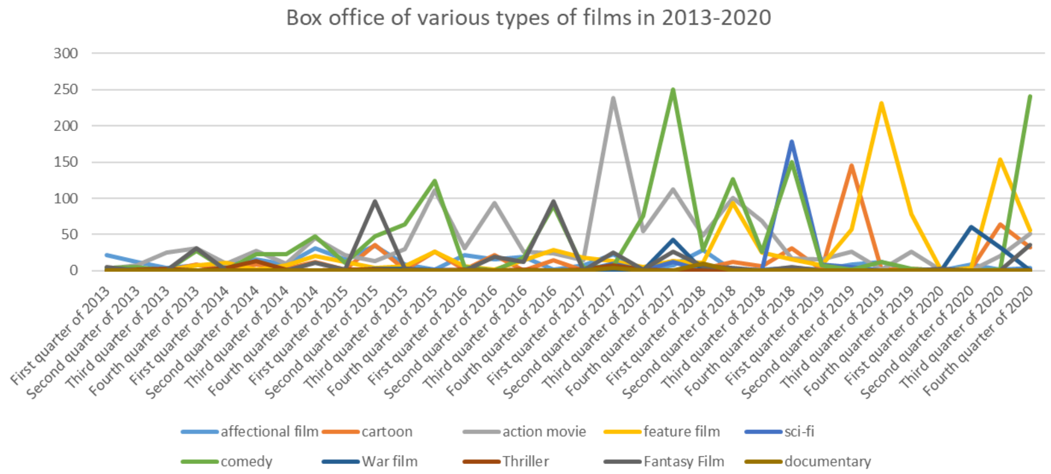

Figure 4 Forecast total box office receipts by genre for films released in cinemas by quarter in 2020

4.1.2 Cluster analysis model building and solving

Clustering organises all data instances into groups of similarities, which are called clusters. The data in the same class are very similar to each other and the data in different classes have a high degree of variability. Clustering is one of the more commonly used tools for data analysis. Cluster analysis can be used during data pre-processing, and for complex structured multidimensional data can be aggregated by means of cluster analysis to standardise complex structured data. It can be used as a stand-alone tool to discover data distribution patterns and implicit information, or as a pre-processing step in other complex algorithms [1].

In this topic we use the K-means clustering algorithm in cluster analysis, which allows the classification of samples using information from multiple variables, and the results are clear and intuitive and more comprehensive than traditional classification methods.

The steps of the K-means clustering algorithm are as follows.

(1) Specify the number of clusters K. In K-means clustering, the number of classes to be clustered is given first, and the number of clusters is determined by considering both the actual needs of the problem and the final clustering effect.

(2) Determine K initial class centroids. Selecting the appropriate initial cluster centres is the key to the K-means clustering algorithm and affects the final clustering result.

(3) The Euclidean distance, or Euclidean distance, is calculated for each sample point to the K initial class centres in turn, and all sample points are assigned to the class with the smallest distance according to the closest principle, forming the K classes. Where the Euclidean distance in dimensional space is given by:

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

In the above equation x_i is the i variable value for point x and y_i is the i variable value for point y.

(4) The K class centroids are redefined, and in general the mean value point of all data point variables in the K classes is used as the centroid of the K classes.

(5) Repeat steps 2, 3 and 4 above until the cluster centroids no longer change or the maximum number of iterations is reached before stopping [2].

Figure 5 below shows the flow chart for classifying sample data using the K-means clustering algorithm:

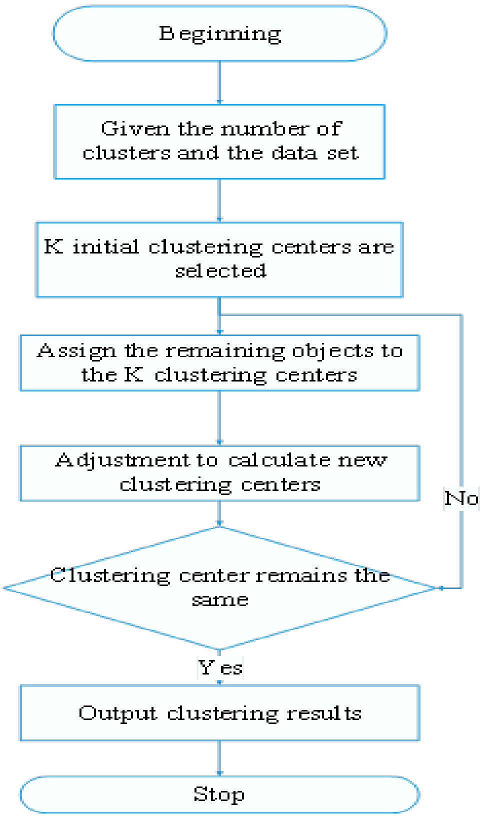
Beginning
Given the number of clusters and the data set
K initial clustering centers are selected
Assign the remaining objects to the K clustering centers
Adjustment to calculate new clustering centers
No
Clustering center remains the same
Yes
Output clustering results
Stop

Figure 5 K-means clustering algorithm sample data classification flow chart

In this paper, K-means clustering was performed using SPSS data analysis software. Through simulation training and optimisation analysis, the final cluster analysis model of $K=3$ was selected to classify the film data into three main categories.

The first category is IMAX romance and IMAX science fiction films released by ordinary Chinese production companies in the summer season, with higher influence of the lead actors, higher influence of the directors and higher competitiveness in the same period, and therefore achieved lower box office.

The second category consists of 2D dramas and 2D comedies released by smaller Chinese production companies in general release slots, with less influential lead actors and more influential directors, and less competitive in the same time period. Although in 2D format, the films are more objective in terms of box office because they are shown more often throughout the year, but still lower than the third category.

The third category is IMAX comedies and IMAX action films released by well-known Chinese production companies in the general slot, where the influence of the lead actor and the influence of the director are both higher, but the competitiveness of the same period is lower, thus leading to a higher box office for these films.

In order to analyse the degree of correlation between the factors influencing the box office, this paper uses the Apriori algorithm for correlation analysis. The association rules are as follows: an association rule is an implication expression shaped as $X \rightarrow Y$, where X and Y are uncorrelated sets of terms, $X \cap Y = \varnothing$. The strength of an association rule can be measured in terms of its support and confidence.

Its support formula is：

$$Support(X,Y)=P(X,Y)=\frac{num(xy)}{num(allsamples)}$$

Its confidence formula is：

$$Cindidence(XY)=P(x\,|\,Y)=\frac{P(xy)}{P(y)}$$

This paper has conducted a correlation analysis of the box office data found, and the results show that if a film has a low lead actor appeal and is released in a normal time slot and in 2D, the film's box office is average in most cases. The reasons for this are likely to be due to the low appeal of the lead actor, the fact that he or she does not have an advantage in the promotion of films released in normal time slots, and the fact that people's consumer demand is low at this stage and romance and science fiction films do not greatly attract people to spend money, but because 2D films have lower ticket prices and more venues, the box office is more average and not too high or too low. If a film's lead actor and director are both less influential, the film is likely to be less popular at the box office. The reason for this may be that the director is not a professional and the lead actor is a mediocre performer, resulting in a low quality film, and thus low box office.

4.1.3 Verifying the validity of cluster analysis

The results of the correlation analysis between the variables using statistical software are shown in Table 1 below.

Table 1 Results of correlation analysis between variables

	a	b	c	d	e	f	g	h	i	j
a	1	.622	.411	.257	.056	.032	-.083	-.035	.361	.284
b	.622	1	.134	.145	.155	-.078	0.000	-.156	.355	.307
c	.411	.134	1	.426	-.161	.031	-.165	.337	.167	-.114
d	.258	.145	.426	1	-.158	.032	-.063	.132	.210	.123
e	.056	.155	-.161	-0.158	1	-.273	-.221	-.247	.156	.216
f	.032	-.078	.031	.032	-.273	1	-.297	-.322	.020	.075
g	-.083	.000	-.165	-.063	-.221	-.297	1	-.269	-.137	-.152
h	-.035	-.156	.337	.132	-.247	-.332	-.269	1	.012	-.298
i	.361	.355	.167	.210	.156	.020	-.137	.012	1	.278
j	.284	.307	-.114	.123	.216	.075	-.152	-.298	.208	1
k	.361	.355	.167	.171	.035	.210	.156	.020	-.137	.278

Where a indicates total box office， b indicates First day box office， c indicates Score，d indicates The length of the film，e indicates Comedy，f indicates Action movie，g indicates Affectional film， h indicates Feature film， i indicates Sci-fi， j indicates Schedule， k indicates The creation of brand degree.

At the level of $P < 0.01$ and Variables that are significantly and positively correlated with total box office are: first day box office, rating, length and slot. At the level of $P < 0.01$ and The variables that are significantly and positively correlated with first day box office are: total box office and date. It is clear that comedies and action films are strongly correlated with first day box office, but not with total box office. Ratings, on the other hand, are not correlated with first day box office, but are correlated with total box office. At the same time it was found that No significant correlation between total box office and film genre （$P < 0.05$）， Comedies, action films and first-day box office are positively and weakly correlated, while romances, science fiction films and first-day box office are negatively and weakly correlated. This indicates that film genre is related to first day box office, but not to total box office, and that all types of films may have high box office[3].

By comparing the results of the cluster analysis with the results of the correlation analysis between the variables, it was found that the results did not differ significantly, so the results of the cluster analysis data can be proved to be reliable.

4.2 Solution to question two

4.2.1 Pre-processing of data

As can be seen from question one, our group has grouped romance and science fiction films into the first category, drama and comedy films into the second category and comedy and action films into the third category. Since animated films, thrillers, war films and science fiction films are less watched, the number of films is lower and the box office is correspondingly lower, so we will not discuss them in this topic and we will eliminate them through Excel. (as long as a film has one of these genres, it can be divided into each of the three categories), for example, if a film is in the genre of romance and drama, we will record the film once in the first category and once in the second category.

In order to ensure that the data is comprehensive and that it can be used as valid data for predicting box office research, the data will be counted and analysed in a number of ways.

The first study was the number of films distributed at the box office, as shown in Table 2 below.

Table 2 Breakdown of the box office

Box office number	Box office type		
	First day box office	First week box office	Total box office
<50(million yuan)	101	40	11
50<100(million yuan)	88	28	22
100<200(million yuan)	85	72	23
200<400(million yuan)	53	91	88
400<600(million yuan)	37	51	55
600<1000(million yuan)	50	34	64
>1000（million yuan）	194	292	354
Total	608	608	608

Subsequently, a separate proportional count of the film formats was conducted and it was found that 3D films accounted for approximately 82% of the 608 films, with ticket prices for 3D formats being higher than those for 2D films, making 3D films more likely to achieve higher box office than 2D films.

4.2.2 Development and solution of a time-smoothed series forecasting model

A time series smoothing model is a basic model that uses time series smoothing to construct a time series. A time series is a sequence of quantitative changes that occur in a phenomenon, arranged in chronological order, to reveal the pattern of development of the phenomenon over time, in order to predict the direction of the phenomenon and its quantity.

Our group used MATLAB to make box office forecasts for the first, second and third categories over the next three years for the more representative seventy of the above data, and obtained the results shown in Table 3 below.

Table 3 Box office forecast results for the next three years for three categories of films

	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
A	25.39	52.04	115.01	97.35	103.81	223.91	211.79	241.3622	269.6753	297.9884
B	59.34	64.6	148.1	148.71	178.73	249.29	95.12	156.9238	160.1258	163.3278
C	34.56269	47.573	44.56	38.09	20.06	19.21	60.79	48.35435	51.49278	54.63121

Define 2D Feature film and Comedy as A, IMAX Comedy and Action movie as B, IMAX Affectional movies and Sci-fi as C.

By analysing Table 3, it was found that the predicted box office for each category showed a year-on-year increase in the next three years, and the increase was large.

To make the forecast results more reliable, we again analysed the data from 2007 to 2019 to obtain the total box office trend and box office growth rate, and then analysed the forecast box office for the next three years, as shown in Figure 6 below.

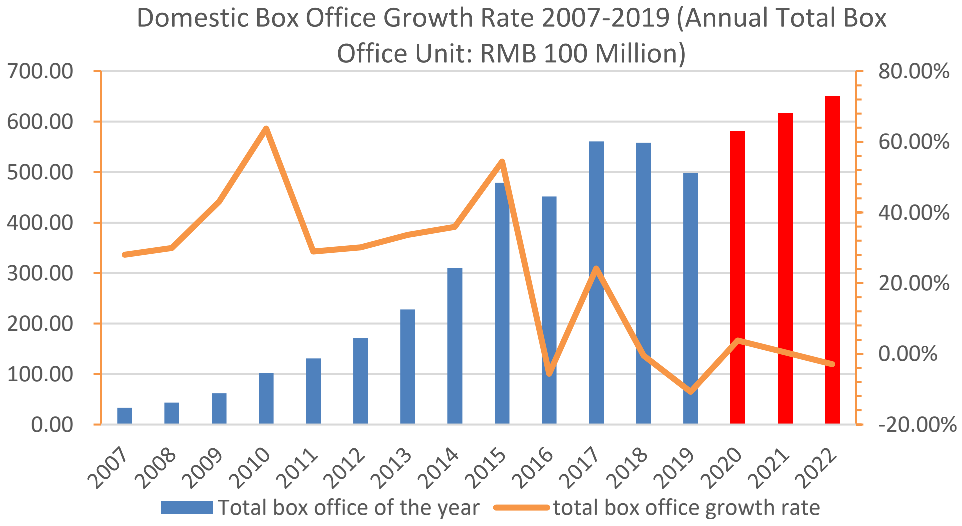

Figure 6 Forecast results for total box office for the next three years

By analysing the above graphs, it can be seen that the total box office is generally growing, but the growth rate of box office is fluctuating significantly and is becoming less and less volatile, possibly because the film industry is becoming saturated in recent years.

4.3 Solution to question three

4.3.1 Python crawler

In this paper we will model the crawling of opinion data based on crawler technology. Web crawlers are a very critical component of search engines, not only through the process of mobile extraction of internet specific page content, but also with the help of search engine web crawler work programs, which are extremely convenient for web data information acquisition benefits. Crawlers are automatic downloading programs that can quickly crawl target data information through user needs and thus perform selective web access.

The language of crawling technology is simple and straightforward, and the process is relatively simple and easy to use, so the process of writing a Python-based crawler will take less time and effort.

Python crawlers are mainly used for data acquisition process analysis and web image acquisition analysis. In the case of data acquisition process analysis, the system uses a web crawler based on Python to log in and crawl web pages such as Douban and Weibo for relevant data information, as well as querying keywords to find the corresponding information and storing this dynamic information in local files.

Crawlers can be divided into generic crawlers and focused crawlers.

One of the generic web crawlers is used to collect web pages and gather information from the Internet. This web page information is used to index and thus support the search engine, which determines whether the entire engine system contains rich content and timely information, and therefore the degree of its performance directly affects the effectiveness of the search engine. It works in four main steps: crawling web pages, data storage, data pre-processing, providing search services and web ranking. However, it also has certain limitations: the search engine cannot provide search results specific to a particular user; it cannot do much to discover and access data such as images, databases and audio; and it is difficult to accurately understand the specific requirements of the user based on the queries provided by semantic information.

A focused web crawler is a web crawler that selectively crawls pages that are relevant to a pre-defined topic.

Compared to general purpose web crawlers, focused web crawlers try to ensure that they only crawl the web information that is relevant to their needs and crawl a smaller area.

Figure 7 below shows the focus crawl process:

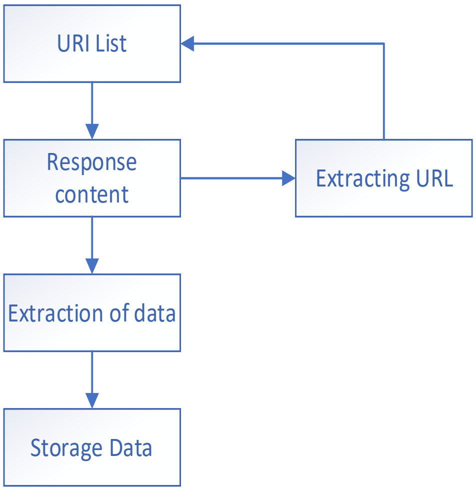

Figure 7 Spotlight on the crawler workflow

The next step in the data crawling process was carried out by our team for Weibo. Firstly, we logged into the Weibo interface, added "#" before and after the topic name, then searched for the required content, opened the browser to review the elements and captured the XHR formatted URL in the web interface as the source address for the crawled data. The content of the json file at that source address is then parsed. The data is stored under data, with cards representing tags, and there are 10 files per page, each with mblog as the detailed topic content of the comment, mblog.user.id as the user ID, mblog.user.name as the user nickname, mblog.reposts_count as the number of reposts on the topic, and mblog. comments_count is the number of comments on the topic, mblog.attitudes_count is the number of likes, mblog.text is the topic text, mblog.created_at is the time of publication, and mlbog.source is the source device [4].

This is followed by an analysis of the impact of online film public opinion on Chinese films, which can be viewed from four aspects: analysis of the box office of low and high scoring Chinese films, analysis of the continued influence of low and high scoring Chinese films, media and netizen attitudes towards Chinese films, and professional critics' concerns about Chinese films.

(1)Analysis of the box office of low and high scoring Chinese films: Analysis of the data within Question 1 reveals that, assuming a film is rated out of 10, we will find that films with a score of 5 or below are low scoring films, and their box office is relatively low; while for films with a score of 7 or above, their box office is high, and there is a positive relationship between high and low scores and box office. There is a huge difference between films with a high rating and those with a low rating.

(2) Analysis of the sustained influence of low- and high-rated Chinese films: Through research and analysis, it is found that the influence of low-rated Chinese films is much less than that of high-rated Chinese films, and the higher the rating, the higher the sustained influence.

(3) Media and netizens' attitudes towards Chinese films: The media generally maintain a neutral attitude towards films out of their own interests, and their articles with positive and negative sentiments are relatively rare, and most media will not issue articles with strong sentiments for the sake of "excessive eye-candy". Unlike the media, netizens have a different attitude, with most of them expressing strong emotions.

Negative attitudes on the internet were not far behind the positive ones. Take the case of "Birds of a Feather", which was released on 6 May 2016 and grossed a mere $280,000 on its first day, and only $1.54 million a week later. However, the subsequent kneeling incident of Fang Li triggered a publicity explosion, with celebrity vloggers forwarding comments and a large number of media and self-publishing outlets reprinting them. The strong public attention led to a significant increase in attendance the next day, and the box office skyrocketed on 14 May, eventually breaking the 60 million mark. Figure 8 below shows the trend of public opinion and box office revenue for "Birds of a Feather", which clearly shows that public opinion has an overriding effect on box office.

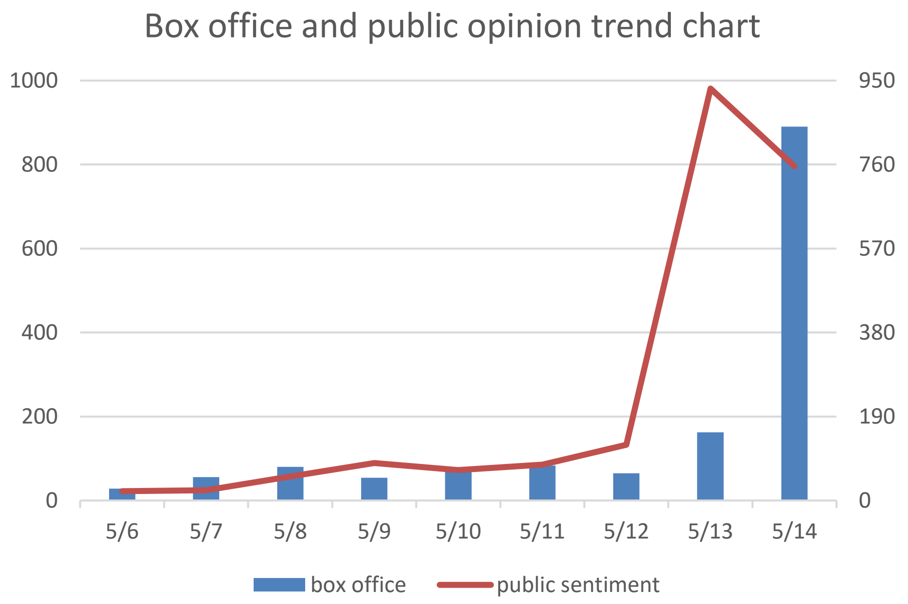

Figure 8 Public opinion chart and movie box office trends

This shows that when Chinese moviegoers make their movie-going choices, external cues such as word-of-mouth and public opinion often influence the audience's choices to some extent, which in turn drives the extent to which box office rises and falls.

（4）Professional critics' attention to Chinese films: Professional critics now have more channels to express their views on a particular film, and there are also an extremely large number of trolling film bloggers and drama pushing bloggers, most of whom have millions of followers, and their reviews of a particular film can greatly influence how much netizens like the film. But accordingly, if a film is hotter at a certain time, these critics will also launch a more friendly review of the film in order to gain heat, even if its production effects are not going to be considered unusually stunning. Therefore, film opinion may also lose its authenticity [5].

4.3.2 Bi-LSTM model

It is not enough to analyse a random tweet to analyse a user, so it is necessary to analyse multiple tweets of a user, i.e. to perform sentiment analysis on multiple tweets, and put these n outputs into the network to get the final classification. The following diagram shows the input and output flow of the model:

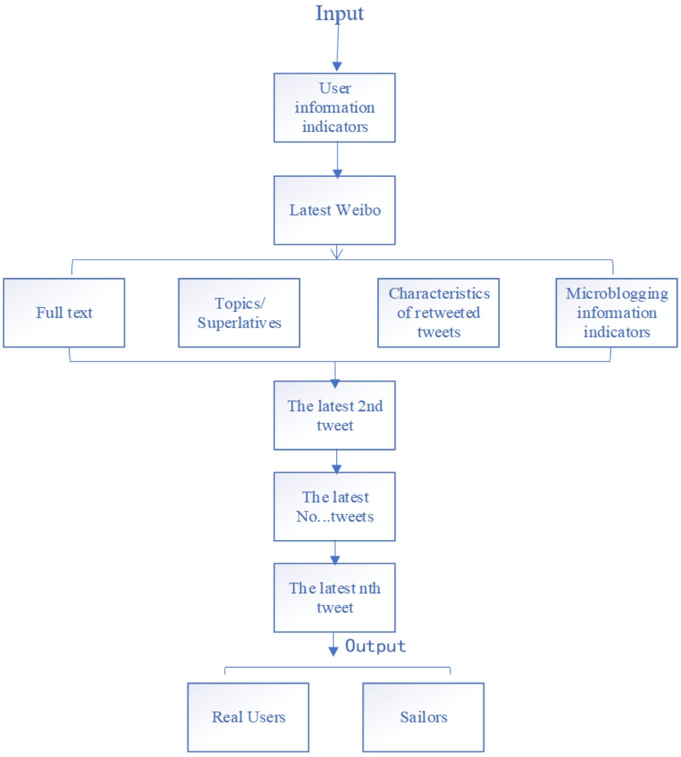

Figure 9 Input-output flow of the Bi-LSTM model

The network used to connect the n outputs will also be a recursive model. As this is a nested parallel LSTM model, most of the activation functions used are Tanh in order to prevent gradient disappearance and a 40% Dropout is performed on some layers to prevent overfitting. The model structure is shown in Figure 9 below.

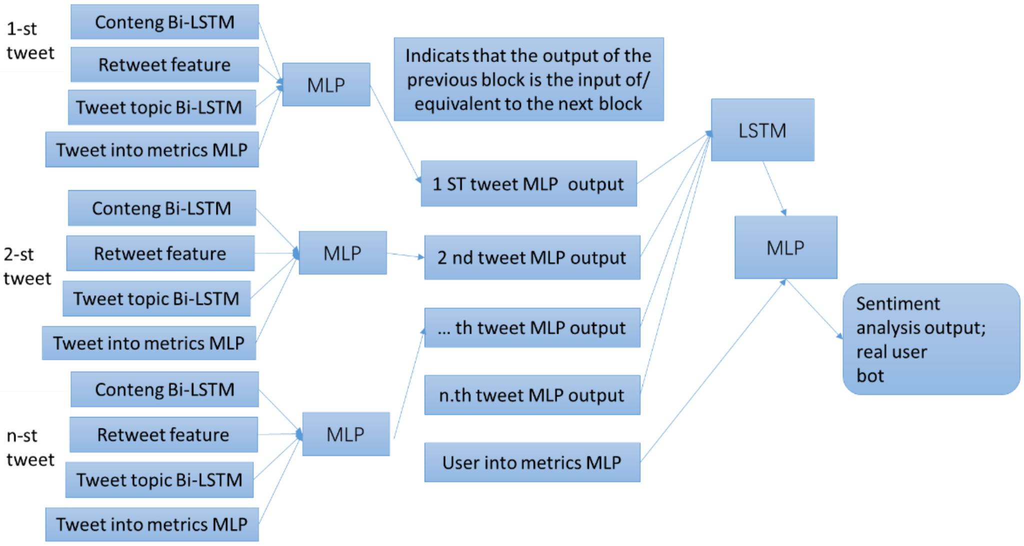

Figure 10 Results of the Bi-LSTM model

A user_id dataset of 568 samples (274 water users and 294 real users) was used for testing. All samples were annotated by manual inspection to ensure as much logic and distribution of the dataset as possible, thus objectively guaranteeing fairness in the accuracy of the test set. The user_id dataset is fed into the crawler code and it outputs a new dataset for the model. Also, {body} and {topics/hyper-talk; features of retweeted tweets} have very different grammars, vocabularies and sentence lengths. Therefore, when embedding, separate lexicons are created for them, which also allows the {topics/hyper-talk; features of retweeted tweets} to be downscaled before being input as one-hot codes for embedding. The following table 4 shows the results of some of the baselines：

Table 4 Selected baselines results

The split rate of the training set	Test set accuracy
85%	98.84%
50%	90.14%
15%	90.48%

Each different training set has a different lexicon and the accuracy of all the test sets is higher than 90% even though there are most unknown words in the test set. Therefore, the model results are reliable [6].

The analysis of the data shows that the audience cannot know the plot of the film before they watch it, so when they make a choice to watch a film, they often rely on publicity and film reviews to judge whether the content of the film is to their liking, which may influence the audience's choice to watch a film and also affect their value judgment of the film to a certain extent. For the viewer, publicity and advertising play a role in informing the viewer of the film, which is used to make the final decision on whether or not to watch it. They are therefore often influenced by the word of mouth of others as well as their own personal assessments and decisions.

4.4 Solution to question four

4.4.1 Time series modelling and solving

Figure 6, based on the results of Question 2, shows that if there were no external factors to interfere with the box office, it would show a year-on-year increase. However, due to the sudden new crown virus, all industries are in a stagnant trend, which has dealt a huge blow to China's development, and has also dealt a serious blow to the development of the film industry as the country has strict attendance requirements for the opening of cinemas. In this paper we analyse the total box office obtained by the Chinese film industry in the absence of the epidemic by querying single day box office data from 2013 to 2019 and using MATLAB to time series predict the set of data to obtain the following Figure 11.

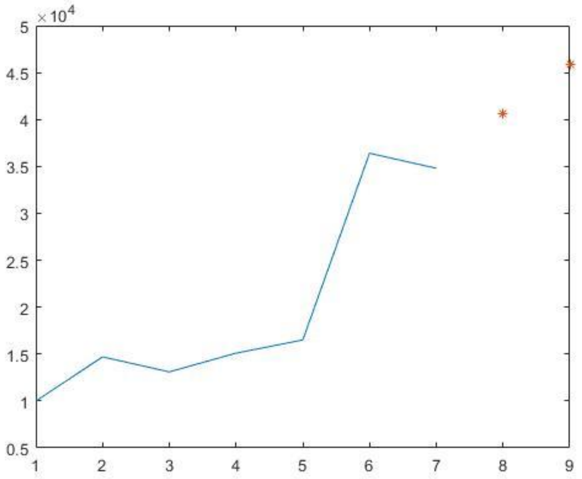

Figure 11 Box office forecast in the absence of an epidemic

In Figure 11, the two red dots indicate the projected gross box office of the film industry in 2020 and 2021 in the absence of epidemic disruption. However, the sudden outbreak of the epidemic in 2020 broke the original box office growth trend.

Our team analysed the box office data in Excel by finding 12 randomly selected days of movie box office data from 2013 to 2019 to obtain the following Figure 12.

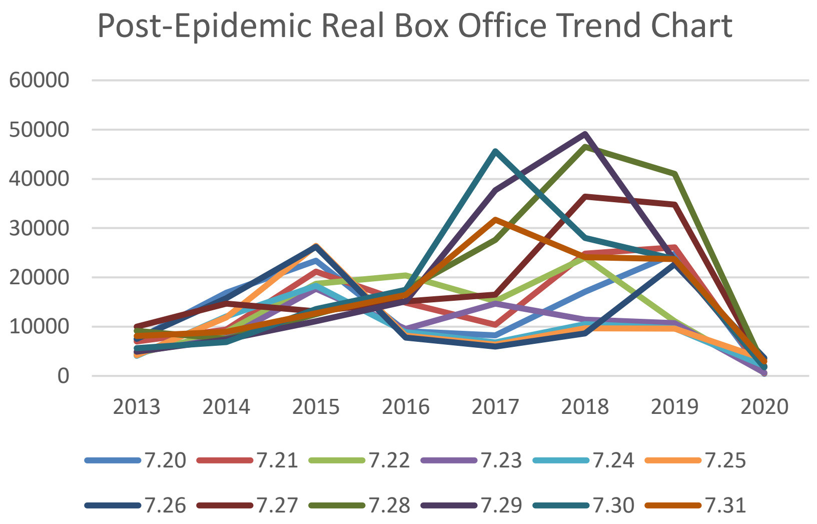

Figure 12 Post-Epidemic Real Box Office Trend Chart

Analysis of Figure 12 shows that the box office for films on the same day of the year fluctuates but mostly shows an increase, but by 2020 each day is infinitely closer to the axis and shows an abrupt decrease.

As a result of national guidelines for epidemic prevention and control, which require attendance at each venue to be no more than 30%, 50% and 75%, the box office has changed with the change in attendance requirements. The first round of increases in cinema attendance began on 14 August, with some cinemas in low-risk areas increasing their attendance to 50%; the second round of increases began on 25 September, with cinema attendance increasing to 75% and a big box office explosion during the National Day, but it still could not fully recover from the trauma caused by the epidemic to cinemas [6].

Therefore, our group looked up box office data for films from 20 July to 7 October within the years 2013 to 2019 and analysed the data according to the corresponding attendance rates to obtain predicted box office data and actual box office data for 2020, for which the degree of impact can be found according to the following equation.

$$c=\left(1-\frac{b}{a}\right)*100\%$$

Where c is the level of impact, a is the projected box office in 2020 and b is the actual box office in 2020.

The results are shown in Table 5 below.

Table 5 Extent of impact of different attendance rates on box office in 2020

	Forecast (2020)	Actual (2020)	Level of impact
30%	795134.86	60721.50	92.36%
50%	625239.47	491629.50	21.37%
75%	563765.56	419236.30	25.64%

By analysing Table 5, it was found that at a cinema attendance rate of 30%, the actual box office revenue was extremely low relative to the predicted box office revenue, with an impact of around 92%, greatly affecting the movie box office; at a cinema attendance rate of 50% or 75%, the actual box office revenue was nearly 20% less than the predicted box office, but the amount of box office loss was much less than that at a 30% attendance rate, so the impact was average.

5.Strengths and Weakness

5.1 Advantages of the model

(1) The time series forecasting model is very simple, requiring only endogenous variables and no other exogenous variables.

(2) K-means is relatively simple and easy to implement, with fast convergence; better clustering results; stronger interpretability of the algorithm; only the number of clusters needs to be called for.

(3) The Python crawler separates the generation task from the crawling data, with a clear division of labour, reducing the exchange of data between the Master and Slaver side; the implementation is simple, and the non-Scrapy framework crawler is also applicable.

5.2 Disadvantages of the model

(1) The time series prediction model is limited by the difficulty in obtaining box office related data, and there are some potential influencing factors of box office, such as film production cost, promotion cost, film ticket price and other indicators are not included in the model, and the number of samples selected is still not large enough.

(2) The selection of values in k-means is also not easy to control.

5.3 Optimisation of the model

(1) Optimisation of the time series prediction model:The accuracy and persuasiveness of the findings would be further enhanced in future studies if the sample size could continue to be expanded and internal data related to the film industry could be obtained.

(2) Improvement of K-means: The clustering centres can be obtained once by giving a suitable value to , at the beginning, through the K-means algorithm.

(3) Optimisation of the Python crawler: using the lxml library runs faster than the bs4 library, and can increase the speed by about 40%; using multiple threads, it can increase the speed by 68 times.

6.References

[1] Xi Jiawei. Data mining-based box office analysis of movies.
https://s.wanfangdata.com.cn/ . 13 November 2021

[2] Wang Qiuping. Research on movie box office prediction based on K-means clustering and BP neural network. https://s.wanfangdata.com.cn/ . 3 November 2021
[3] Yu Lanting. Analysis of factors affecting the box office of domestic films, https://s.wanfangdata.com.cn/ .13 November 2021
[4] The 5th Mathematical Modelling Competition for University Students, 2020 Team#202001502 . Modeling Public Opinion Monitoring and Sentiment Analysis. https://s.wanfangdata.com.cn/ 13 November 2021
[5] Li Jianfeng. A Study on the Impact of Online Film Public Opinion on Chinese Film Ecology. https://s.wanfangdata.com.cn/. 13 November 2021
[6] timmmGZ. Weibo sentiment analysis and crawler, water army detection with stable 95% accuracy. https://juejin.cn/post/6933146654551998472. 13 November 2021