**Government of the Russian Federation**

**National Research University Higher School of Economics**

RESEARCH PROJECT

*Analysis of the list of cult films by Kinopoisk in order to identify*

*patterns in the development of the film industry using the SNA method*

Written by

*Butrimova Maria*

*Panasyuk Sophia*

*Fedorov Alexander*

Tutor

*Karpov Ilya*

Moscow, 2022

Contents

[INTRODUCTION 3](#_Toc106885779)

[RELATED WORKS 5](#_Toc106885780)

[GRAPH ANALYSIS 7](#_Toc106885781)

[CONCLUSION 16](#_Toc106885782)

[LITERATURE REVIEW 17](#_Toc106885783)

INTRODUCTION

Nowadays movies are not the only kind of entertainment – it is one of the major sources of global commerce. Not only movie directors and box office officials are concerned with the success of movies but general people also. People used to talk about their emotions after films’ watching in social medias. Moreover, analysis of social media data about movies is recently popular among the data analysts.

Some users who often rate movies on the portal are film enthusiasts. However, the preferences of users can vary greatly - someone likes films of one genre or a combination of genres, someone - quite different. These dependencies can be implicit, hidden - not only genres, but also many other factors, various characteristics of motion pictures can have an impact here. Other than this there remains some other scopes like analyzing a director’s previous success histories or actor’s previous popularity etc.

The purpose of our work is to identify common trends in the 1000 must-watch films using SNA method for analyzing relations between directors and actors. Moreover, results of the research will be helpful in determining film preferences of the Russian audience.

Our project deals with analyzing the 1000 must-watch films from list made by Kinopoisk To achieve the goal, we used web scraping for gathering data from the service. For this purpose, we used specific Python libraries like Pandas, BeautifulSoup, Mathplotlib to work with data and analyze it. Moreover, it was necessary to go through Kinopoisk protection system, that’s why we used Python features (UserAgent, functions time and random etc.) for that. After that, we created several graphs in Gephi that helped us to identify the links between collected data and then to conduct the analysis for finding out the relationships between actors and directors. In addition to that, we used data about the main film genre for directors and actors from the list (based on their highest film in the rating) for understating genre distribution for cult films and perception of it by the audience.

It is important to say that as part of our studies at the Faculty of World Economy and International Affairs, we worked on clustering this category of Kinopoisk films in the IBM SPSS Statistics program. Thus, it is also in our scientific interests to compare the results of the previous analysis with the new one.

RELATED WORKS

There are a good number of studies to predict the movie success rate as people are too much excited about ﬁlms. For example, W. R. Bristi, Z. Zaman, N. Sultana explored the influence of such features as directors of the movie, screenwriters of it, the main actor and actress, country of origin, genre, IMDb rating on the audience perception of the future releasing film. For research purpose, authors used data collected form IMDb and Wikipedia related to 250 Hollywood films which were made in 2018. Eventually, they created a model which evaluates the potential success and the influence of different features. Thus, it was revealed that «some features having more influence on movie success and some other features having less or no influence»[[1]](#footnote-1). Authors mentioned that «budget is having a small positive influence but cast or actor/actress doesn’t have any influence on Russian film industry»[[2]](#footnote-2)

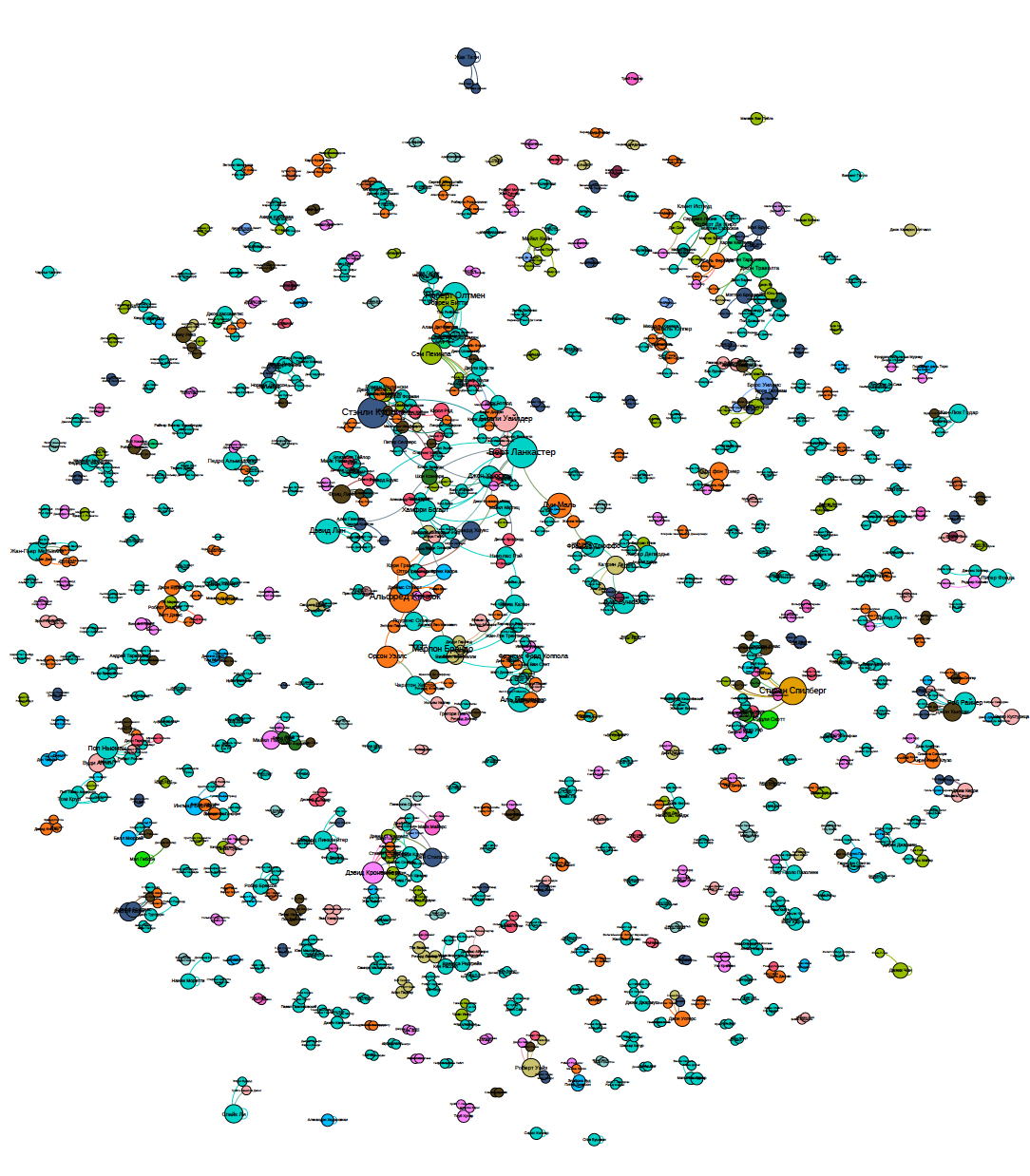
J. S. Krauss and P. Gloor in their work tried to predict «the success of new movies over their first four weeks in the box office after opening»[[3]](#footnote-3) and «identify movies as potential flops or blockbusters». For that purpose, researchers used data collected from IMDb (rating, comments), Rotten Tomatoes (rating, comments) and Box Office Mojo (box office performance data). As a result, it was found that the chance of a movie of getting nominated for an Academy Award was strongly connected with IMDb community discussion level: if the discussion level is higher, than movie has more chances to become an Oscar nominee. By the way, it needs to be mentioned that the main important feature of IMDb for that kind of analysis is that IMDb presents the audience perception of the movie in general, but this service also has one remarkable problem: users can give their rating to the film biased as the voting is open. That’s why for more reasonable results of the research it is necessary to use different metrics in addition to IMDb movie rating.

F. Chang in his research tried to create an IMDb rating prediction model. For that purpose, he used SNA method and studied «cooperative relationship between the director and the lead actor, the community relationship between directors and actors, the relationship between film workers' participation in social network activities and IMDB scores of their films»[[4]](#footnote-4). The researcher had chosen such features for analysis as the number of audience comments, the film title, the IMDb score of the film, director, the leading actor, film title. As the result, it was revealed that «he interaction and social activities between the staff and the audience itself are a kind of film publicity, which has an impact on the publicity of the film»[[5]](#footnote-5). Moreover, the analysis indicated that Steven Spielberg, Woody Allen, Martin Scorsese, Clint Eastwood, Ridley Scott and Morgan Freeman have strong influence and resource control ability in the whole Hollywood film industry and *users review* takes the biggest part of factors for IMDb score.

In conclusion, we can say that in the field of film industry research the growing interest of applying SNA method can be seen. Specificity of the film data and its potential for gaining an extra profit is the main reason for this tendency. Moreover, it really works. Saying that, we can mention the experience of Netflix in this field. Netflix collects lots of data from users such as time of watching, rating list, comments, the number of searches and what is searched for and even screenshots of scenes people might have viewed repeatedly. With this data, Netflix can create a detailed profile on its users. To collect all this data and harness it into meaningful information, Netflix requires data analytics. After analyzing it, Netflix uses this information for creating a new project. For example, Netflix show «House of Сards» was created based on audience preferences analysis and gained a great success in the USA and Europe.

GRAPH ANALYSIS

In order to achieve the goal of our project, we built a graph using Gephi and carried the SNA methods. We used directors and actors as the nodes and relations between them as the edges. Besides, we used the variable genre as the group variable to indicate the main genre for every director and actor and group them by genres. The graph below presents the results.

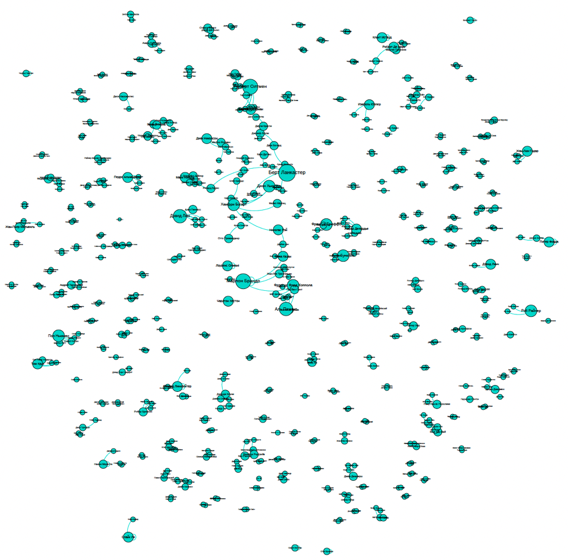
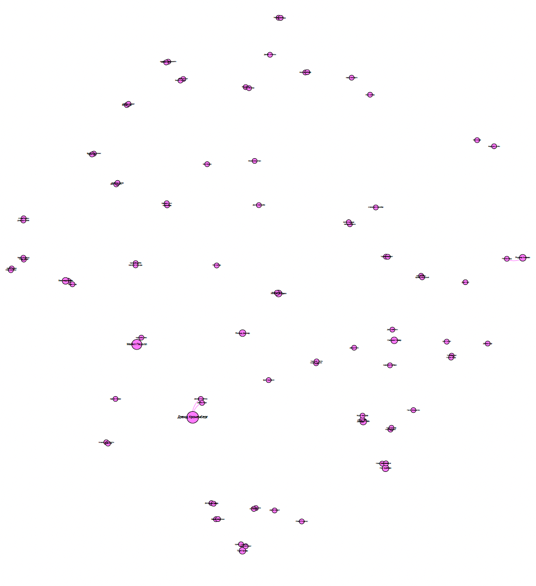


As we can see, it is hard to say something about edges between actors and directors based on their main genre, but also it is clearly visible that the films of different genres are presented in the list. For the research purpose, it is necessary to identify genre distribution. We used Gephi instruments to select which genre the most cult films belong to, determine their percentage of the total number (results are presented in the table below)

Изображение выглядит как текст

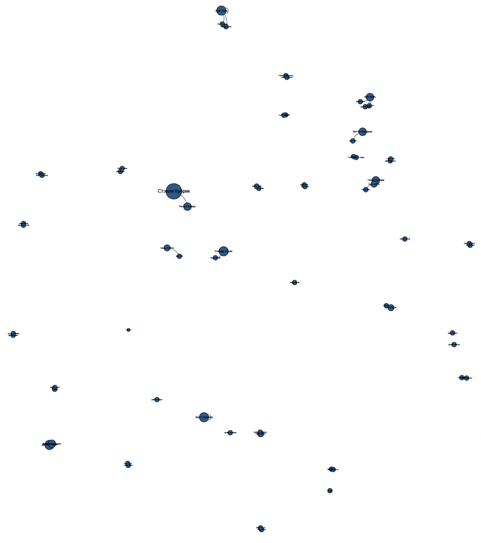
Автоматически созданное описание

As you can see from the table above, the most cult films are dramas (46%), followed by thrillers (10.16%), horror (6.98%), action (6.67%), comedies (5.04%) and fantasy (4.03%). Then, we decided to identify the links between film-makers with the same main genre. You can see the graphs below.

Dramas Horrors

Изображение выглядит как природа, дождь

Автоматически созданное описание 

Actions Comedies

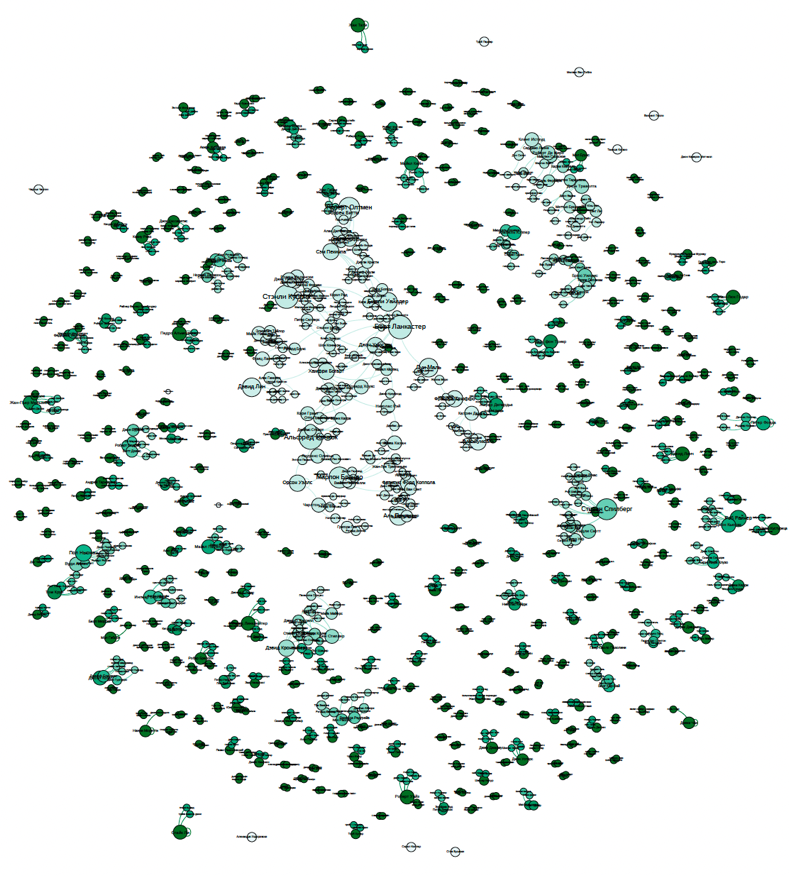


Sci-Fi

As you can see from the graphs above, the biggest number of edges between directors and actors can be seen for dramas. It depends, firstly on the genre distribution of the list for cult films, which allows us to say that the cult actors and directors are working mostly on dramas and, moreover, films of this genre can find the higher level of the audience approval. Among the most popular directors who made films in this genre are François Truffaut, Marlon Brando, David Lean, Robert Altman, Francis Ford Coppola and Luis Buñuel. Speaking of actors, Al Pacino, Marlon Brando, and Burt Lancaster should be mentioned here.

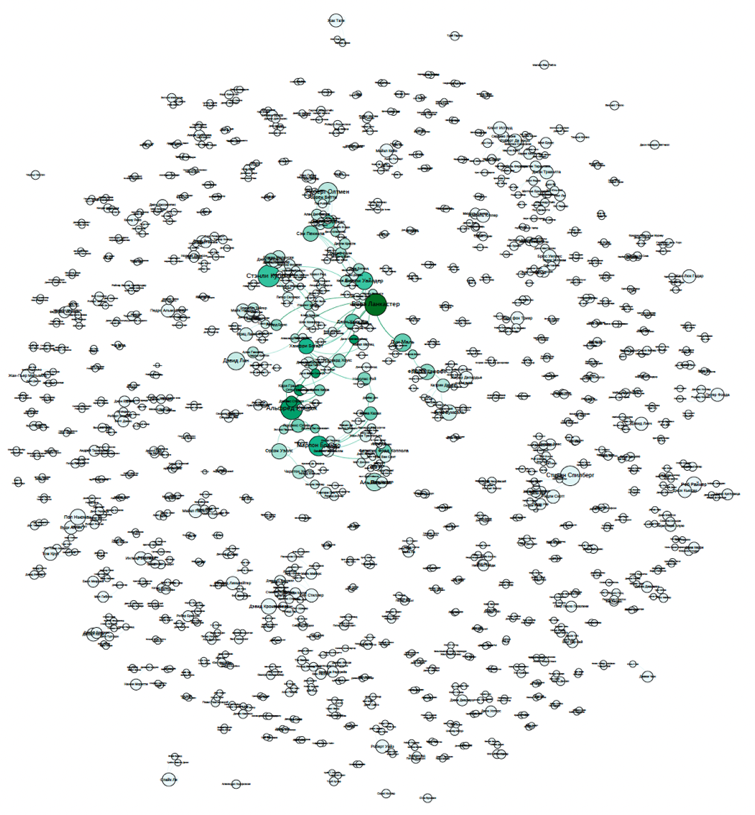
The next step for our research is to identify the most powerful clusters in the network. For that we used centralities and modularity analysis. Among the important metrics we must consider:

**• Closeness centrality**: the average length of the shortest path between the node and all other nodes in the graph. So, important nodes are close to other nodes.



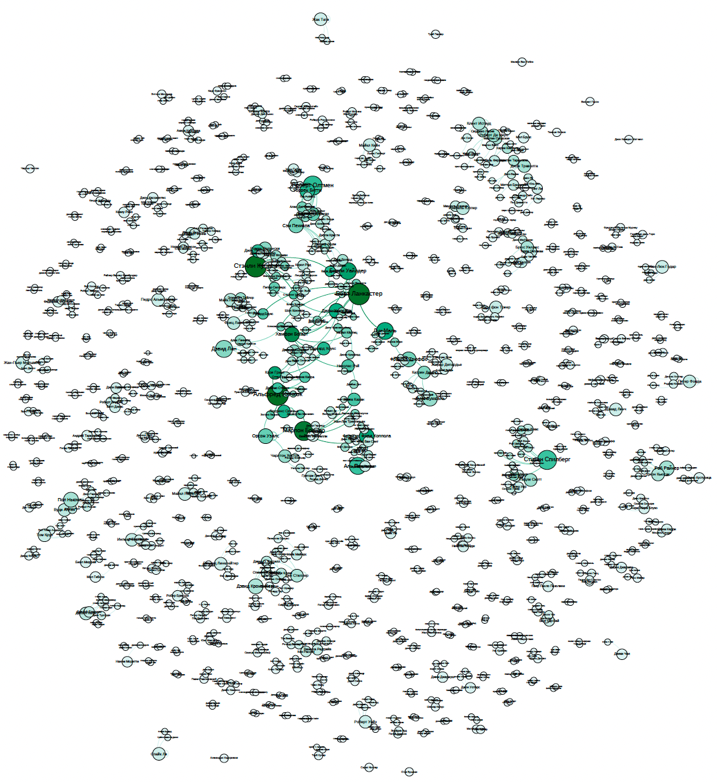
Closeness centrality graph

**• Betweenness centrality**: the number of times a node acts as a bridge along the shortest path between two other nodes. So, important nodes connect other nodes.



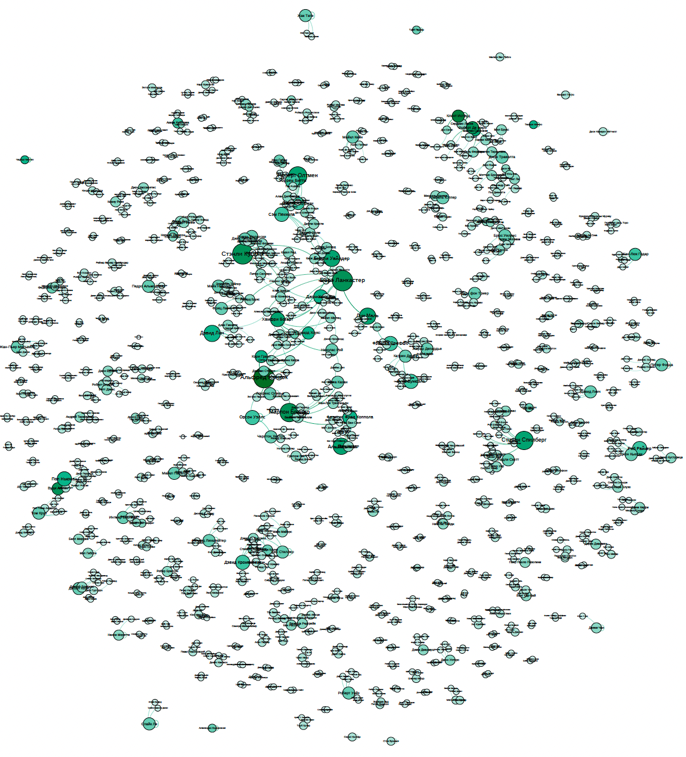
Betweenness centrality graph

**• Eigenvector centrality**: influence of a node in a network. It assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes



Eigenvector centrality graph

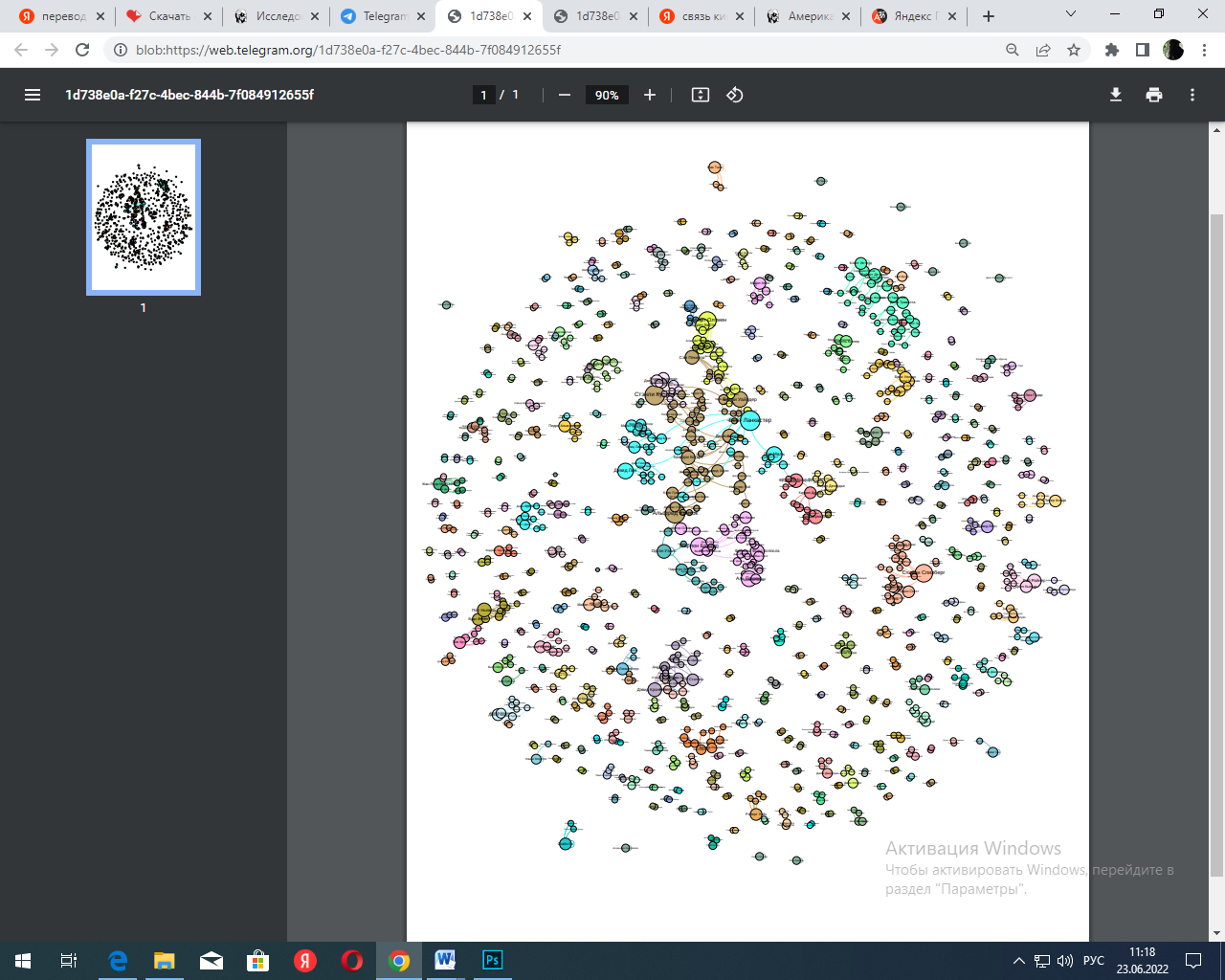
Then we decided to see which of the directors are the most prominent and famous. To solve this problem, we resorted to the strength of a graph. This one below illustrates the most famous film directors in accordance with our database. Therefore, the more famous the director is, the darker the green color.



Strength of a graph

Thus, having selected the most popular films by the film search rating and carried out a cluster analysis using graphs, we can conclude that the most prominent directors are Alfred Hitchcock, Stanley Kubrick, Sam Peckinpah, David Lean. As for the actors, here one can name Burt Lancaster and Humphrey Bogart.

Using various metrics such as centrality, we managed to identify possible clusters, and based on this data we decided to analyze the modularity through Gephi. As a result, we got 15 most interesting clusters which are analysed throughout our research project. Results are presented on the graph below.



In it, we identified 15 clusters developed by the following features:

1. Turquoise (Altman, Beatty, Hoffman): it differs in that it includes directors and actors working on films from 1960 to 1990 in the USA;

2. Orange (Peckinpah, Kubrick, McDowell, Connery): this graph includes the teams of directors and actors producing films in the USA in the period from 1960 to 1980;

3. Blue (Lin, Lancaster, Taylor): the blue graph is grouped by people working on films in the 50s of the last century in the USA;

4. Pink (Polanski, Foreman): It is divided into teams working on films from the 1970s to the 2000s in the USA;

5. Green (Ford Koppola): this graph includes teams of directors and actors producing films in the USA in the period from 1970 to 1980;

6. Pink-gray (Wells Heston): reflects the period of work on the films from the 1940s to the 1960s

7. Lilac (Truffaut, Deneuve): it differs in that it includes directors and actors working on films from 1960 to 1990 in France;

8. Bright Pink (Tarantino, Travolta, de Niro): shows the period of the directors' work on films from the 1990s to the 2010s in the USA

9. Ochre (DiCaprio, Williams): shows the period of work of actors on films from the 1990s to the 2010s in the USA

10. Cream (Spielberg, Lucas, Scott): differs in that it includes directors and actors working on films from the 1980s to the 2000s in

11. Light green (Kronberg, Myers, Stiller): shows the period of directors' work on films from the 1990s to the 2000s in the USA

12. Peachy (Redgrave, Russell, Ivory, Dunning): shows the period of the directors' work on films from the 1980s to the 2000s in the USA

13. Green-swamp (Newman, Allen): reflects the period of work on films in different years of the 20th century in the USA (generalizing cluster)

14. Yellow (Howard, Finney): It is divided into teams working on files from the 1970s to the 1990s in the UK

15. Gray-green (Aldrich, Davis): shows the period of directors' work on films from the 1940s to the 1970s in the USA and the UK

Thus, the generalizing trends that unite all of the above clusters can be attributed to the division of the graph into clusters according to the country-time principle. Note that the main content-makers (the largest elements of the graph) in the list of cult films are the directors of the mid-20th century. We can also see separate clusters with directors and actors from the 2000s. We can also see several actors capturing several eras of cinema development, which caused such a cluster division.

The resulting graph shows that the largest number of films which were produced in the United States, which is why interconnected clusters are formed for different periods.

It is important to identify the relationship between American and European cinema, which can also be traced in clusters. The spread of American cinema in Europe, and in particular in France and the UK, not only brought huge profits to American film studios, but also significantly influenced the consciousness of Europeans, created a positive opinion about the United States, about American society and culture. And from this point of view, the concept of the distribution of cinema during the process of globalization and the concept of "cultural imperialism" do not seem contradictory: the huge success of American cinema in Europe after World War II met the goals of both private film studios and TNCs, and the State Department, seeking to prevent the ideological victory of communism in Europe and to position the societies of European countries in favor of USA.

CONCLUSION

As the business market of ﬁlm industries are becoming huge day by day, competition here is also growing complex. Accordingly, this determines the importance of conducting a clustering analysis of the most popular films of all time among Kinopoisk users.

The main cult films, according to Kinopoisk viewers, were created in the period 1960-1980, since we can observe a closely interconnected sector within the graph. different directions of cinema (by country-time period) influence each other, and thus, a single cultural space of cinema is formed.

Besides, as the result of our analysis we identified that the greatest numbers of cult films were made in the USA and Western Europe. Moreover, we found the strong connection between American and French cinematograph, which is interesting for the future research in term of liberal arts field.

Thus, having selected the most popular films by the film search rating and carried out a SNA method using graphs, we can conclude that the most prominent directors are Alfred Hitchcock, Stanley Kubrick, Sam Peckinpah, David Lean. As for the actors, here one can name Burt Lancaster and Humphrey Bogart. Drama is the most popular genre. Among the most popular directors who made films in this genre are François Truffaut, Marlon Brando, David Lean, Robert Altman, Francis Ford Coppola and Luis Buñuel. Speaking of actors, Al Pacino, Marlon Brando, and Burt Lancaster should be mentioned here.

That’s why it is important to mention that for future success of the film industry products it would be useful to study the main features of films which were made by mentioned directors. Probably, it would be helpful to identify the main features of their works and to impose them into new product. Of course, we are not talking about plagiarism but as the area of art film industry can use the experience of the past, which can help to develop future product in terms of their quality and artistic content. Films, which will be made using results of research, probably would be more successful and profitable not only in a short-term but in a long-term as well.

LITERATURE REVIEW

Bristi W. R., Zaman Z., Sultana N. Predicting iIMDB rating of movies by machine learning techniques // 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). 2019.P. 1–5, 2019

Chang F. Film Project Analysis Based on SNA: https://zhuanlan.zhihu.com/p/171923925

Krauss J. S., Fischbach K., Simon D., Gloor P. Predicting Movie Success and Academy Awards through Sentiment and Social Network Analysis **//** 16th European Conference on Information Systems (ECIS). 2008. P. 1-12

1. Bristi W. R., Zaman Z., Sultana N. Predicting iIMDB rating of movies by machine learning techniques // 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). 2019.P. 1–5, 2019 [↑](#footnote-ref-1)
2. Ibid. [↑](#footnote-ref-2)
3. Krauss J. S., Fischbach K., Simon D., Gloor P. Predicting Movie Success and Academy Awards through Sentiment and Social Network Analysis **//** 16th European Conference on Information Systems (ECIS). 2008. P. 1-12 [↑](#footnote-ref-3)
4. Chang F. Film Project Analysis Based on SNA: https://zhuanlan.zhihu.com/p/171923925 [↑](#footnote-ref-4)
5. Ibid. [↑](#footnote-ref-5)