Final project NYPD 23/24

1. Final project (our proposal) - FILM RANKING

Create a model to measure "cinematic impact" on the data collected from IMDb (the international movie database; see section Data Sources). Next, use the created model to perform at least three analyses. Two of them should be exactly as they are defined below. However, for the other(s) ... the sky is the limit (sample questions in the assignment text below could be an inspiration, but it is even better to ask your own questions). See section Data Analysis.

See also sections Detailed Requirements, Technical Requirements and Exam Details.

1.1 Introduction to Model Idea

Below, we define the model on which your analysis should be performed.

Film Hegemony / GOAT Director / Actor

Our aim is to create a model to measure "cinematic impact" by numbers. As a result, we can order countries/directors/actors and finally answer the following questions:

- Which director is the greatest of all time?
- Which country is closest to film hegemony[1]?

But how do you measure "cinematic impact"?

Quality vs Quantity (two alternative approaches)

Popular cinematic services provide the numbers to describe a film, like the average rating, the number of votes, the film budget, etc. A sample idea is to consider films' cinematic impact:

1. Quantity - the number of votes matters; the film should have a big audience.
2. Quality - bad films are bad (even worse than no film because of the negative impact);

HINT: From *`title.ratings`*, we can obtain an *`averageRating`* and *`numVotes`*.

Potential difficulty in creating and using such a model is in the details. The choice between quality and quantity will influence an answer:

- Who is better, Kieślowski or Vega, Wes Anderson or James Cameron?
- Are films from Italy better or films from the USA?

One can combine such measures or use a selected one. However, even quality ranking has its challenges: How should two movies be treated with an average rating of 8 both and a number of votes of 10 and 100,000? And what if we want to compare films from Italy with US movies?

[1] Civilisation game series enthusiasts (in the game one develops civilisation to rule the world) already know that victory can be achieved without military actions, only by dominating the world's culture.

Totality or Fraction (create own metric)

Assume we want to compare directors' or countries' "cinematic impact." If we focus on quality, selection is crucial for our measurement.

We can compute the average score from all the movies.

The first difficulty: But what if we obtain the same score?
In our model, should an unpredictable director responsible for masterpieces and flops (9 1 9 1) be better than a mediocre director (5 4 6 5)?
Possible solutions: We can assign special weights to masterpieces and/or flops and focus only on masterpieces. On the other hand, we can compute the variance to find a predictable director and minimise the chance of disappointment. Imagine going on a date. Find the director with the best score among all directors with an acceptable level of predictability.

The second difficulty: Number of films considered.
If we consider the quality impact of directors, what should we do if their total number of films does not match? Intuitively, a director with more films seems to be better for the same score. Another idea is to choose a number X arbitrarily, focus on only the top X best movies, and compute the score only for them. What do you do when a director has fewer than X films?

Universality or specialisation (genre)

Let's focus on the arbitrary classification of films into genres (*`title.basics`*).

- Are Italian westerns better than American westerns (due to Sergio Leone)?
- Are Czech comedies really the best in the world?
- Is Artur Barciś (due to roles in Dekalog) the best dramatic actor of all time? What if we compare that score with Barciś, a comedy actor?

Language

Idea: test obtained score (based on total votes / average rating and all movies / chosen N movies) for the language of a movie for different genres.

- For example, is the order of languages preserved for comedies and dramas? Or are Czech comedies way better than Czech dramas? In general, for one language, what is the difference in scores for dramas and comedies?

"Cinematic impact"

To measure the quality of a “cinematic impact” of a country, one may consider additional economic parameters.

- Countries with small GDPs potentially invest less in cinematography. Is it true? Can we observe any trends? Intuitively, Czech "cinematic impact" is greater than Belgian despite similar population potential. Is Rwanda's "cinematic impact" large among African countries with low GDP? (It is an intuition based on the fact that there was a

Rwandan movie at the Warsaw Film Festival). Is Belgium "weak" or The Czech Republic "strong" considering EU countries or countries with similar GDP/population? Are there countries that beat such a trend?

Let's define:

1. Weak impact - as the total number of votes.
2. Strong impact - as the quality score of models metric.

A **fictional** example

Let’s assume that:

A. GDP total: 1. China 2. USA ... 23. Belgium ... 34. The Czech Republic ...;
B. Weak "cinematic impact": ... 15. The Czech Republic ...;
C. Strong "cinematic impact": ... 7. The Czech Republic ….

That would mean that The Czech Republic is +19 and +27 better than their GDP total potential.

1.2. Data Sources

Datasets are available at - https://datasets.imdbws.com/
The description can be found here - https://developer.imdb.com/non-commercial-datasets/

1.3 Data Analysis

Task 1 - Quality of movies
Analyse the quality of movies by country. Create the order of countries for the representation set of the size of 10, 20, ..., 200 best films.
What are the top 10 countries for each set? Do you observe any trends?

Task 2 - "Cinematic impact" hegemony (country order)
Compute: (I) weak impact (total number of votes) and (II) strong impact (quality score of models metric) for all the countries, use: (a) population, (b) GDP, (c) GDP/population

Find the film hegemon countries with great "cinematic impact" wrt population, GDP, and GDP/population.

An example of GDPs and a fictional order of countries with weak and strong impacts has already been given before. The hegemony is computed as the difference between GDP rank and impact rank. You can compute hegemony defined by a different formula.

Task 3 - Own Analysis
Any analysis. The previously mentioned ideas can inspire it. Compare American vs. Italian westerns OR compare Sam Peckinpah vs. Sergio Leone westerns OR ….

1.4 Detailed Requirements

You need to write a Python program that:

- Takes (using *`argparse`*) file paths from the command line for files (in CSV format) containing data (GDP, population, …).
- Assumes the files are in the same format as today (although there may be more data in the future, meaning you cannot assume the program will work exactly on the files available today, but you should assume the format will be preserved) and reads them in.
- Cleans the data (depending on files, there could be little to do in this step).
- Selects only data from the years that are present in all tables. Note: this refers to the years that actually appear in the tables, even in the future, i.e., you cannot rigidly assume in the final version (hardcode) that the data is, for example, for the years 1960-2024, although this may be convenient in the initial versions of the solution.
- Combines the data from the read files.
- Performs analyses (using the NumPy or pandas libraries).

1.5 Technical Requirements

- The program should be divided into (at least) two Python modules (they don't have to be large): one containing a library of operations for analysing data according to the assignment description and the other one (it might be a Python file or a notebook) calling these operations.
- We expect the program's source code to be in a Git repository with at least five (5) commits.
- We expect one pull request to be submitted (and accepted).
- Unit tests of the code should be provided (where it makes sense).
- You should profile the program's performance (a file with the profiling results should be included with the solution. However, you are not obligated to perform the optimisations suggested by the performance analyses.
- The program should allow for specifying the range of years on the command line (parameters -start year and -end year), in which case the analysed period is additionally shortened to the given range. If the range after this operation turns out to be empty, an exception should be raised (and handled somewhere in the program).
- The program can be a regular program or a Jupyter Notebook.
- The program should be delivered as a package that can be installed with the pip command (in the form of `pip install ./path/to/package`).
- If the program detects any inconsistencies in the data, it should print a clear message and then continue running. The program might issue a message for each error, or the error messages might take a more cumulative form, like the number of errors for each error category.

1.6 Exam Details

- This course has an oral exam, which involves presenting your final project to the teacher (usually from your group).

During this conversation, we expect:

- A presentation of the program's functionality.
- Presentation and justification of your design decisions (e.g., program division into modules).
- Presentation of your code repository.

As regards the program, we take into account:

- How your code makes use of modules, functions, and parameters.
- Whether it is readable (variables and functions names, comments).
- We suggest using tools that automatically check (some) recommendations regarding code readability. Such tools may be built-in into your IDE (e.g., PyCharm), or you can use external tools (pylint, flake8, ruff, etc.).
- It's also worth reading the Python Style Guide PEP-8 recommendations.
- To what extent tests cover your code.
- Use of libraries presented during classes.
- The delivered files (e.g., profiling results).

2. Your own project

You can choose a different topic for the assignment that meets the following criteria:

- It is not less difficult than the task presented here.
- It has been agreed upon with and accepted by the teacher in your group.

3. Grading rules

Submitting the program to the repository and emailing the teacher should happen:

- **For the first deadline:** by July 14th.
- **The second deadline is** September 8th. Submitting the solution after the first deadline will lower the grade **by one**.

After the email, the instructor and student will schedule a personal/Zoom meeting. These meetings from the first deadline do not have to happen before the first protocol is closed - in that case, we will enter the grade (without lowering, of course) in the second one.

We wish you success!