SOLENE

A geographical analysis of Culture

This subject tackles the question of cultural diffusion and availability. We are everyday confronted to cultural ideas and concepts. In our globalized world, who creates the cultural content on which a population is confronted? Is it uniform over the globe or are there disparities? Using the movie dataset, we can try to dive into a snippet of this broad question by looking at culture revolving around movies.

Who produces the films? On what geographical area are the movies available to the public? Is there a monopoly of a few country (US films for example) or is the movie production more geographically distributed? We could extend the research to actor's nationalities as well as producer's. Are actors only playing in movie studios of their own nationality or is there crosspollination? This analysis would use the language in which a film is translated, as well as potential additional databases such as the Wikipedia pages of actors to know about their nationality, or info about cinemas frequentations and visas (https://www.data.gouv.fr/fr/organizations/centre-national-du-cinema-et-de-l-image-animee/)

History of Memories

The way historic events are vulgarized and popularized evolves with time and space. By analysing the plot summaries of movies revolving around historic events, we could try to understand how those historic periods are portrayed. Is it an action movie? A drama? On what element is the movie focused? On what type of characters is it focused? Heroic players, villains, nobles and elites or more popular social classes? This character analysis could be based upon already existing researches such as the one presented in the Learning Latent Personas paper.

Interesting periods to be used for this analysis could be World War I and II, the Cold War, the Colonies. In some cases, it may be interesting to expand the dataset with more recent movies in order to observe evolution over the 2000-2020s.

How has a specific genre evolved over the years?

The dataset contains movies that were filmed over a more-than-60-years period. Ranging from the 1950s until the 2010s, the dataset usually presents at least 200 movies per year. Most of them include genre categorisation.

By leveraging the plot summaries available in the dataset, we could try to analyze how a specific genre (for example horror or science-fiction) has evolved over the years. How many characters are displayed? Are the personas presented similar to each other or is there an evolution trend? What kind of events happen in the plot? We could relate the evolution in plots and subjects with technical improvements in the filming industry (cameras, digitalisation, etc.) and see whether technological progress influences plots in any manner.

Guess the story's genre

In a reverse way, we could also try to extract a movie genre from a textual plot summary and fill in the gaps in the dataset. By analysing the vocabulary used, the types of situations happening in the plot, we could create a model that would attribute a genre to a textual plot by aggregating keywords to specific categories. The method could be tested using the data already provided about genre.

Great and original ideas, just be careful that your ideas are feasible given the available resources. (1) Interesting idea, and good that you already explored where to obtain the data you need. A small note, you may want to look for data regarding cinema visits in other countries as well. (2) Original idea, with good potential. You should however be careful about linking movies to specific historical events, as this may be quite difficult to do given the plot summary. Also do you already have an idea of what you would like to look for in the plot summary to support your analysis. With the plot analysis, also be careful that you find a way to extract what you want before choosing this topic. (3) How would you plan to measure the similarity between different characters? Also, it may be difficult to pinpoint exact events in the plot summaries of the movies. Moreover, what link would you expect between the technological advances and the plot, and how would you measure this? (4) Interesting, just make sure you will have enough findings to visualize in the final data-story. Your project mentor throughout the semester will be Marija Sakota: marija.sakota@epfl.ch. For future discussions specific to your P2 and P3 deliverables, you are encouraged to be in touch with your mentor.

AZIZ

Idea 1 : Cultural diversity in main characters .

Cultural diversity in movie characters has always been a controversial topic . From Ben Affleck criticised for playing Mexican-American lead character in

Argo to backlash over Ariel character in "The Little Mermaid" being played by a black woman to promote inclusion , there has always been this feeling

that the cinema industry and especially Hollywood lead characters have been dominated by a certain social category , usually white men . Are main characters really predominantly white men ?

And how did this trend change over the years ? If this trend really exist , is it specific to Hollywood which dominated the movie industry or do we see more diversity in other countries cinema industries .

To achieve this task , we will first have to identify the main characters from the plot summary of each movies using an NLP library and then assign to them an ethnicity and gender according to the actor

who portray them . Since quite a few actor ethnicities are missing from the main CMU Movie Summary corpus , we will have to add data from wikipedia about the actors .

Idea 2 : How to choose a movie title

A great movie title can make a major contribution to its success because it conveys to the consumer, in just a few words, what the movie will be about. It therefore needs to be carefully crafted so that it

conveys just enough information without giving away too much.We can use machine learning to give you the right movie title for your movie. We take into account movie metadata, such as success ,genres,

character names and other actors involved in the movie and then we try to label the plot summary using NLP (natural language processing) techniques to identify common patterns .

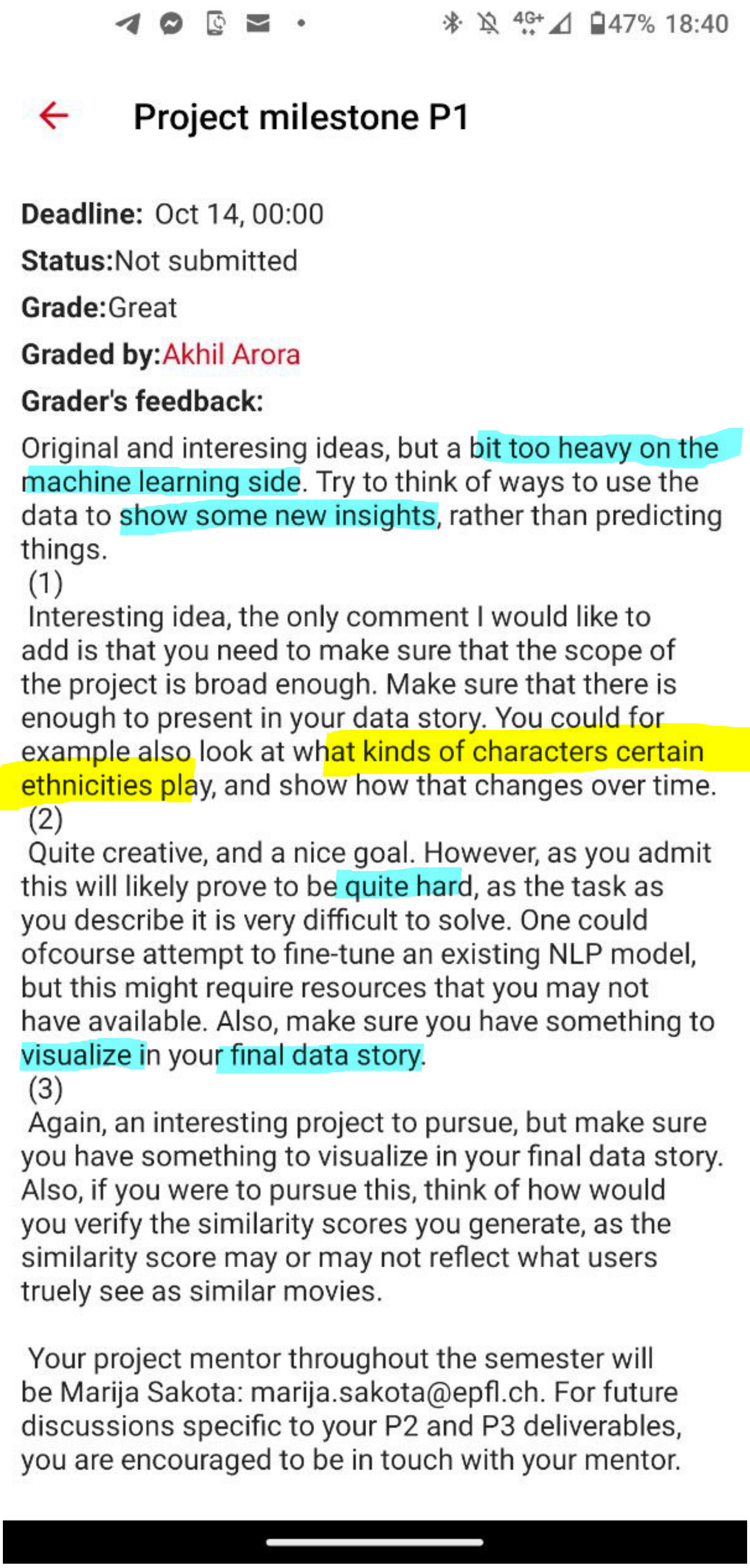
This is a very difficult task and one that is not yet fully solved. To do this, the idea is to find words that are similar in meaning and their contexts, and then assign each word an index based on how often it

appears in a given context (or sentence). This process is called 'vectorization' . Going from there, we will try to generate the most appropriate movie title .

Idea 3 : Movie recommendation system

How many of us have gone on netflix looking for a specific movie to find out it is not available on the platform but Netflix still manages to display as results very similar movies but with a different titles .

We will try in this project to recreate this movie recommendation system . Our movie recommendation system is based on the idea that we can discover new movies for users by finding similarities between their favorite movies and other similar movies. We accomplish this by establishing a similarity score between different movies and analyzing movie characteristics like genre, year of release , character types , actors and plot summaries .



Project milestone P1

Deadline: Oct 14, 00:00

Status: Not submitted

Grade: Great

Graded by: Akhil Arora

Grader's feedback:

Original and interesing ideas, but a bit too heavy on the machine learning side. Try to think of ways to use the data to show some new insights, rather than predicting things.

(1)

Interesting idea, the only comment I would like to add is that you need to make sure that the scope of the project is broad enough. Make sure that there is enough to present in your data story. You could for example also look at what kinds of characters certain ethnicities play, and show how that changes over time.

(2)

Quite creative, and a nice goal. However, as you admit this will likely prove to be quite hard, as the task as you describe it is very difficult to solve. One could ofcourse attempt to fine-tune an existing NLP model, but this might require resources that you may not have available. Also, make sure you have something to visualize in your final data story.

(3)

Again, an interesting project to pursue, but make sure you have something to visualize in your final data story. Also, if you were to pursue this, think of how would you verify the similarity scores you generate, as the similarity score may or may not reflect what users truely see as similar movies.

Your project mentor throughout the semester will be Marija Sakota: marija.sakota@epfl.ch. For future discussions specific to your P2 and P3 deliverables, you are encouraged to be in touch with your mentor.

LOIC

Gender and ethnic evolution in the cinema. The recent drama about “The little Mermaid” shows us that cinema has an impact on our vision of the world. Each film is part of a certain vision of the world that will influence spectators by challenging or comforting it. Among the most distinguishable visions, is how the minorities are represented. We can analyze genders and ethnicities look at their evolution during time, across countries and according to films. We might also investigate the roles given to those characters and their importance. For this, we could use the summary with the character name and make a small analysis with the number of occurrences, for example. The aim is to compare those evolution over time, countries and genre and see if we can find similarities. Identifying the movie release date. How can we recover the release date of a movie? As we already have the release date, the idea is to find different solutions and compare them with each other’s. We could use the summary to identify keywords and train a machine learning algorithm on them. A second way would be to use the actors’ names, date of birth and machine learning. About 8% of the values are missing in the dataset. How would the two models matches for the missing values? The network of actors. Since we have movies that connect actors together, we can show how actors are interconnected. The idea is to map the actor’s network, where the bonds will correspond to films where both actors have played. Then we could analyze that network, identify sub-networks, find if they correspond to real-life aspects like studios or countries. Is the language the biggest factor of sub-networks? What are the links between those hubs? The aim of this subject is to investigate relations between actors and how we can rely them on other things than movies played together like gender, locations or movie genre.

Interesting ideas to pursue, however some small things need to be addressed. Also work on the details of the proposals, as you only used 300 out of 500 words. (1) Think a bit more clearly about how you would like to quantify "importance" in a movie. Indeed, word count may be one option, however, this would discard any information about what kind of character is protrayed. To this end you may be able to also look at te latent character type in the dataset. (2) Interesting approach to fill missing data. However, it is missing some data-science centered ideas. If you would like to pursue this topic, make sure you do a thorough analysis of why a certain model came to a prediction, and present this in a nice fashion to get a complete data story. Also, how would you like to compare different output dates of different methods? (3) Original idea to pursue, not much to add. Your project mentor throughout the semester will be Marija Sakota: marija.sakota@epfl.ch. For future discussions specific to your P2 and P3 deliverables, you are encouraged to be in touch with your mentor.

AHMED

1. are certain characters portrayed in a different light in movies produced in different countries?

2. are movie plots getting more complex over time? Are certain genres more complex than others? Do certain countries produce more complex plots than others?

3. are certain character archetypes a good predictor of the genre of a movie? Or of movie success (box office revenue, IMDb review)?

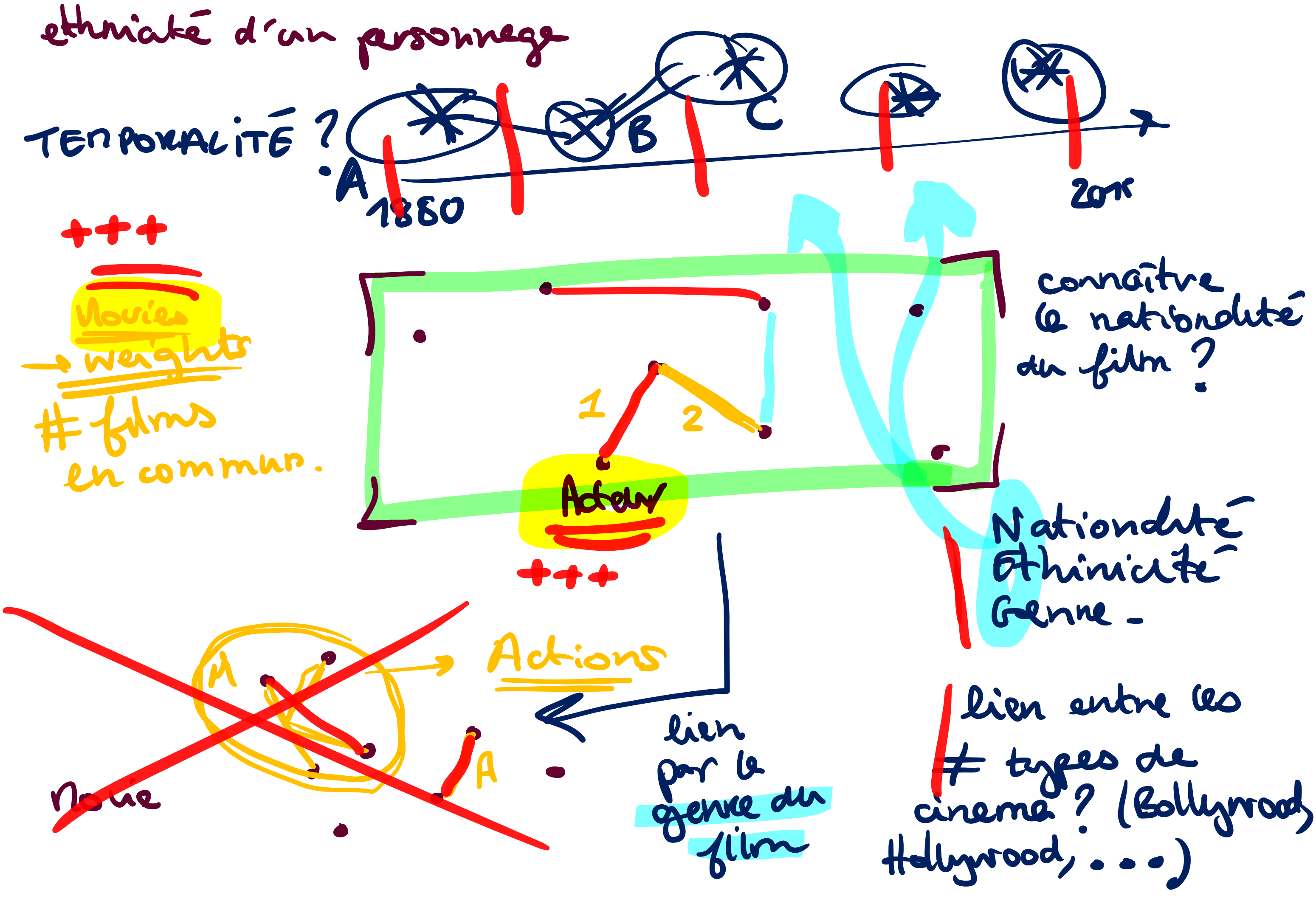
It is no secret that in the process of translation many details could be lost, or added due to human error or language limits among other factors. This is especially true in the film industry, where the translation team should be careful to accommodate the target country's culture. This ranges from minor dialogue modifications to whole scenes being cut or edited. It also goes without saying that the same characters that figure in movies produced across different countries should at least be a little different. An interesting idea for this project could be to make a collection of unique characters that appear in movies of two different languages at a minimum and extract features from the corresponding movie's summary plot in order to quantify the extent to which the movie's language could influence character portrayal. This could be done in many ways, either by identifying character archetypes or by introducing new metrics based on the description and actions of the character. However, a potential setback is that the amount of data we end up with after satisfying the collection criteria might be insufficient to draw any significant conclusion.

The idea of measuring the complexity of a film is a very captivating topic. There is no doubt that movies are growing more complex over time, but how much? Does it vary by genre? Are certain countries producing more or less complex films than others? In order to answer these questions, we need to first define complexity. The simplest way to do this is by looking at the number of characters and actions in the summary plot. The film duration can also be considered a factor, since typically, complex films take longer to tell their story. Finally, we can look at the number of genres that are represented in the film. If a film has a number of genres, then it is likely to be more complex than one that only focuses on one topic.

Potential setbacks could be that our data contain much more American movies than foreign films and that there are many more blockbuster movies than indie ones, for instance. It is also worth mentioning that there are many missing values in the film's length column, and we could not find the exact start year of many films. This, however, may be resolved by fetching other data sources, like IMDb.

A character archetype is a “universal symbolic pattern” that is often used in literature, film, and other forms of popular culture. One could speculate that There are many character archetypes that imply a specific genre. For example, the “dumb blonde” archetype implies a comedy, while the “tough guy” archetype implies an action movie. However, it is unclear whether these archetypes are a good predictor of the genre of a movie or of movie success. An interesting use case for the dataset would be to try identifying the different character archetypes from the summary plot, then to explore correlation between these archetypes and the movie genre. We could also correlate the archetypes with the movie's success, but this requires additional data sources, such as the box office revenue (the box office revenue provided in the dataset comprises a lot of missing values) and the IMDb review score. Another option would be to try to build a model that predicts the movie genre from the character archetypes. In all cases, this exercise would require NLP techniques to vectorize the text data in order to be able to use it in a machine learning model.

Good ideas, (Warning) but make sure you stick to the word limit for the next milestone! (1) I am not sure what you mean by "a collection of unique characters that appear in movies of two different languages at a minimum". Do you mean by this that you would like to find 2 different movies starring the same character or a translation of a movie? In the first case, you may have a very difficult time finding these, as I could imagine that they are quite rare. A small issue with using character archetypes, is that we can reasonably expect that the archetype of a character will stay the same across laguages but that rather some small details differ, e.g. a villain will stay a villain. You could potentially look at the description and actions, but these may be similar in the plot summary as well. (2) Good idea, but be careful with how you define complexity. Since counting the characters and actions may not be sufficient to explain complexity, as a movie with few characters could have a very complicated plot, while the plot of a heist movie with a lot of characters could be considered linear. Also make sure you are able to extract actions from the movie plot summary. (3) Interesting idea, just be careful that the idea is feasible. For example prediction archetypes by analyzing the plot summary is probably quite difficult to do, and would likely require training/fine-tuning an NLP model. However trying to predict the genre by looking at the archetypes could be an easier approach. In any case, make sure you have enough data to visualize in your final data story. Your project mentor throughout the semester will be Marija Sakota: marija.sakota@epfl.ch. For future discussions specific to your P2 and P3 deliverables, you are encouraged to be in touch with your mentor.


ethnicité d'un personnage
TEMPORALITÉ ?
A
B
C
1880
2015
+++
Movies
→ weights
# films en commun.
1
2
Acteur
+++
connaître la nationalité du film ?
Nationalité
Ethnicité
Genre.
M
A
Actions
~~Movie~~
lien par le genre du film
lien entre les # types de cinéma ? (Bollywood, Hollywood, ...)