**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.*

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 .

