ADA Project: milestone 2 propositions

Dataset: CMU

Ideas:

- Is the cinema database representative of our society? Are the jobs, names, ages, genders, ethnicities of people we see in movies representatives of the composition of our society? We can draw a ‘portrait robot’ of the average human being.

- Do the names of cinema actors influence the names of the babies in the USA? or is it the opposite? Study by region, by year, …

- Is a random pool of these movies realistic? can we define a realism metric system, based on an arbitrary number of criteria that help characterize our society, and rate various random or chosen pools of N movies among the dataset. For example, are the rates better if we remove Sci-Fi movies from the dataset ?...

PLAN pour P2 :

Proposition HUGO : (cf en dessous)

Proposition HUGO

Intro :

La société humaine s'est éteinte ne laissant derrière elle aucune trace à part la CMU database. Quelle image pourrait se faire de nous des extraterrestres trouvant cette base de données ?

I. Le portrait-robot de l'humain moyen

a. Métier : on peut les extraire à partir des noms (typiquement pour les militaires, leurs grades, pour des médecins, un filtre sur le prefixe 'dr', pour les prof, un filtre sur 'prof' etc.

b. Physique : origine ethnique, taille,

c. Prénom : cf ce dataset

d. Qu'est- ce qu'il aime bien ?

i. Analyse sur les genres/thèmes de films les plus aimés ou abordés.

PS : possibilité de faire des analyses causales ici en regardant si ces paramètres sont vraiment déterminants ou est-ce que c'est autre chose. Par exemple est-ce que les humains sont plutôt attirés par le thème du film, l'acteur, ….

e. (Idée PM) : regarder un portrait "psychologique" : extraire l'intégralité des verbes des synopsis (play, cry, love), faire matcher ces verbes à un effet (positif - négatif - neutre) et en déduire si les humains sont plutôt "gentils/joyeux" ou "méchants/dépressifs" en fonction du temps. Ensuite, on pourrait chercher à classer le temps en fonction des périodes joyeuses aux US (chute de l'URSS, 30 glorieuses…) et tristes (11 septembre, 2e guerre mondiale) - ça peut éventuellement se faire avec une sorte de courbe de "bonheur" des gens, et regarder s'il y a une bonne corrélation ?

Ou et quand se passe l'action ? genre bcp aux US, et pdt certains événements, guerres

L'idée de cette partie est de présenter le dataset, et d'extraire des statistiques. Pour ORIGINE et PRÉNOM, c'est très simple. Pour MÉTIER des rôles joués, il y a déjà un mini-cluster qui a été fait et sur lequel on peut travailler pour la P2. Mais je pense que pour la P3, il faudra clairement agrandir le pool de données.

II. Cette image est-elle réaliste ?

a. Métier/Origine : comparaison des moyennes, possibilité de faire des intervalles de confiances et tout le bordel

b. Peut-on vraiment avoir une idée des prénoms populaires avec les films ? Est-ce que les films ont vraiment un impact sur le choix des prénoms ? Est-ce qu'ils reflètent ou anticipent la réalité ?

c (proposition Guillaume) : tenter des trucs élaborés sur un système de métrique de réalisme d'un pool de films (ex de critères : distributions (et pas seulement valeurs moyennes, si on y arrive), des : tailles, âges, ethnies, sexes, métiers, …), et noter des pools randoms de N films. Puis, enlever certaines catégories de films (ex, la SF), et reprendre des pools random et voir si ça correspond mieux.

Rq : On peut faire scanner automatiquement (via un script python) un synopsis à chatGPT pour 0.0010$. + d'infos ici (page pricing) et là (page API). Donc en gros si on veut scanner les 43000 scripts comme des bourrins pour en tirer des infos, on en a pour 50€ au total (en comptant le prix des réponses). Un truc malin serait de lui donner le nom des films et lui demander d'en sortir les infos directement (sans doute moins cher, aussi), mais ça casse le principe "les extraterrestres n'ont accès qu'à ce dataset".

Plan d'action (Proposition PM & Jean) :

A FAIRE AVANT LE RDV DE VENDREDI 10.11 :

- Extraire les statistiques du portrait robot sur les point suivants :
  - taille
  - distribution fréquentielle des prénoms (pas d'analyse temporelle, genre quand les prénoms sont donnés, je pense) (séparer entre prénoms des personnages et prénom des acteurs, sans doute personnages principaux/secondaires – selon leur présence dans le synopsis du film ?)
  - métier en se basant sur le cluster préexistant
- Extraire les mêmes informations pour l'humanité (sortir des références solides), et notamment avec le dataset des prénoms.

A FAIRE POUR LA P2 :

- Présenter un portrait robot issu de l'ensemble des films du dataset et le comparer au portrait robot de la réalité. En tirer quelques conclusions simples. Notamment les films ne reflètent pas la vérité, sans doute parce que certains thèmes sont beaucoup trop "extrapolés" (SF) ou trop anciens (regarder l'impact des quotas de diversité - pour les afro-américains ça apparaît en 1970 par exemple), et ça nuit à la représentativité. Proposer d'explorer une métrique pour étudier la représentativité d'un pool de films

- Faire une étude causale :

A FAIRE POUR LA P3 :

- Proposer une métrique associée à un pool de films sur les points suivants : ethnie simplifiée (parce que par pays ça n'a pas de sens), sexe, taille, métiers, age.
- Sampler des échantillons de 1000 films, évaluer la métrique et regarder si en supprimant des catégories on serait pas plus représentatif.

Liens peut-être utiles :

- dataset public des films concernés
- plein de truc sur les films : https://data.world/datasets/movie
-
-
-

P2 deliverable (done as a team): GitHub repository with the following:

Readme.md file containing the detailed project proposal (up to 1000 words). Your README.md should contain:

- Title

- Abstract: A 150 word description of the project idea and goals. What's the motivation behind your project? What story would you like to tell, and why?

- Research Questions: A list of research questions you would like to address during the project.

- Proposed additional datasets (if any): List the additional dataset(s) you want to use (if any), and some ideas on how you expect to get, manage, process, and enrich it/them. Show us that you've read the docs and some examples, and that you have a clear idea on what to expect. Discuss data size and format if relevant. It is your responsibility to check that what you propose is feasible.

- Methods

- Proposed timeline

- Organization within the team: A list of internal milestones up until project Milestone P3.

- Questions for TAs (optional): Add here any questions you have for us related to the proposed project.

Notebook containing initial analyses and data handling pipelines. We will grade the correctness, quality of code, and quality of textual descriptions.