Documentation Technique T-DAT-902

Introduction :

Le projet Nostradamovies consiste à prédire les genres de films à partir d'un ensemble de données comprenant 80 000 affiches de films, des synopsis et des informations complètes provenant du site IMDB. L'objectif est de développer un modèle capable de prédire les genres de films uniquement à partir des affiches. Le projet comprend également la normalisation des genres existants dans l'ensemble de données et la possibilité d'ajouter de nouveaux genres personnalisés. De plus, une visualisation statistique et un outil interactif doivent être fournis, ainsi qu'une méthodologie détaillant les algorithmes et les caractéristiques utilisées pour les extractions et les prédictions.

1. Prétraitement des données :

- Nettoyage de l'ensemble de données : suppression des valeurs manquantes, des doublons, etc.

- Normalisation des genres existants : remplacement des genres existants par des genres normalisés (par exemple, remplacer "comédie horreur" par "comédie" ou "horreur" uniquement).

- Ajout de nouveaux genres personnalisés : possibilité d'ajouter de nouveaux genres qui ne sont pas présents dans l'ensemble de données d'origine (par exemple, "blockbuster", "teen movie", "Cannes Palme d'Or", etc.).

- Extraction des caractéristiques des affiches et des synopsis : identification des caractéristiques importantes pour les genres de films, telles que les couleurs, les éléments visuels, la taille des titres, etc.

2. Modélisation :

- Utilisation de réseaux de neurones et d'apprentissage profond : utilisation de bibliothèques éprouvées pour développer un modèle de prédiction de genres basé sur les affiches des films.

- Utilisation de techniques d'apprentissage supervisé : entraînement du modèle à partir des données étiquetées avec les genres de films correspondants.

- Utilisation de techniques d'apprentissage non supervisé : application de clustering et d'analyses non supervisées pour extraire des caractéristiques importantes des affiches et des synopsis.

- Utilisation de SHAP values : utilisation des valeurs SHAP pour évaluer l'importance des caractéristiques extraites des affiches et des synopsis.

3. Visualisation et statistiques :

- Création d'un document synthétisant la visualisation et les statistiques des données.

- Utilisation d'un outil interactif pour la visualisation des données, similaire à l'exemple fourni.

- Affichage des données pertinentes considérées comme significatives pour l'analyse des genres de films.

- Utilisation de clustering et d'analyses non supervisées pour visualiser les caractéristiques importantes extraites des affiches et des synopsis.

4. Extraction des affiches :

- Détermination des caractéristiques les plus importantes pour chaque genre de film.

- Sélection de l'affiche la plus typique pour chaque genre en fonction de l'importance des caractéristiques.

- Création d'un algorithme capable de sélectionner l'affiche contenant le plus grand nombre d'éléments importants, dans l'ordre d'importance, par exemple, l'affiche adjacente.

5. Prédiction des genres pour de nouvelles affiches :

- Développement d'une fonction capable de traiter une image PNG ou JPEG et de prédire les genres de films correspondants.

- Affichage des trois prédictions les plus probables avec les probabilités associées.

- Extraction des caractéristiques de l'affiche et calcul de leur importance pour le genre prédit.

6. Technologies utilisées :

a. Langage de programmation :

- *Python* : utilisé comme langage principal pour le développement du projet en raison de sa richesse en bibliothèques et de sa facilité d'utilisation.

b. Bibliothèques et frameworks :

- *TensorFlow* : utilisé pour l'implémentation de réseaux de neurones et d'apprentissage profond pour la prédiction des genres de films.
- *Keras* : une interface de haut niveau pour TensorFlow, utilisée pour la construction et l'entraînement des modèles de prédiction.

- *OpenCV* : utilisé pour le traitement des images, notamment pour extraire des caractéristiques des affiches.

- *NumPy* : utilisé pour les opérations mathématiques et le traitement des tableaux de données.

- *Pandas* : utilisé pour la manipulation et l'analyse des données tabulaires.

- *Matplotlib* : utilisé pour la visualisation des données sous forme de graphiques et de diagrammes.

- *Scikit-learn* : utilisé pour l'implémentation de techniques de clustering et d'analyses non supervisées.

- *SHAP* (SHapley Additive exPlanations) : utilisé pour évaluer l'importance des caractéristiques extraites des affiches et des synopsis.

7. Outils utilisés :

a. IDE (Environnement de développement intégré) :

- *Jupyter Notebook* : utilisé pour le développement interactif et la création de rapports, en combinant du code exécutable, des visualisations et des explications.

b. Système de contrôle de version :

- Git : utilisé pour le suivi des modifications du code source, la gestion des versions et la collaboration avec une équipe.

c. Stockage et gestion des données :

- *Base de données* : utilisée pour stocker les données de l'ensemble de données, y compris les affiches de films, les synopsis et les informations IMDB.

- *Système de fichiers* : utilisé pour stocker les fichiers d'images des affiches de films.

d. Outils de visualisation :

- *Tableau de bord interactif* : utilisé pour créer une visualisation interactive des statistiques et des caractéristiques des données, permettant aux utilisateurs d'explorer les informations de manière conviviale.

e. Outils de documentation :

- *Diagrammes UML* : utilisés pour représenter l'architecture et les interactions des différents composants du système.

8. Méthodologie :

La méthodologie adoptée pour la réalisation du projet comprend les étapes suivantes :

a. Prétraitement des données :

- Nettoyage des données en supprimant les valeurs manquantes, les doublons, etc.
- Normalisation des genres existants dans l'ensemble de données et ajout de nouveaux genres personnalisés.
- Extraction des caractéristiques des affiches et des synopsis.

b. Modélisation :

- Utilisation de TensorFlow et de Keras pour développer un modèle de prédiction de genres basé sur les affiches de films.
- Utilisation de techniques d'apprentissage supervisé et non supervisé pour l'entraînement du modèle et l'extraction des caractéristiques.

c. Visualisation et statistiques :

- Utilisation de bibliothèques telles que Matplotlib pour créer des visualisations interactives des statistiques et des caractéristiques des données.