Analyse de données de films

BOURAI Assia & BENMESSAOUD Hamza
Sorbonne université, UPMC, Informatique, 3I026

Introduction

Dans ce projet nous nous intéressons à l'analyse de données de films tirées de deux base de données (MovieLens/Tmdb) des films, On essayera dans ce projet de toucher au trois points suivants:

- Visualisation de données,
- Classification supervisée,
- Classification non-supervisée.

Organisation des informations

Afin d'exploiter au maximum les deux bases de données fournies, nous avions sélectionné les informations qui nous semblaient les plus pertinentes.
Pour la suite du traitement nous avons choisi les information suivantes:

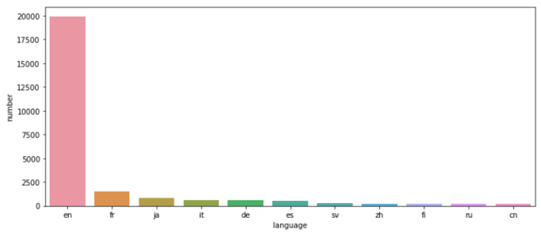
- L'identifiant du film,
- Le titre,
- Le genre du film (One Hot Coding [Action, Adventure, Animation, Children, Comedy, Crime, Documentary, Drama, Fantasy, Film-noir, Horror, Musical, Mystery, Romance, Sci-fi, Thriller, War, Western]
- L'acteur principal,
- Le réalisateur,
- La popularité,
- Le langage original du film,
- Le titre original,
- La moyenne de vote,
- Le nombre de vote,
- La moyenne d'évaluation,
- Le nombre d'évaluation.

Visualisation de données

Dans cette section nous allons avoir une visualisation des données que nous avons citées en haut.

1) Représentation de la langue dans la base

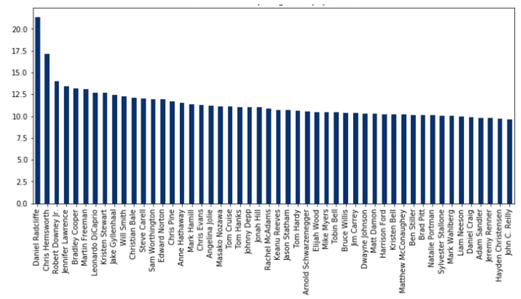
On remarque que la langue la plus représentée de la base est l'anglais avec plus de 19000 films dont la langue originale est l'anglais.



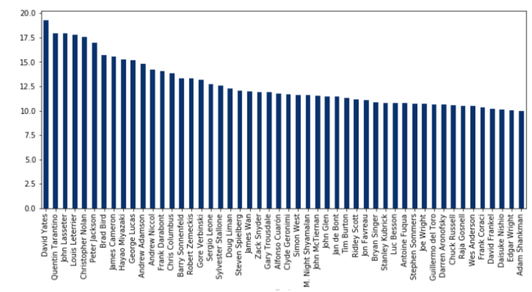
2) Meilleurs films par rapport au nombre de vote

15336	Inception (2010)	21060	2010-07-15
17625	Avengers, The (2012)	18191	2012-04-25
12397	Dark Knight, The (2008)	17998	2008-07-16
14423	Avatar (2009)	17817	2009-12-10
22459	Interstellar (2014)	17521	2014-11-05
23241	Guardians of the Galaxy (2014)	16698	2014-07-30
2833	Fight Club (1999)	15417	1999-10-15
19786	Django Unchained (2012)	14902	2012-12-25
12503	Iron Man (2008)	14399	2008-04-30
292	Pulp Fiction (1994)	14296	1994-09-10

3) Acteurs avec la plus grande popularité

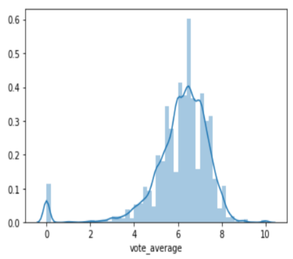


4) Réalisateurs avec la plus grande popularité



5) Distribution des notes de films

On remarque que la majorité les films de notre base ont une note entre 6 et 8, la Distribution est représentée par la courbe ci-contre

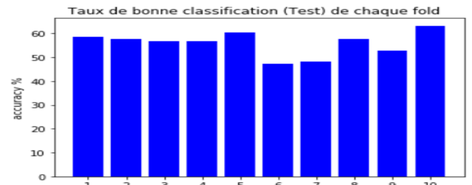


Classification supervisée

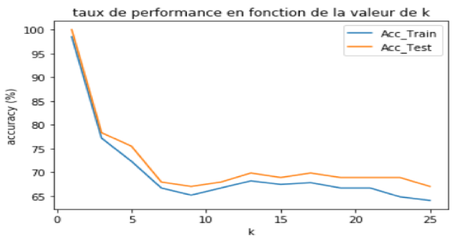
Dans cette section, on essayera de prédire si un film appartient a un genre donnée ou pas, À partir de la moyenne et du nombre des notes attribuées, ainsi que la moyenne et le nombre de vote, le genre en question ici est "DRAMA" car il est le plus représenté dans notre base.

1) Algorithme KNN

Avec Cross Validation nous avons eu les résultats suivants:
Taux de bonne classification (Test) = 55%
Ecart Type = 4,81

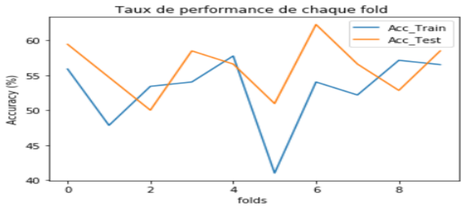


En modifiant le K nous avons eu les résultats suivants:
Taux de bonne classification (Test) = 72%



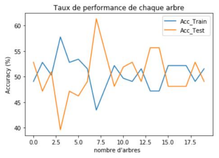
2) Classifieur Perceptron Grandient Sto

Avec Cross Validation nous avons eu les résultats suivants:
Taux de bonne classification (Test) = 56%
Ecart Type = 3,70

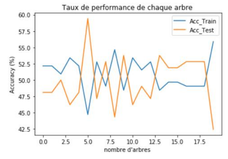


3) Arbre de décision

Arbre de décision Simple
Taux de bonne classification (Test) = 50%
Ecart type = 4,43

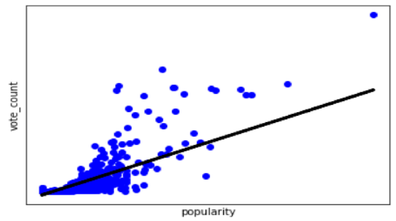


Arbre de décision Bagging
Taux de bonne classification (Test) = 49%
Ecart type = 3,88



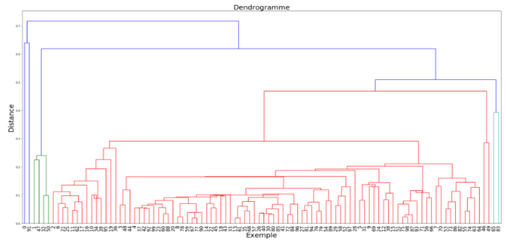
4) Régression linéaire

Nous allons essayer de prédire le nombre de vote par rapport a la popularité d'un film en utilisant la régression linéaire.
La droite noire correspond à la prédiction tandis que les points sont les vraies valeurs, nous avons obtenu un score de coefficient de détermination de 0,53 sachant que le meilleur score possible est de 1.



Classification non supervisée

Nous avons regrouper un échantillon de 100 films avec l'aide du Clusturing hiérarchique, Le premier dendrogramme représente les clusters(chaque couleur) des films les plus proche par rapport aux notes qu'ils ont obtenu et leurs popularité.
Nous avons en tout 4 clusters, et on remarque que le plus grand cluster contient 92 films (couleur Rouge)



Le dendrogramme ci-dessous représente les clusters(chaque couleur) des films les plus proche par rapport au genre
Nous avons en tout 7 clusters, et on remarque que le plus grand cluster contient 67 films (couleur Rouge), on remarque aussi qu'il y a des film qui ont exactement le même genre représenté ici par des droite dans l'axe des abscisses 0,0 (par exemple entre 13 et 89)