

# **PROJET WEB**

# **SCRAPING**

**Présenté par :**

Mélanie Yakoub

Mounira Arbouch

Joseph Willson

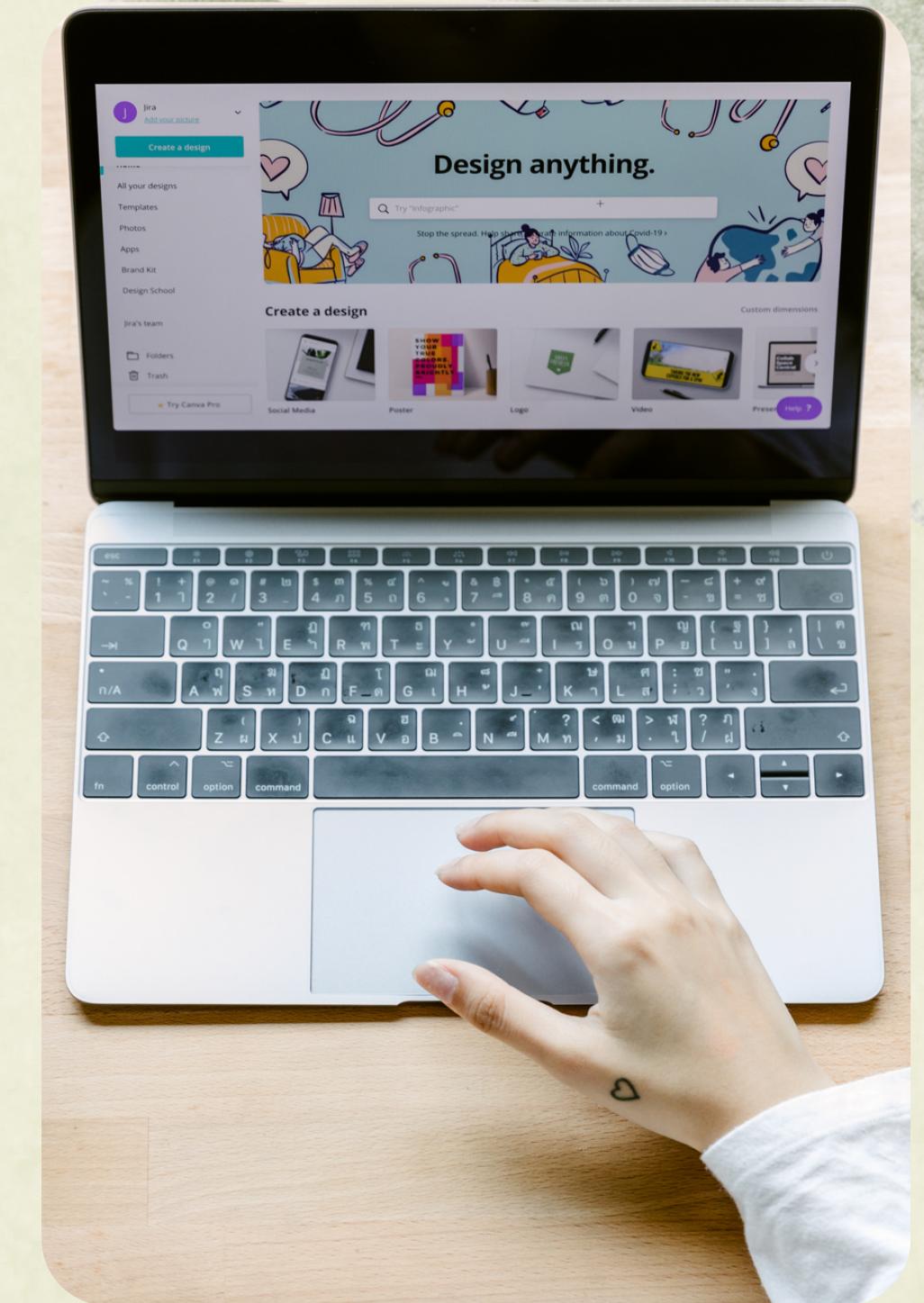
# TABLE OF CONTENTS

Introduction	----- 03
Web scrapping	----- 04 - 07
Feature engineering	----- 08 - 12
Analyse Statistique	----- 13 - 22
Filtrage et Modèle NLP	----- 23 - 30

# INTRODUCTION

Dans ce projet, nous avons décidé de faire du web scraping sur la page "Allociné" dans le but d'obtenir des informations pertinentes sur les films pour pouvoir les traiter.

À cette fin, une analyse minutieuse a été réalisée, suivie d'une modélisation afin de déterminer les genres de films que les personnes préfèrent.



# WEB SCRAPPING

Comme expliqué dans l'introduction, dans le cadre de ce projet, nous avons décidé de web scraper la page "Allociné".

Pour ce faire, nous avons récupéré les informations sur les films qui nous semblaient les plus importantes, que nous détaillons dans la suite de cette présentation.

## **Titre du film**

Le titre du film est évidemment un indispensable car il représente l'élément que le spectateur verra en premier.

## **Note des spectateurs et de la presse**

Ces variables nous ont semblé pertinentes car elles permettent d'évaluer le sentiment général des spectateurs ainsi que la tendance générale des genres de film.

## **Durée**

Une variable pertinente, surtout quand on sait aujourd'hui que certains sont prêts à voir Barbie qui dure 1h54 et d'autres qui préfèrent Oppenheimer qui dure 3h 😅.

## **Lien des bandes d'annonces**

Le lien de la bande-annonce est présent pour chaque film pour permettre aussi aux personnes d'avoir une idée de ce qu'ils pourraient retrouver dans le film.

## Réaliseurs

La variable "réalisateur" avec les noms des producteurs, utile dans la fonction de filtrage, nous permettant de retrouver les informations sur les films.

## Résumés

Cette variable nous permet d'avoir une idée du film que nous allons voir, comme les bandes d'annonce. À noter que cette variable a été fortement utilisée dans notre modèle de NLP que nous introduirons plus tard.

## Types / Genres

La variable "genre" est sans doute la variable la plus importante dans notre travail d'analyse, car elle nous permet de comprendre les préférences des spectateurs.

## Acteurs

Nous retrouvons les noms des acteurs qui ont joué dans les films. Cette variable est aussi très intéressante dans la fonction de filtrage que nous avons créée.

# WEB SCRAPPING : BASE

Titre	Réalisateur	Acteurs	Note Presse	Note Spectateurs	Types	Durée
Top Gun: Maverick	Joseph Kosinski	Tom Cruise, Miles Teller, Jennifer Connelly	3,7	4,3	Action	\n2h 11min\n
The Fabelmans	Steven Spielberg	Gabriel LaBelle, Michelle Williams, Paul Dano	4,9	4,3	Biopic, Drame	\n2h 31min\n
The First Slam Dunk	Takehiko Inoue	Shugo Nakamura, Subaru Kimura, Kenta Miyake	3,9	4,2	Animation, Comédie, Drame	\n2h 04min\n
Le Bleu du Caftan	Maryam Touzani	Lubna Azabal, Saleh Bakri, Ayoub Missiouï	3,8	4,2	Drame, Romance	\n2h 02min\n
Avatar : la voie de l'eau	James Cameron	Sam Worthington, Zoe Saldana, Sigourney Weaver	4,1	4,2	Science Fiction, Aventure, Fantastique	\n3h 12min\n
Suzume	Makoto Shinkai	Lévanah Solomon, Nanoka Hara, Benjamin Jungers	3,8	4,2	Animation, Aventure, Drame	\n2h 02min\n
Le Retour des hirondelles	Li Ruijun	Wu Renlin, Hai-Qing	4,1	4,2	Drame	\n2h 13min\n
Le Chat Potté 2 : la dernière quête	Januel P. Mercado, Joel Crawford	Boris Rehlinger, Antonio Banderas, Diane Dassigny	3,1	4,2	Animation, Comédie, Aventure	\n1h 42min\n
Babylon	Damien Chazelle	Brad Pitt, Margot Robbie, Diego Calva	4,0	4,2	Historique, Drame	\n3h 09min\n
As bestas	Rodrigo Sorogoyen	Denis Ménochet, Marina Foïs, Luis Zahera	4,2	4,1	Thriller, Drame	\n2h 17min\n
Elvis	Baz Luhrmann	Austin Butler, Tom Hanks, Olivia DeJonge	3,8	4,1	Biopic, Musical	\n2h 39min\n
Presque	Bernard Campan, Alexandre Jollien	Tiphaine Daviot	3,5	4,1	Drame, Comédie	\n1h 32min\n
En corps	Cédric Klapisch	Marion Barbeau, Hofesh Shechter, Denis Podalydès	3,4	4,1	Comédie dramatique, Drame, Comédie	\n1h 58min\n
Treize vies	Ron Howard	Colin Farrell, Viggo Mortensen, Joel Edgerton	3,2	4,1	Drame	\n2h 27min\n
The Quiet Girl	Colm Bairéad	Carrie Crowley, Andrew Bennett, Catherine Clinch	3,7	4,1	Drame	\n1h 36min\n
Close	Lukas Dhont	Eden Dambrine, Gustav De Waele, Emilie Dequenne	3,3	4,1	Drame	\n1h 44min\n
About Kim Sohee	July Jung	Doona Bae, Kim Si-eun, Choi Hee-jin	3,8	4,1	Drame, Policier	\n2h 17min\n
Jujutsu Kaisen 0	Sung-ho Park	Alexandre N'Guyen, Megumi Ogata, Alice Orsat	3,9	4,1	Animation, Action, Drame	\n1h 45min\n
A l'Ouest, rien de nouveau	Edward Berger	Felix Kammerer, Albrecht Schuch, Aaron Hilmer	4,1	--	Drame, Guerre, Historique	\n2h 28min\n
Simone, le voyage du siècle	Olivier Dahan	Elsa Zylberstein, Rebecca Mader, Élodie Bouchez	3,1	4,1	Biopic, Drame	\n2h 20min\n

Résumé

\nAprès avoir été l'un des meilleurs pilotes de chasse de la Marine américaine, Pete "Maverick" Mitchell continue à repousser ses limites en tant que pilote d'essai. Il est chargé de former un détachement d'élite. \nPortrait profondément intime d'une enfance américaine au XXème siècle, The Fabelmans de Steven Spielberg nous plonge dans l'histoire familiale du cinéaste qui a façonné sa vie personnelle et professionnelle. \nLe meneur de jeu de Shohoku, Ryota Miyagi, joue toujours intelligemment et à la vitesse de l'éclair, contournant ses adversaires tout en gardant son sang-froid. Ryota fait partie de l'équipe de basket-ball de l'école secondaire de Shohoku. \nHalim est marié depuis longtemps à Mina, avec qui il tient un magasin traditionnel de caftans dans la médina de Salé, au Maroc. Le couple vit depuis toujours avec le secret d'Halim, son homosexualité qu'il cache à tous. \nSe déroulant plus d'une décennie après les événements relatés dans le premier film, AVATAR : LA VOIE DE L'EAU raconte l'histoire des membres de la famille Sully (Jake, Neytiri et leurs enfants), les éprouvant dans diverses situations. \nDans une petite ville paisible de Kyushu, une jeune fille de 17 ans, Suzume, rencontre un homme qui dit voyager à la recherche d'une porte. Décidant de le suivre dans les montagnes, elle découvre une portière qui semble être l'entrée d'un monde magique. \nC'est l'histoire d'un mariage arrangé, entre deux êtres méprisés par leurs familles. Entre eux, la timidité fait place à l'affection. Autour d'eux, la vie rurale se désagrège... \nLe Chat Potté découvre que sa passion pour l'aventure et son mépris du danger ont fini par lui coûter cher : il a épuisé huit de ses neuf vies, et en a perdu le compte au passage. Afin de retomber sur ses pieds, il décide de se faire passer pour un chat errant dans les rues de Los Angeles des années 1920. Récit d'une ambition démesurée et d'excès les plus fous, BABYLON retrace l'ascension et la chute de différents personnages lors de la création d'Hollywood, une ère de décadence et de folie. \nAntoine et Olga, un couple de Français, sont installés depuis longtemps dans un petit village de Galice. Ils ont une ferme et restaurent des maisons abandonnées pour faciliter le repeuplement. Tout devrait bien se passer jusqu'à ce qu'ils rencontrent un mystérieux visiteur. \nLa vie et l'œuvre musicale d'Elvis Presley à travers le prisme de ses rapports complexes avec son mystérieux manager, le colonel Tom Parker. Le film explorera leurs relations sur une vingtaine d'années, de leur première rencontre à leur séparation définitive. \nDeux hommes prennent la route, de Lausanne vers le sud de la France, dans un corbillard. Ils se connaissent peu, ont peu de choses en commun, du moins le croient-ils... \nElise, 26 ans, est une grande danseuse classique. Elle se blesse pendant un spectacle et apprend qu'elle ne pourra plus danser. Dès lors sa vie va être bouleversée, Elise va devoir apprendre à se réparer et à rebondir. \nUne équipe de sauvetage est réunie en Thaïlande où une dizaine de jeunes hommes et leur entraîneur sont piégés dans un système de grottes souterraines, peu à peu victimes de la montée des eaux. \nEn Irlande, 1981, Cáit, une jeune fille effacée et négligée par sa famille, est envoyée vivre auprès de parents éloignés pendant l'été. Mais dans cette maison en apparence sans secret, où elle trouve l'épanouissement, Cáit se sent mal à l'aise. \nLéo et Rémi, 13 ans, sont amis depuis toujours. Jusqu'à ce qu'un événement impensable les sépare. Léo se rapproche alors de Sophie, la mère de Rémi, pour essayer de comprendre... \nPour son stage de fin d'étude, Kim Sohee intègre un centre d'appel de Korea Telecom. En quelques mois, son moral décline sous le poids de conditions de travail dégradantes et d'objectifs de plus en plus élevés. \nLorsqu'il était enfant, Yuta Okkotsu a vu son amie Rika Orimoto perdre la vie dans un accident. Depuis, Rika vient hanter Yuta qui a même souhaité sa propre mort après avoir souffert des années de cette maladie. \nL'histoire poignante d'un jeune soldat allemand sur le front occidental pendant la Première Guerre mondiale. En première ligne, Paul et ses camarades voient l'euphorie initiale se muer en désespoir et en tristesse. \nLe destin de Simone Veil, son enfance, ses combats politiques, ses tragédies. Le portrait épique et intime d'une femme hors du commun qui a bousculé son époque en défendant un message humaniste.

# WEB SCRAPPING : BASE

Lien de bande-annonce

```
https://www.allocine.fr/film/fichefilm\_gen\_cfilm=186636.html
https://www.allocine.fr/film/fichefilm\_gen\_cfilm=255726.html
https://www.allocine.fr/film/fichefilm\_gen\_cfilm=306133.html
https://www.allocine.fr/film/fichefilm\_gen\_cfilm=297895.html
https://www.allocine.fr/film/fichefilm\_gen\_cfilm=178014.html
https://www.allocine.fr/film/fichefilm\_gen\_cfilm=299413.html
https://www.allocine.fr/film/fichefilm\_gen\_cfilm=300250.html
https://www.allocine.fr/film/fichefilm\_gen\_cfilm=228395.html
https://www.allocine.fr/film/fichefilm\_gen\_cfilm=275675.html
https://www.allocine.fr/film/fichefilm\_gen\_cfilm=297016.html
https://www.allocine.fr/film/fichefilm\_gen\_cfilm=228681.html
https://www.allocine.fr/film/fichefilm\_gen\_cfilm=284465.html
https://www.allocine.fr/film/fichefilm\_gen\_cfilm=287738.html
https://www.allocine.fr/film/fichefilm\_gen\_cfilm=282400.html
https://www.allocine.fr/film/fichefilm\_gen\_cfilm=299438.html
https://www.allocine.fr/film/fichefilm\_gen\_cfilm=294372.html
https://www.allocine.fr/film/fichefilm\_gen\_cfilm=303802.html
https://www.allocine.fr/film/fichefilm\_gen\_cfilm=291677.html
https://www.allocine.fr/film/fichefilm\_gen\_cfilm=182953.html
https://www.allocine.fr/film/fichefilm\_gen\_cfilm=271339.html
```

# DÉTAILS SUR LE WEB SCRAPING

Nous sommes repartis des méthodes vues en cours et avons généralisé ces différentes méthodes afin d'obtenir toutes les informations nécessaires.

Des fonctions telles que **html** ou **xpath** de la bibliothèque **lxml** ont été très utilisées, mais pas seulement. En effet, lors de notre web scraping, la bibliothèque **bs4** a été utilisée et a été fortement appréciée pour la récupération de liens pour les bandes-annonces.

# FEATURE ENGINEERING

La partie feature engineering est l'une des parties les plus importantes dans ce projet car elle nous a d'une part permis de comprendre le jeu de données récolté, mais aussi de nettoyer la base et de la rendre plus interprétable pour pouvoir commencer notre analyse.

Nous avons commencé par nettoyer notre jeu de données. Pour cela, des méthodes vues en cours, notamment avec des méthodes regex, ont été utilisées.

# En voici un exemple avec la colonne durée :

AVANT

Durée
\n2h 11min\n
\n2h 31min\n
\n2h 04min\n
\n2h 02min\n
e   \n3h 12min\n
\n2h 02min\n
\n2h 13min\n
\n1h 42min\n
\n3h 09min\n
\n2h 17min\n
\n2h 39min\n
\n1h 32min\n

APRÈS

Durée
2h 11min
2h 31min
2h 04min
2h 02min
3h 12min
2h 02min
2h 13min
1h 42min
3h 09min
2h 17min
2h 39min
1h 32min

## Mais aussi...

La création de nouvelles colonnes nous permettant de faciliter l'analyse de données est très indispensable, essentiellement pour les méthodes de filtrage, mais pas seulement.



# Un exemple avec la colonne types :

AVANT

Types
Action
Biopic, Drame
Animation, Comédie, Drame
Drame, Romance
Science Fiction, Aventure, Fantastique

APRÈS

Type_1	Type_2	Type_3
Action	NULL	NULL
Biopic	Drame	NULL
Animation	Comédie	Drame
Drame	Romance	NULL
Science Fiction	Aventure	Fantastique

À noter que cette variable est utilisée dans notre modèle de NLP.

# D'AUTRES MÉTHODES

Nous avons, par une imputation, remplacé les valeurs manquantes de la colonne "Note spectateurs" en les remplaçant par la moyenne de la colonne.

De plus, à la suite de cela, nous avons créé une nouvelle colonne appelée "Note moyenne" qui est la moyenne entre la note presse et la note spectateur dans le but de créer une colonne "Catégorie Générale".

Dans cette colonne, nous avons fixé des seuils à partir de la note minimale et maximale pour catégoriser les films en "faible", "moyen" et "bien".

# ANALYSE STATISTIQUE

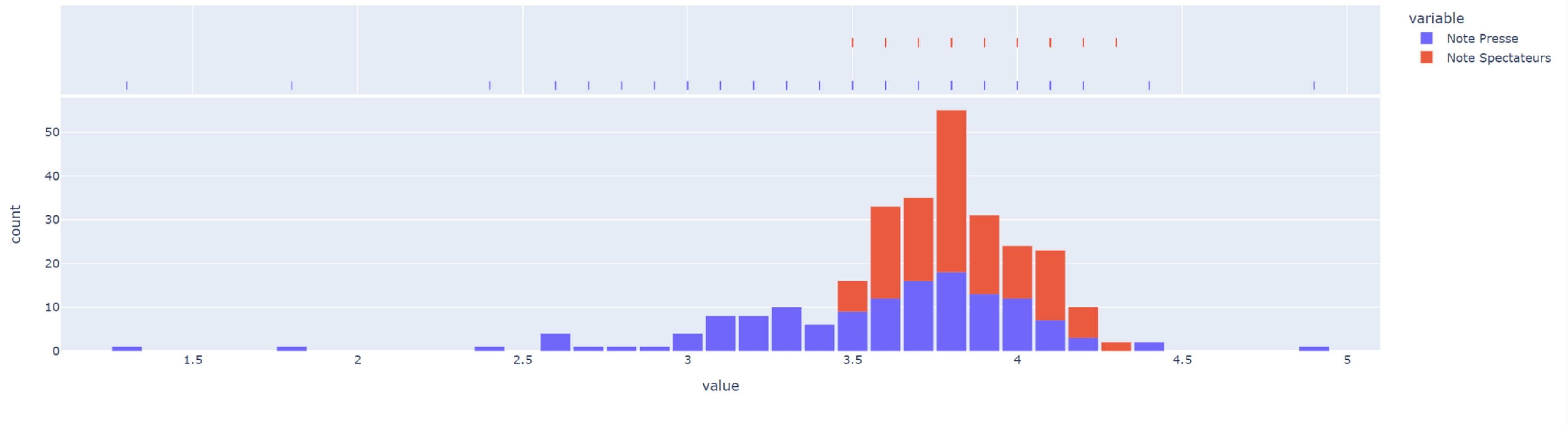
Dans cette partie, nous avons analysé les données. Basées sur les informations qui nous sont données, nous avons pu tirer certaines conclusions.

Ces dernières nous ont permis de comprendre le comportement des spectateurs et leurs préférences.

# Analyse globale des notes

Note globalement bonne

Distribution des notes



Comme le montre le graphique, la distribution des notes semble être normale et est centrée autour de 3,8.

# Analyse du nombre de films par type

On observe que le type de film le plus présent dans cette dataframe est sans aucun doute les films de type drame.

Unique_Word	Frequency
Judiciaire	1
Romance	13
Thriller	19
Famille	7
Historique	10
Comédie	23
Epouvante-horreur	1
Aventure	11
Fantastique	7
Comédie dramatique	10
Policier	11
Musical	2
Animation	15
Biopic	12
Science Fiction	2
Guerre	4
Drame	98
Évènement Sportif	1
Action	13

# Analyse globale des notes par catégorie



En fonction des seuils que nous avons fixés, nous pouvons voir que les notes sont répertoriées essentiellement dans la catégorie "moyen".

# Aperçu globale

On s'aperçoit que globalement, les films les plus appréciés sont les films d'action et d'animation.

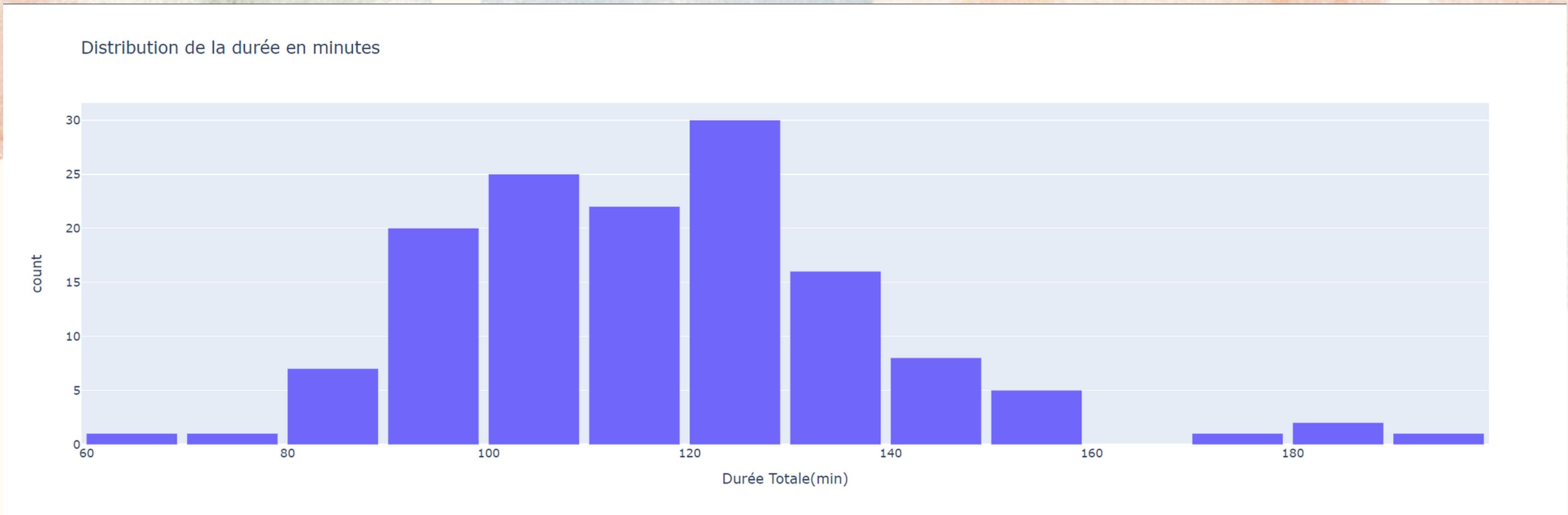
Type_1	Note_Moyenne_Par_Type1
Romance	3.775
Thriller	3.7600000000000002
Historique	3.46
Comédie	3.4555555555555553
Aventure	3.3
Fantastique	3.4
Comédie dramatique	3.6799999999999997
Policier	3.7
Animation	3.8266666666666667
Biopic	3.8125
Science Fiction	3.8
Guerre	3.7
Drame	3.7030303030303022
Action	3.84

# Aperçu particulier

Type_1	Moyenne_Note_Spectateurs	Moyenne_Note_Presse
Romance	3.774999976158142	3.80000011920929
Thriller	3.9	3.609999966621399
Historique	3.9199999809265136	3.0199999809265137
Comédie	3.6888888941870794	3.244444396760729
Aventure	3.599999046325684	3.04999952316284
Fantastique	3.599999046325684	3.20000047683716
Comédie dramatique	3.830000190734864	3.50999990463257
Policier	3.599999046325684	3.79999952316284
Animation	3.95999942779541	3.68000003178914
Biopic	3.94999988079071	3.6874999701976776
Science Fiction	3.849999046325684	3.849999046325684
Guerre	3.79999952316284	3.599999046325684
Drame	3.804545420588869	3.598484844872446
Action	3.920000286102297	3.760000381469725

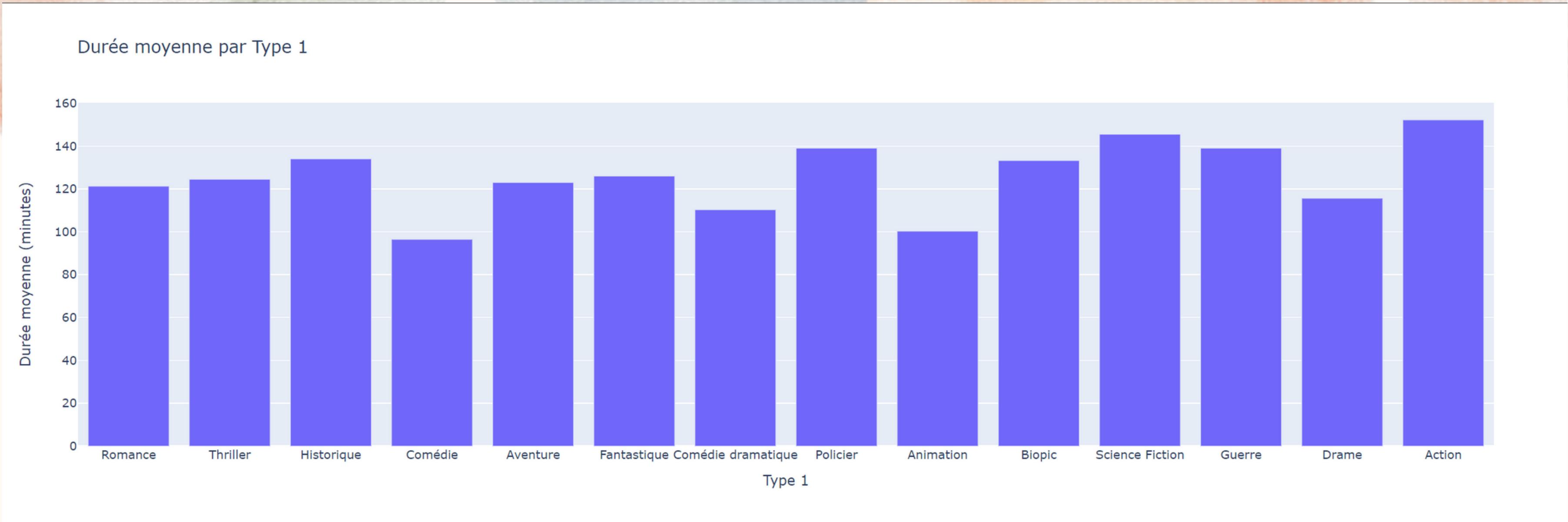
Cependant, quand nous regardons en détail, on peut voir que les spectateurs ont aussi une préférence pour les films historiques, contrairement à la presse.

# Parlons de durée



De même que dans l'analyse globale des notes, la distribution de la courbe ci-dessus semble être normale, et nous pouvons observer que les personnes n'ont aucun problème à regarder des films qui durent environ 2h. Cependant, les films avec une durée supérieure à 2h sont moins appréciés.

# Parlons de durée



Globalement et quel que soit le type de film, ils durent tous en moyenne 2h. Les films d'action ont une durée un peu plus longue.

# FILTRAGE ET MODÈLE NLP

Dans cette partie, nous allons présenter une fonction que nous avons implémentée et que vous pourrez retrouver dans le notebook, permettant de filtrer à votre guise.

Ainsi qu'un modèle de prédiction NLP basé sur les résumés des films qui permet d'associer chaque film à son type.

# FILTRAGE

```
def filter_and_display(colonne, director_name, df_select):
    # Filtrer le DataFrame en fonction du réalisateur
    filtered_df = df_select.filter(col(colonne) == director_name)

    # Afficher les résultats
    if filtered_df.count() == 0:
        print(f"Aucun film trouvé pour le {colonne} {director_name}.")
    else:
        print(f"Films du {colonne} {director_name}:")
        filtered_df.show(truncate=False)
```

La fonction ci-dessus est très simple. Vous mettez la colonne sur laquelle vous souhaitez filtrer, l'individu, le titre ou tout autre chose présente dans cette colonne, et une sélection des colonnes que vous souhaitez voir.

# UN EXEMPLE

```
# Exemple d'utilisation  
filter_and_display("Types", "Drame", df_final.select("Titre", "Réalisateur", "Acteurs", "Durée", "Lien de bande-annonce"))
```

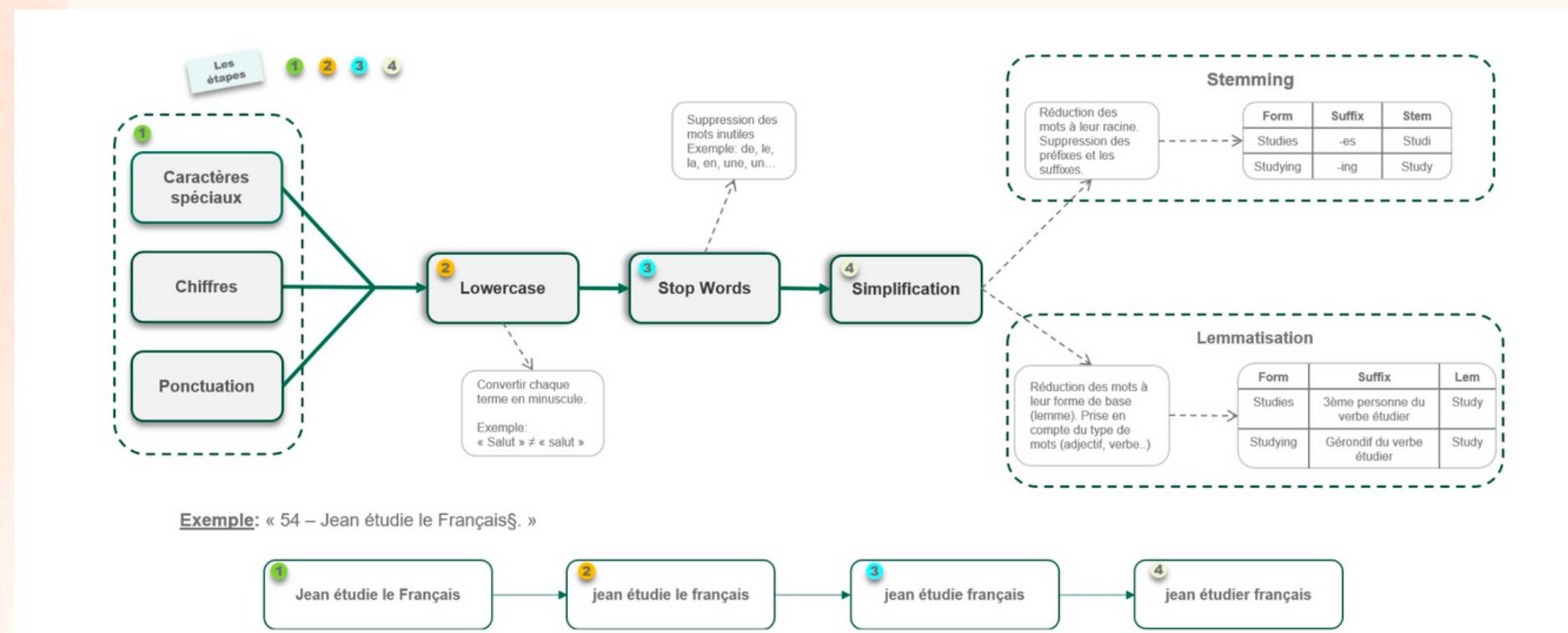
Dans la ligne de code, nous voulons filtrer sur la colonne "Types", avoir tous les films de type drame et en sortie, nous voulons les colonnes : "Titre", "Réalisateur", "Acteurs", "Durée", "Lien de bande-annonce"

# OUTPUT

Titre	Réalisateur	Acteurs	Durée	Lien de bande-annonce
Le Retour des hirondelles	Li Ruijun	Wu Renlin, Hai-Qing	2h 13min	<a href="https://www.allocine.fr/film/fichefilm_gen_cfilm=300250.htm">https://www.allocine.fr/film/fichefilm_gen_cfilm=300250.htm</a>
Treize vies	Ron Howard	Colin Farrell, Viggo Mortensen, Joel Edgerton	2h 27min	<a href="https://www.allocine.fr/film/fichefilm_gen_cfilm=282400.htm">https://www.allocine.fr/film/fichefilm_gen_cfilm=282400.htm</a>
The Quiet Girl	Colm Bairéad	Carrie Crowley, Andrew Bennett, Catherine Clinch	1h 36min	<a href="https://www.allocine.fr/film/fichefilm_gen_cfilm=299438.htm">https://www.allocine.fr/film/fichefilm_gen_cfilm=299438.htm</a>
Close	Lukas Dhont	Eden Dambrine, Gustav De Waele, Emilie Dequenne	1h 44min	<a href="https://www.allocine.fr/film/fichefilm_gen_cfilm=294372.htm">https://www.allocine.fr/film/fichefilm_gen_cfilm=294372.htm</a>
Leila et ses frères	Saeed Roustaee	Taraneh Alidoosti, Navid Mohammadzadeh, Payman Maadi	2h 39min	<a href="https://www.allocine.fr/film/fichefilm_gen_cfilm=303686.htm">https://www.allocine.fr/film/fichefilm_gen_cfilm=303686.htm</a>
The Whale	Darren Aronofsky	Brendan Fraser, Sadie Sink, Ty Simpkins	1h 57min	<a href="https://www.allocine.fr/film/fichefilm_gen_cfilm=289056.htm">https://www.allocine.fr/film/fichefilm_gen_cfilm=289056.htm</a>
Les Nageuses	Sally El Hosaini	Nathalie Issa, Manal Issa, Matthias Schweighöfer	2h 14min	<a href="https://www.allocine.fr/film/fichefilm_gen_cfilm=292214.htm">https://www.allocine.fr/film/fichefilm_gen_cfilm=292214.htm</a>
Les Huit Montagnes	Charlotte Vandermeersch, Felix Van Groeningen	Luca Marinelli, Alessandro Borghi, Filippo Timi	2h 27min	<a href="https://www.allocine.fr/film/fichefilm_gen_cfilm=290592.htm">https://www.allocine.fr/film/fichefilm_gen_cfilm=290592.htm</a>
Revoir Paris	Alice Winocour	Virginie Efira, Benoît Magimel, Grégoire Colin	1h 45min	<a href="https://www.allocine.fr/film/fichefilm_gen_cfilm=290439.htm">https://www.allocine.fr/film/fichefilm_gen_cfilm=290439.htm</a>
Le Sixième enfant	Léopold Legrand	Sara Giraudeau, Benjamin Lavernhe, Judith Chemla	1h 32min	<a href="https://www.allocine.fr/film/fichefilm_gen_cfilm=289527.htm">https://www.allocine.fr/film/fichefilm_gen_cfilm=289527.htm</a>
Le Serment de Pamfir	Dmytro Sukholtykyy-Sobchuk	Oleksandr Yatsentyuk, Stanislav Potiak, Solomiya Kyrylova	1h 42min	<a href="https://www.allocine.fr/film/fichefilm_gen_cfilm=285787.htm">https://www.allocine.fr/film/fichefilm_gen_cfilm=285787.htm</a>
The Son	Florian Zeller	Hugh Jackman, Laura Dern, Vanessa Kirby	2h 03min	<a href="https://www.allocine.fr/film/fichefilm_gen_cfilm=291247.htm">https://www.allocine.fr/film/fichefilm_gen_cfilm=291247.htm</a>
Le Tourbillon de la vie	Olivier Treiner	Lou de Laâge, Raphaël Personnaz, Isabelle Carré	2h 01min	<a href="https://www.allocine.fr/film/fichefilm_gen_cfilm=288544.htm">https://www.allocine.fr/film/fichefilm_gen_cfilm=288544.htm</a>
Les Bonnes étoiles	Hirokazu Kore-edo	Song Kang-Ho, Dong-won Gang, Doona Bae	2h 09min	<a href="https://www.allocine.fr/film/fichefilm_gen_cfilm=286075.htm">https://www.allocine.fr/film/fichefilm_gen_cfilm=286075.htm</a>
Argentina, 1985	Santiago Mitre	Ricardo Darín, Alejandra Flechner, Paula Ransenberg	2h 20min	<a href="https://www.allocine.fr/film/fichefilm_gen_cfilm=295782.htm">https://www.allocine.fr/film/fichefilm_gen_cfilm=295782.htm</a>
Notre-Dame brûle	Jean-Jacques Annaud	Samuel Labarthe, Jean-Paul Bordes, Mikaël Chirinian	1h 50min	<a href="https://www.allocine.fr/film/fichefilm_gen_cfilm=284864.htm">https://www.allocine.fr/film/fichefilm_gen_cfilm=284864.htm</a>
R.M.N.	Cristian Mungiu	Marin Grigore, Judith State, Macrina Bârlădeanu	2h 05min	<a href="https://www.allocine.fr/film/fichefilm_gen_cfilm=299722.htm">https://www.allocine.fr/film/fichefilm_gen_cfilm=299722.htm</a>
War Pony	Gina Gammell, Riley Keough	Stanley Good Voice Elk, Jojo Bapteise Whiting, Steven Yellow Hawk	1h 54min	<a href="https://www.allocine.fr/film/fichefilm_gen_cfilm=303685.htm">https://www.allocine.fr/film/fichefilm_gen_cfilm=303685.htm</a>
Blue Jean	Georgia Oakley	Rosy McEwen, Kerrie Hayes, Lucy Halliday	1h 37min	<a href="https://www.allocine.fr/film/fichefilm_gen_cfilm=306743.htm">https://www.allocine.fr/film/fichefilm_gen_cfilm=306743.htm</a>
Petites	Julie Lerat-Gersant, François Roy (III)	Pili Groyne, Romane Bohringer, Victoire Du Bois	1h 30min	<a href="https://www.allocine.fr/film/fichefilm_gen_cfilm=295064.htm">https://www.allocine.fr/film/fichefilm_gen_cfilm=295064.htm</a>

# MODÈLE NLP

Avant de vectoriser le résumé de chaque film (conversion de données textuelles en représentations numériques), nous avons créé une fonction de nettoyage textuel.



```
def clean_text_french(text):
    # Convertir le texte en minuscules
    text = text.lower()

    # Supprimer les accents, les apostrophes, et les guillemets
    text = unidecode(text).replace("'", "").replace("\"", "")

    # Supprimer la ponctuation
    text = text.translate(str.maketrans("", "", string.punctuation))

    # Supprimer les chiffres sauf ceux suivis de quatre chiffres (années)
    text = re.sub(r'\b\d{1,3}\b|(?<!\\d)\d{5},(?!\d)', '', text)

    # Tokenization (division en mots)
    words = word_tokenize(text, language='french')

    # Supprimer les stopwords en français
    stop_words = set(get_stop_words('fr'))
    #words = [word for word in words if word not in stop_words]

    # Stemming (racinisation) en français
    stemmer = SnowballStemmer('french')
    words = [stemmer.stem(word) for word in words]

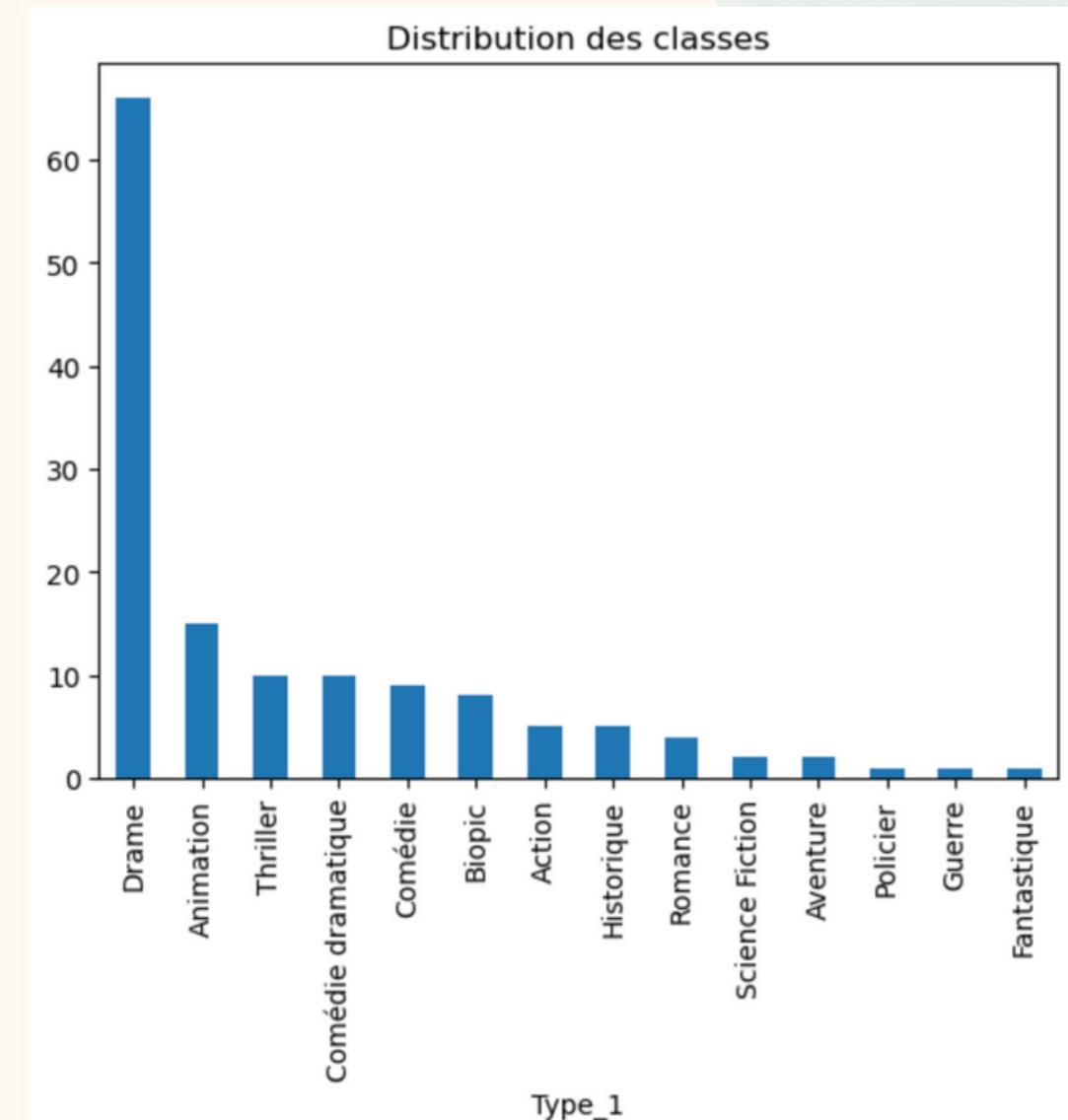
    # Rejoindre les mots pour reformer le texte
    cleaned_text = ' '.join(words)

    return cleaned_text
```

# DÉSÉQUILIBRE

Face au problème de déséquilibre des données et à l'échantillon faible, nous avons utilisés deux stratégies :

- 1. Utilisation d'un paramètre pour équilibrer les poids dans le modèle de prédiction Random Forest.
- 2. Suppression aléatoire de 40 échantillons de drames pour rééquilibrer les données (meilleure stratégie).



# VECTORISATION

Nous avons tester deux techniques différentes de vectorisation :

- Méthode 1 : TF IDF (paramétré par un grid search)

Formule:

$$\begin{aligned} \text{TF - IDF} &= \text{TF(mot, document)} \times \text{IDF} \\ &= \text{Nombre d'occurrence du mot dans le document} \times \log\left(\frac{\text{nombre de document}}{\text{idf}}\right) \end{aligned}$$

```
# Create TF-IDF vectorizer
tfidf_vectorizer = TfidfVectorizer(norm='l2', ngram_range=(1, 3), max_df=0.9, min_df=1,
#max_df=0.9, min_df=1,)

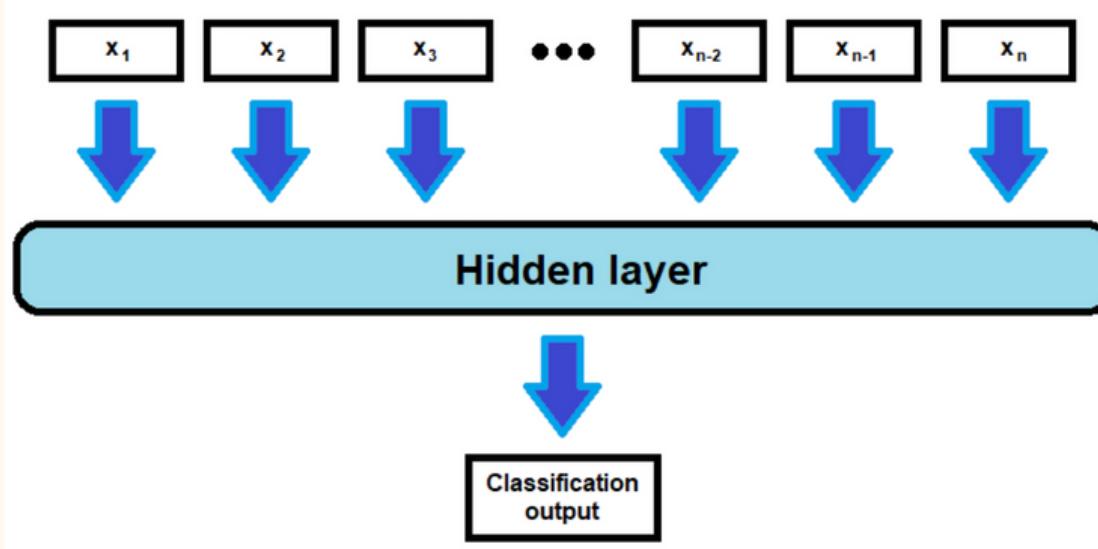
# Fit and transform the data
tfidf_matrix = tfidf_vectorizer.fit_transform(df_nouveau_resume_nlp['Résumé propre'])

# Get feature names (words)
feature_names = tfidf_vectorizer.get_feature_names_out()

# Convert the TF-IDF matrix to a DataFrame
tfidf_df = pd.DataFrame(tfidf_matrix.toarray(), columns=feature_names)

# Display the TF-IDF DataFrame
tfidf_df.head()
```

- Méthode 2 : FastText développé par Facebook AI Research (FAIR)



```
# Creation d'un modèle fasttext sur le corpus
from gensim.models import FastText
from gensim.utils import simple_preprocess

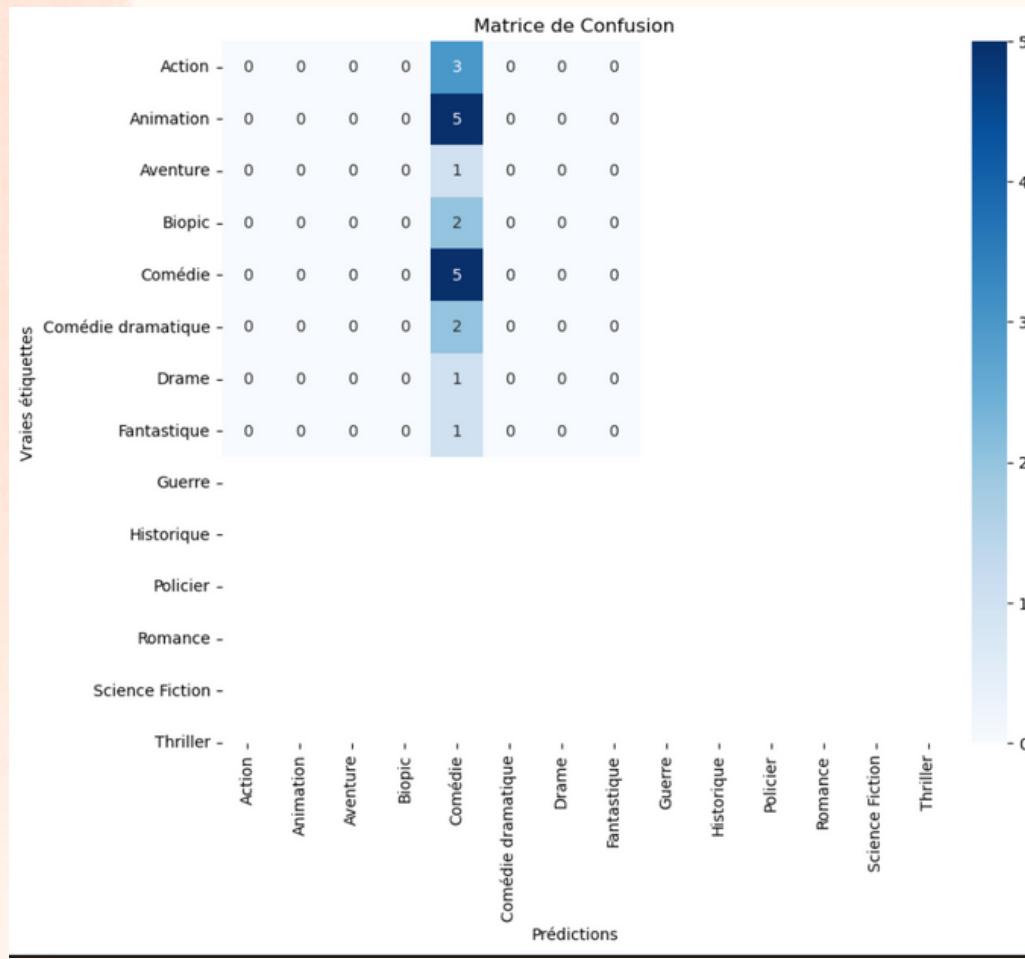
# Tokenisation du corpus
tokenized_corpus = [simple_preprocess(text) for text in df_nouveau_resume_nlp_copy['Résumé propre']]

new_fasttext_model = FastText(
    sentences=tokenized_corpus,
    vector_size=300, # La dimension des vecteurs de mots
    window=5, # La taille de la fenêtre contextuelle
    min_count=1, # Le nombre minimum d'occurrences d'un mot pour être inclus
    workers=4, # Le nombre de threads utilisés pendant l'entraînement
    sg=1 # L'utilisation de Skip-gram (sg=1) ou CBOW (sg=0)
)
```

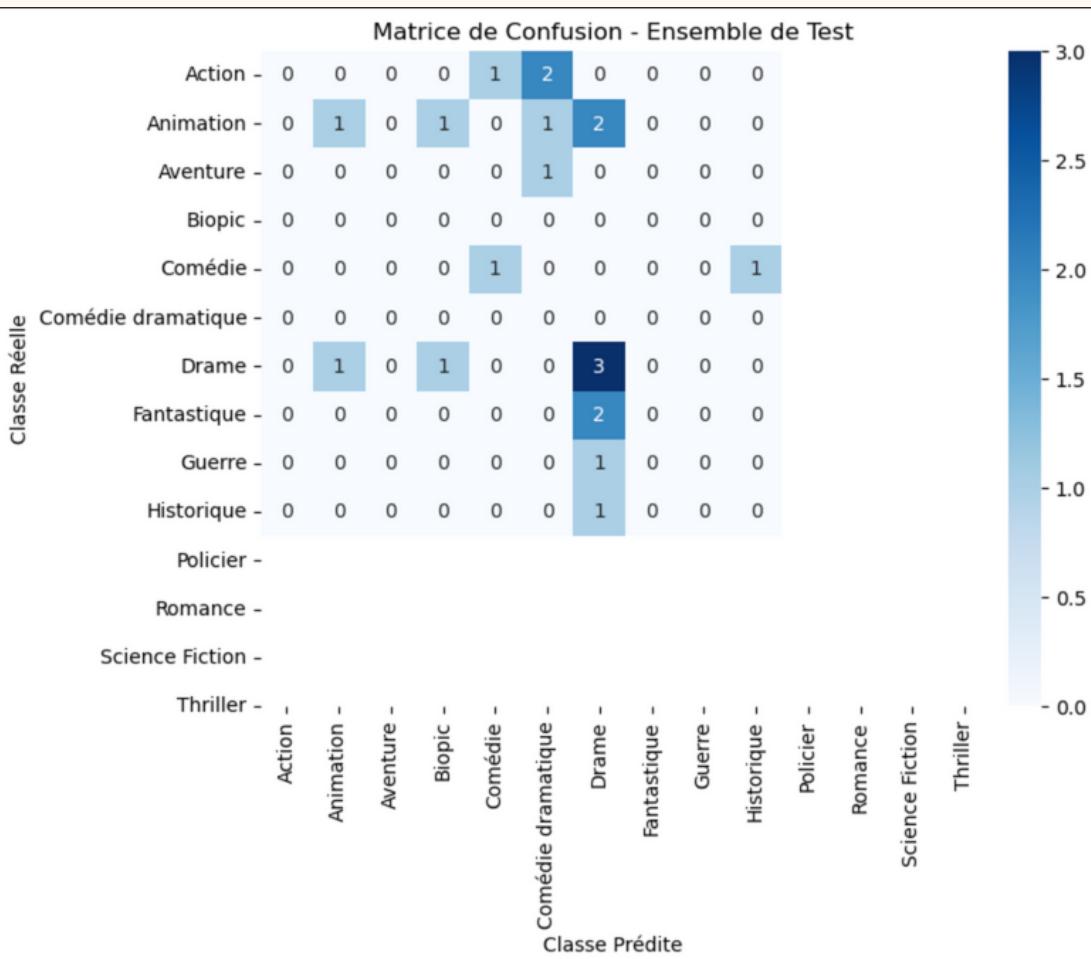
# RÉSULTAT DU MODELE DE PRÉDICTION

## RANDOM FOREST

- Matrice de confusion avec la vectorisation du TF IDF



- Matrice de confusion avec la vectorisation FastText



- Accuracy : 25 % pour les deux modèles
- Modèle avec la vectorisation TF-IDF : prédit toujours la même catégorie "Comédie", (avec 40 échantillons en plus : la prédition était toujours la catégorie majoritaire "Drame")
- Modèle avec la vectorisation FastText : capte davantage la sémantique des mots, d'où des prédictions plus diversifiés
- Explication : Faible échantillon de 139 observations et déséquilibre élevé entre les classes.
- Perspectives d'amélioration : utilisation de modèles de vectorisation pré-entraînés, fine-tuning, Transformers (Bert).

**MERCI !**