

# Recommandations de films

— Projet OpenClassrooms —

Par Xavier Montamat

---

# Problématique

Le but de ce projet est de créer un algorithme générant des recommandations de films pertinentes.

Ce à partir d'informations détaillées sur chaque films.

Il n'y a encore aucune recommandations utilisateurs.

---

# Plan de réalisation

## Nettoyage

- Données initiales
- Valeurs manquantes
- Valeurs aberrantes

## Pistes de modélisation

- Préparer les données
- Reduction de dimensions
- Modèle Kmeans
- Agglomerative clustering
- Matrice de distances

## Modèle final

- Présentation du modèle final
- Exemples
- API

# Pré analyse

## Axes d'approche possibles

- Algorithme non supervisé pour trouver des groupes
- Clustering
- Hierarchique
- Distance (approche choisie)

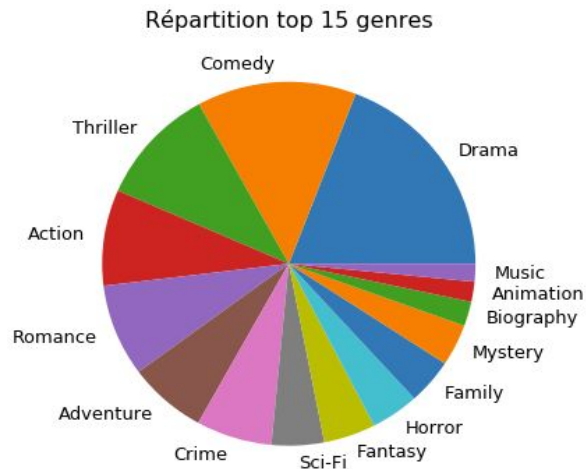
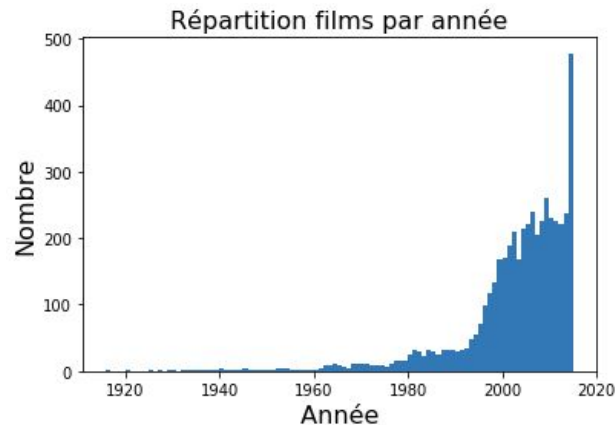
## Problèmes potentiels

- Trop de données manquantes
- Groupes de petite tailles
- Vérification des résultats

# Les données

## Quelques chiffres

- 5043 films et 28 attributs
- 65 pays et 57 langues
- 2398 directeurs différents
- 26 genres de films
- Jusqu'à 7 genres par films
- 8087 descriptions d'intrigue différentes



# Nettoyage des données

## Les valeurs manquantes

- Remplir par la moyenne / plus commun

aspect\_ratio title\_year budget duration gross

- Remplir par 0 / vide

director\_name ,director\_facebook\_likes ,num\_critic\_for\_reviews  
,director\_facebook\_likes ,actor\_3\_facebook\_likes ,actor\_2\_name  
,actor\_1\_facebook\_likes ,actor\_1\_name ,actor\_3\_name ,plot\_keywords  
,num\_user\_for\_reviews ,language ,country ,actor\_2\_facebook\_likes

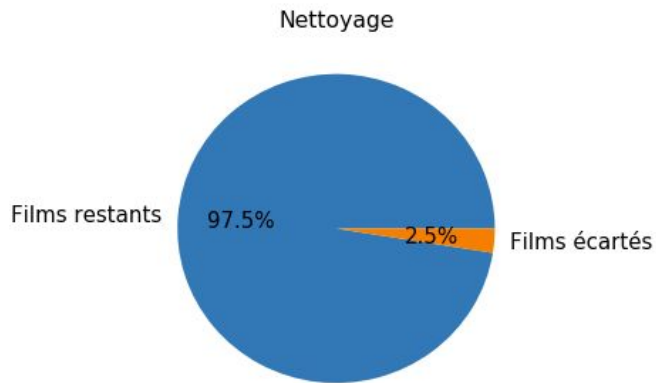
# Nettoyage des données

## Valeurs aberrantes

- Quelques valeurs trop hautes
  - Budget >300M
  - Aspect\_ratio de 16
- Quelques valeurs non corrélées
  - Director\_name + director\_facebook\_likes

## Duplicats

- 45 duplicats parfaits
- + 82 duplicats par nom
- 4916 films restants



# Pistes de modélisation

J'ai réalisé plusieurs modèles au cours de l'étude, dont Kmeans, et Hierarchy clustering.

Une des première étape est de préparer les données. Par exemple extraire les informations de texte les exprimer en valeurs.

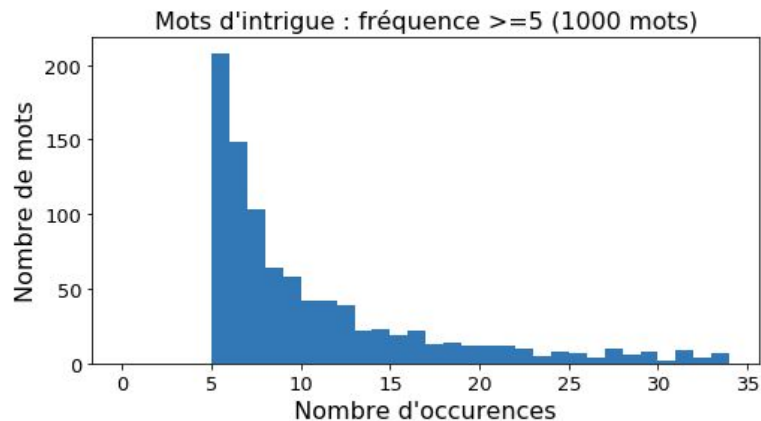
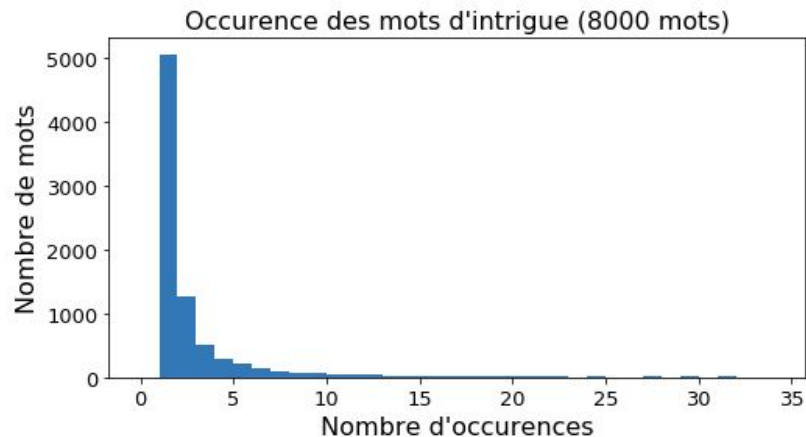
---



# Préparer les données

## Calcul de fréquence des mots (TF-IDF)

1. Unifier les mots type Prénom+Nom / New York City
2. Séparer listes en mots distincts
3. Calcul de fréquence avec le modèle TF-IDF par colonne
  - a. Retirer mots d'intrigue de trop basse fréquence



# Préparer les données

## Scaler les nombres

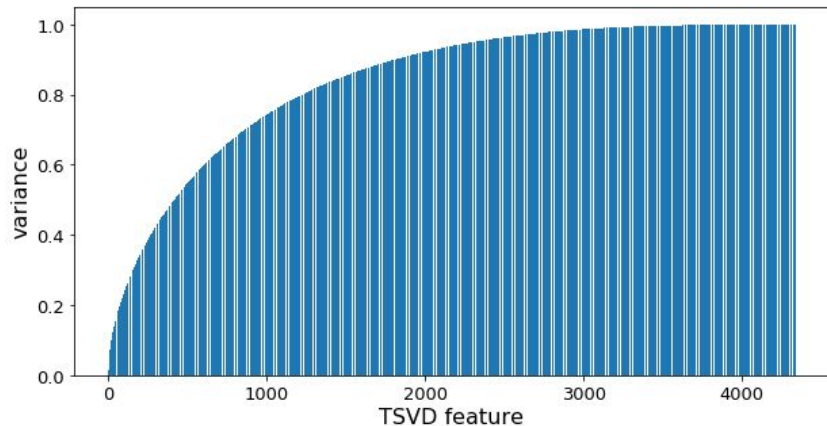
Utilisation d'un MinMax Scaler

- Grandes et faibles valeurs au même niveau
- Pouvoir comparer avec les fréquences de mots

# Réduction dimensionnelle

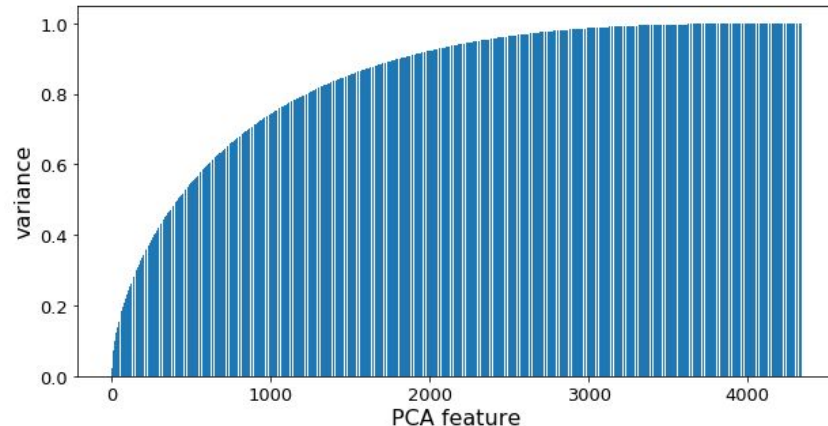
## TruncatedSVD

- Test de Truncated SVD pour les colonnes fréquences de mots.
- Compatible avec les matrices sparses (CSR)
- Variance de 0.85 atteint avec 1474



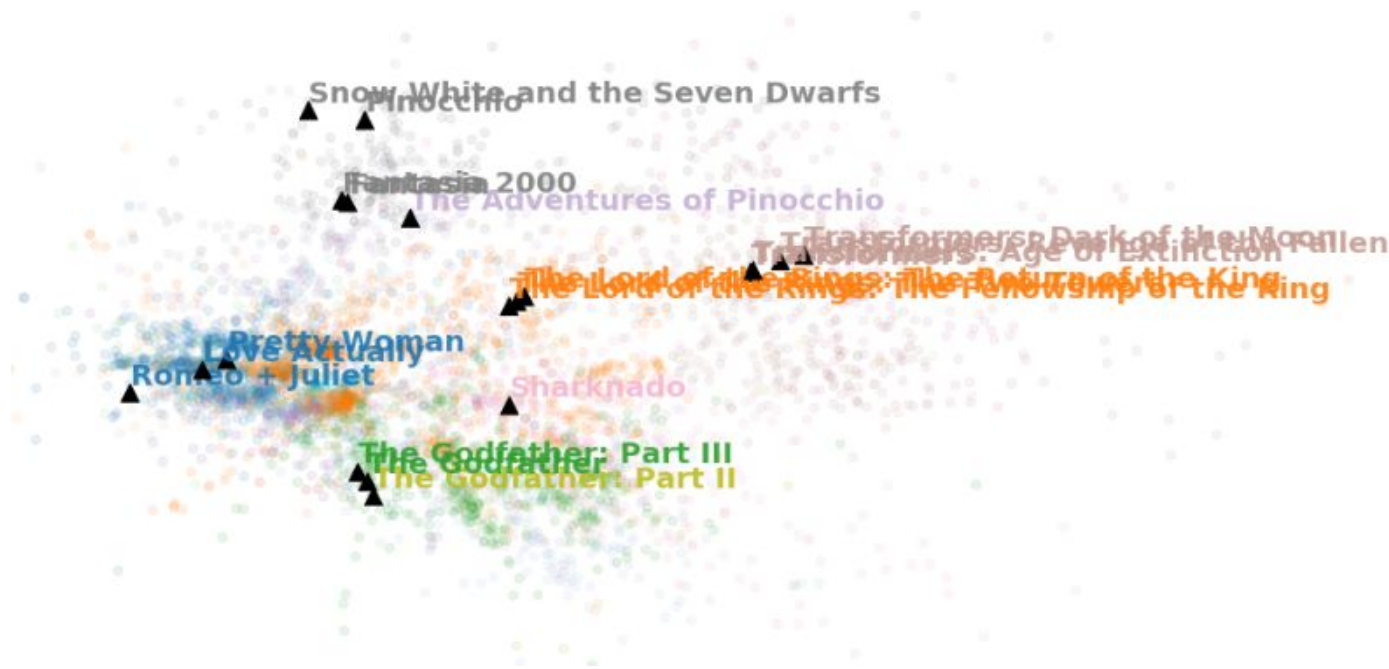
## PCA

- Plus coûteux car incompatible sparse
- Même variance



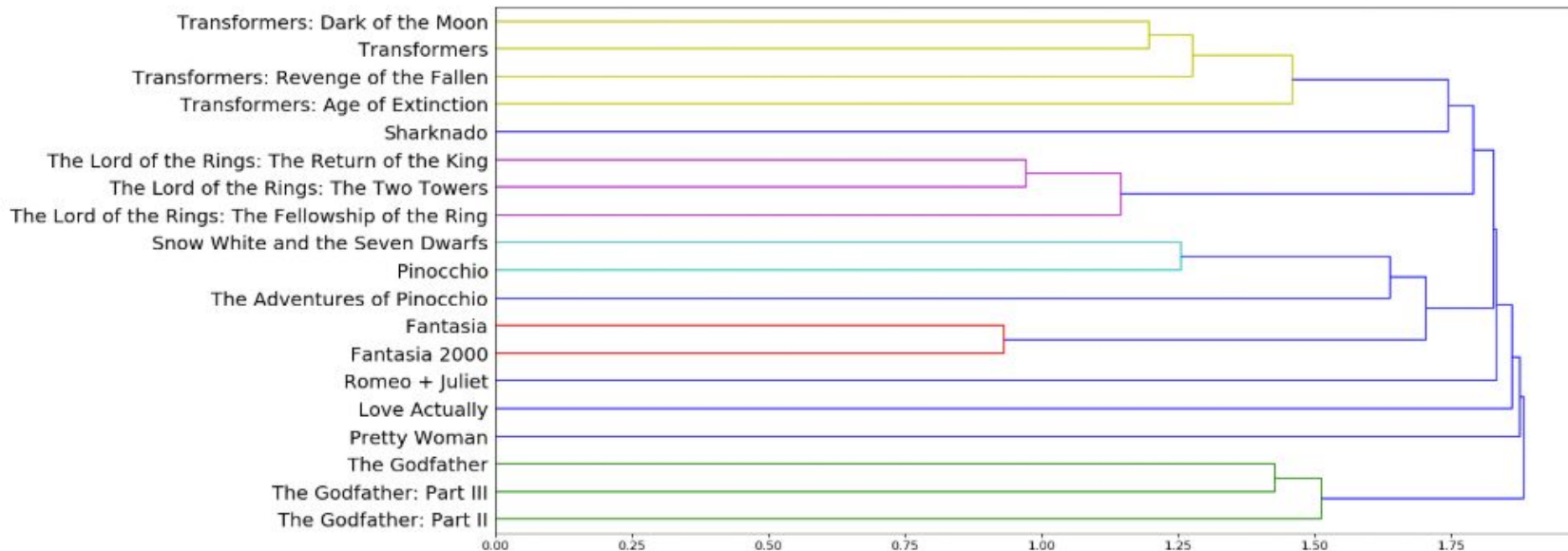
# Kmeans groupes

- Représentation 2D de Kmeans à 20 composants
- Sur 20 films test pré sélectionnés
- Trop lourd et clusters non égaux



# Cluster Hierarchique & Dendrograme

- Association de films par hiérarchie
- Sur 20 films tests pré sélectionnés
- Bon visuel mais n plus proches difficiles à récupérer



# Distance spatiale - Plus proches voisins

## Calculs de distance

- Sklearn NearestNeighbors (55 secs /1500 features)
- Scipy Spatial distance (18 secs /1500 features)
- Retourne une matrice des distances calculées
- Ordonner par plus faible distance

	movie_title	title_year	imdb_score	genres	director_name	actor_1_name	actor_2_name	actor_3_name
Django Unchained	Django Unchained	2012.000	8.500	Drama Western	Quentin Tarantino	Leonardo DiCaprio	Christoph Waltz	Ato Essandoh
Compliance	Compliance	2012.000	6.400	Biography Crime Drama Thriller	Craig Zobel	Dreama Walker	Matt Servitto	James McCaffrey
Skyfall	Skyfall	2012.000	7.800	Action Adventure Thriller	Sam Mendes	Albert Finney	Helen McCrory	Rory Kinnear
The Hateful Eight	The Hateful Eight	2015.000	7.900	Crime Drama Mystery Thriller Western	Quentin Tarantino	Craig Stark	Jennifer Jason Leigh	Zoë Bell
The Social Network	The Social Network	2010.000	7.700	Biography Drama	David Fincher	Andrew Garfield	Dustin Fitzsimons	Marcella Lentz-Pope

# Modèle Final

Le modèle final se base sur les techniques de preprocessing évoquées.

Ainsi que sur des matrices de distance pour sélectionner les recommandations.

J'ai également choisi de faire plusieurs types de recommandations pour chaque film.

---

# Choix de la distance spatiale euclidienne

- Rapide
- Simple
- Précise

Par rapport aux algos clusters

- Pas besoin de groupes
- Ni de visualisation

Récupération simple des recommandations suivantes



# Plusieurs modèles

5 modèles de recommandations :

- **General**(Directeur, acteur principal, année, genres, imdb score, pays...)
- **Genre & Intrigue** (Genres, mots d'intrigue)
- **Direction artistique** (Directeur, couleur, langue, aspect\_ratio...)
- **Acteurs** ( Acteur 1, Acteur 2, Acteur 3...)
- **Succès** (Score Imdb, reviews\_count, facebook likes...)

Plus de variété dans les suggestions

# Examples

## Django Unchained - Recommandations

Similar overall

Similar genre

Similar picture

Similar cast

Similar success

movie_title
The Hateful Eight
Kill Bill: Vol. 1
The Revenant
Reservoir Dogs
The Great Gatsby

movie_title
Unforgiven
The Work and the Glory II: American Zion
Doc Holliday's Revenge
The Legend of God's Gun
The Ballad of Gregorio Cortez

movie_title
Inglourious Basterds
The Hateful Eight
Pulp Fiction
Kill Bill: Vol. 2
Jackie Brown

movie_title
Shutter Island
The Great Gatsby
Body of Lies
The Beach
The Quick and the Dead

movie_title
Mad Max: Fury Road
Black Swan
The Wolf of Wall Street
Argo
Skyfall

# Pertinence de recommandation

77.6% de chance que les films du même directeur soient recommandés dans le top 5 catégories **overall / picture**

81.3% de chance que les films du même acteur principal soient recommandés dans le top 5 catégories **overall / cast**

Variance moyenne du score Imdb :

- Dataset complet : 1.27
- Similaire en général : 0.44
- Succès similaire : 0.01

# API

Export de 5 recommandation par catégorie

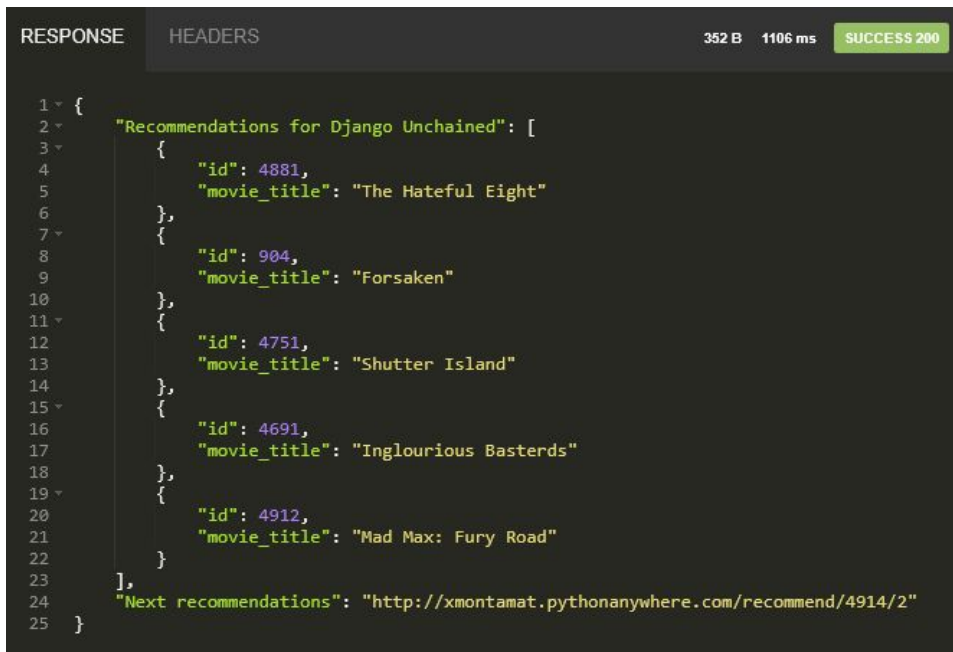
Affichage du premier film par catégorie

Exemple :

<http://xmontamat.pythonanywhere.com/recommend/4914>

Pas de doublons

Pagination possible jusqu'à 5 pages



```
RESPONSE HEADERS 352 B 1106 ms SUCCESS 200

1 {
2   "Recommendations for Django Unchained": [
3     {
4       "id": 4881,
5       "movie_title": "The Hateful Eight"
6     },
7     {
8       "id": 904,
9       "movie_title": "Forsaken"
10    },
11    {
12      "id": 4751,
13      "movie_title": "Shutter Island"
14    },
15    {
16      "id": 4691,
17      "movie_title": "Inglourious Basterds"
18    },
19    {
20      "id": 4912,
21      "movie_title": "Mad Max: Fury Road"
22    }
23  ],
24  "Next recommendations": "http://xmontamat.pythonanywhere.com/recommend/4914/2"
25 }
```

# Conclusion

**Resultats pertinents**  
**Choix variés et réalistes**  
**Facilement évolutif**

# Axes d'amélioration

Pagination des recommandations  
Filtrer films déjà vu  
Nouveau modèle notes utilisateurs

# Questions



Xavier Montamat