DS P3 - Développez un moteur de recommandations de films

Source: https://morioh.com/p/5e461f15f1df

Sommaire

- Contexte
- Nettoyage
- Exploration
- Moteur
- Conclusion

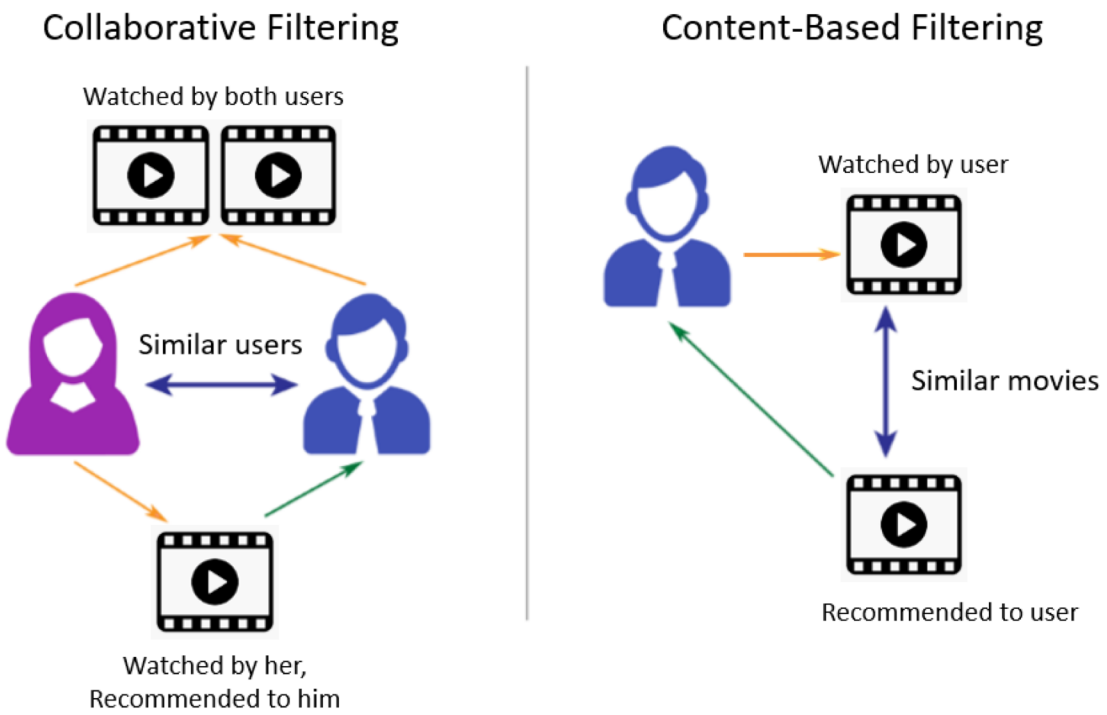
Contexte - Objectif

- Moteur de recommandations de film
  - Api accessible dans le cloud
  - Pour un titre de film donné, retourner 5 titres similaires et/ou intéressants

- Les données
  - Dataset: movie_metadata.csv (5043 titres de films, 28 variables, 1,5Mo)

Contexte - Problématique

- Les moteurs de recommandations les plus populaires se répartissent en 2 grandes familles



  - Pas de retours utilisateurs => recommandations orientées contenu
- Trouver des films similaires
  - Quels critères/variables utiliser?
  - Quel algorithme utiliser?

Nettoyage – Corrélations

	action	adventure	animation	biography	comedy	crime	documentary	drama	family	fantasy	film_noir	game_show	history
action	1.00	0.31	-0.02	-0.09	-0.17	0.15	-0.08	-0.23	-0.06	0.06	-0.02	-0.01	-0.01
adventure	0.31	1.00	0.30	-0.07	-0.03	-0.15	-0.06	-0.24	0.31	0.27	-0.02	-0.01	0.01
animation	-0.02	0.30	1.00	-0.04	0.16	-0.09	-0.03	-0.17	0.54	0.25	-0.01	-0.00	-0.04
biography	-0.09	-0.07	-0.04	1.00	-0.14	-0.01	0.04	0.21	-0.07	-0.08	-0.01	-0.00	0.30
comedy	-0.17	-0.03	0.16	-0.14	1.00	-0.08	-0.09	-0.25	0.21	0.04	-0.03	-0.01	-0.14
crime	0.15	-0.15	-0.09	-0.01	-0.08	1.00	-0.05	0.07	-0.13	-0.15	0.03	-0.01	-0.06
documentary	-0.08	-0.06	-0.03	0.04	-0.09	-0.05	1.00	-0.13	-0.05	-0.05	-0.01	-0.00	0.03
drama	-0.23	-0.24	-0.17	0.21	-0.25	0.07	-0.13	1.00	-0.18	-0.20	0.01	-0.01	0.17
family	-0.06	0.31	0.54	-0.07	0.21	-0.13	-0.05	-0.18	1.00	0.31	-0.01	-0.00	-0.06
fantasy	0.06	0.27	0.25	-0.08	0.04	-0.15	-0.05	-0.20	0.31	1.00	-0.01	-0.01	-0.07
film_noir	-0.02	-0.02	-0.01	-0.01	-0.03	0.03	-0.01	0.01	-0.01	-0.01	1.00	-0.00	-0.01
game_show	-0.01	-0.01	-0.00	-0.00	-0.01	-0.01	-0.00	-0.01	-0.00	-0.01	-0.00	1.00	-0.00
history	-0.01	0.01	-0.04	0.30	-0.14	-0.06	0.03	0.17	-0.06	-0.07	-0.01	-0.00	1.00
horror	-0.06	-0.11	-0.07	-0.09	-0.15	-0.11	-0.06	-0.23	-0.11	0.09	-0.01	-0.01	-0.07
music	-0.10	-0.07	-0.01	0.09	0.04	-0.06	0.08	0.05	0.02	-0.03	-0.01	-0.00	-0.01
musical	-0.08	0.02	0.13	0.02	0.05	-0.05	-0.03	-0.00	0.17	0.06	-0.01	-0.00	-0.00
mystery	-0.05	-0.06	-0.05	-0.08	-0.19	0.12	-0.05	0.00	-0.07	-0.01	0.05	-0.00	-0.07
news	-0.01	-0.01	-0.01	-0.01	-0.02	0.01	0.16	-0.01	-0.01	-0.01	-0.00	-0.00	0.04
reality_tv	-0.01	-0.01	-0.00	-0.01	0.01	-0.01	-0.00	-0.00	-0.01	-0.01	-0.00	0.71	-0.00
romance	-0.17	-0.12	-0.07	-0.02	0.18	-0.12	-0.08	0.16	-0.05	-0.05	-0.00	0.03	-0.01
sci_fi	0.28	0.23	0.06	-0.09	-0.09	-0.13	-0.05	-0.20	0.02	0.03	-0.01	-0.01	-0.08
short	-0.02	0.00	-0.01	-0.01	0.00	-0.01	0.04	-0.03	0.01	-0.01	-0.00	-0.00	-0.01
sport	-0.04	-0.05	-0.01	0.15	0.00	-0.08	0.04	0.07	0.03	-0.06	-0.01	-0.00	0.01
thriller	0.28	-0.03	-0.12	-0.09	-0.36	0.35	-0.10	-0.03	-0.20	-0.08	0.03	-0.01	-0.06
war	0.04	0.01	-0.03	0.07	-0.12	-0.08	0.02	0.16	-0.07	-0.05	-0.01	-0.00	0.33

OLS Regression Results

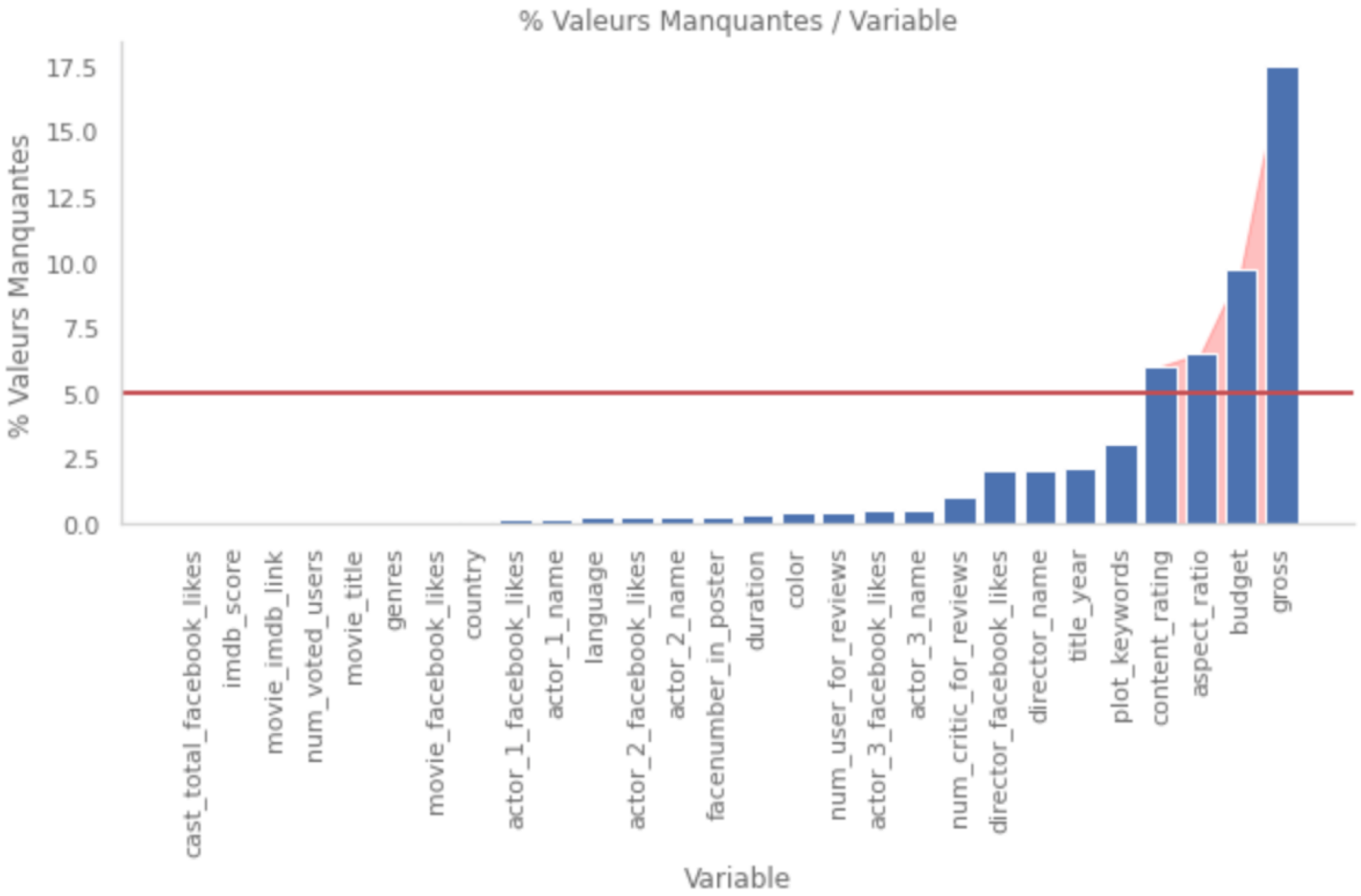
Dep. Variable:	actor_1_facebook_likes	R-squared:	0.897
Model:	OLS	Adj. R-squared:	0.897
Method:	Least Squares	F-statistic:	3.275e+04
Date:	Wed, 07 Apr 2021	Prob (F-statistic):	0.00
Time:	21:17:27	Log-Likelihood:	-37457.
No. Observations:	3772	AIC:	7.492e+04
Df Residuals:	3770	BIC:	7.493e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1105.0395	93.945	-11.763	0.000	-1289.227	-920.852
cast_total_facebook_likes	0.7735	0.004	180.979	0.000	0.765	0.782

- R^2 proche de 0,9 , score significatif
- cast_total_facebook_likes explique 77% actor_1_facebook_likes avec une p-value à 0 (qui confirme l'hypothèse alternative que le coefficient est différent de 0)

➔ on exclue la variable cast_total_facebook_likes de notre étude

Nettoyage – Suppression



- Valeurs manquantes
  - Majorité de films avec - de 5%
  - Maximum 18% (relativement peu)
  - Variables gross (revenu), budget à garder pour leur influence potentielle dans le choix d'un film
- Variables inutiles
  - Imdb_link
  - Aspect_ratio
  - Color
  - Country
  - Content_rating
  - Cast_total_facebook_likes
- Dédoublonnage
  - Par titre de film/réalisateur

	movie_title	director_name	count
489	Ben-Hur	Timur Bekmambetov	3
1558	Halloween	John Carpenter	3
1679	Home	Tim Johnson	3
4602	Victor Frankenstein	Paul McGuigan	3
2618	Pan	Joe Wright	3
1988	King Kong	Peter Jackson	3
...	...	...	...

Nettoyage - Transformation

- Binarisation (One Hot Encoding)
  - 26 genres

	genres
0	Action|Adventure|Fantasy|Sci_Fi
1	Action|Adventure|Fantasy
2	Action|Adventure|Thriller
3	Action|Thriller
4	Documentary

	action	adventure	animation	biography	comedy	crime	documentary	drama	family	fantasy	film_noir	game_show	his
0	1	1	0	0	0	0	0	0	0	1	0	0	
1	1	1	0	0	0	0	0	0	0	1	0	0	
2	1	1	0	0	0	0	0	0	0	0	0	0	
3	1	0	0	0	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	1	0	0	0	0	0	

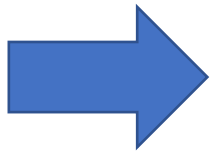
  - 8112 plots

	plot_keywords
4995	estranged_daughter friendship nevada prisoner unlikely_friendship
416	action_hero hero magic ticket video_store
2601	airport child_protagonist child_swearing christmas home_alone
1950	basement creature disappearance interior_designer mansion
4638	death dock longshoreman murder union

	1000000_b.c.	10_year_old	1190s	12_step_program	12_year_old	12_year_time_span	12th_century	13_year_old	13_y
1707	0	0	0	0	0	0	0	0	0
1037	0	0	0	0	0	0	0	0	0
3115	0	0	0	0	0	0	0	0	0
3050	0	0	0	0	0	0	0	0	0
2104	0	0	0	0	0	0	0	0	0

Nettoyage - Transformation

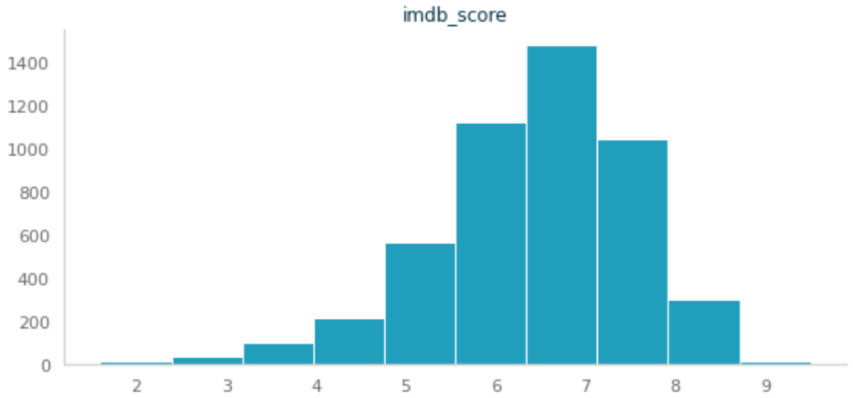
- movie_metadata.csv
  - 5043x28
  - 1,1 Mb

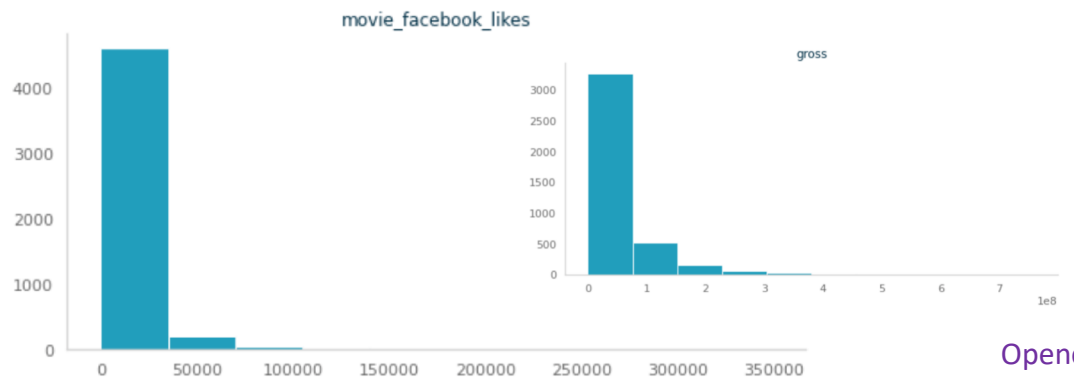


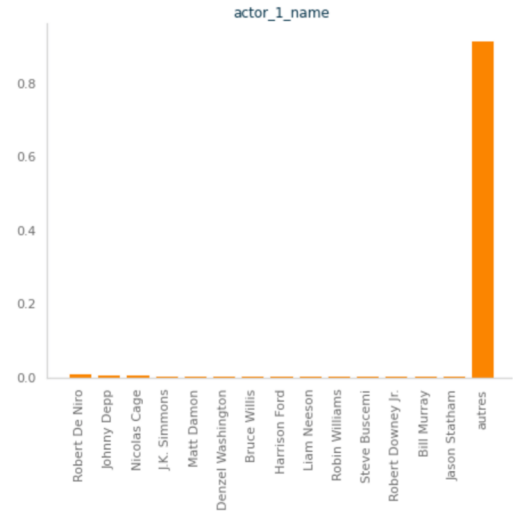
- movie_metadata_clean.csv
  - 4919x21
  - 807 Kb
- movie_genres.csv
  - 4919x23
  - 884 Kb
- movie_plots.csv
  - 4919x84
  - 3.2 Mb

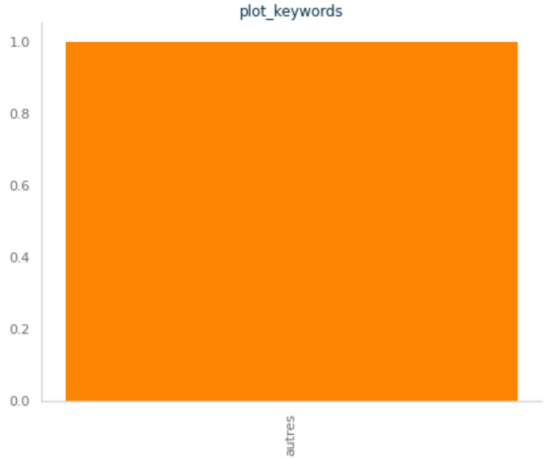
Exploration - Distributions

- Variables Quantitatives
  - Imdb_score – distribution normale
  - Distribution non normale autrement

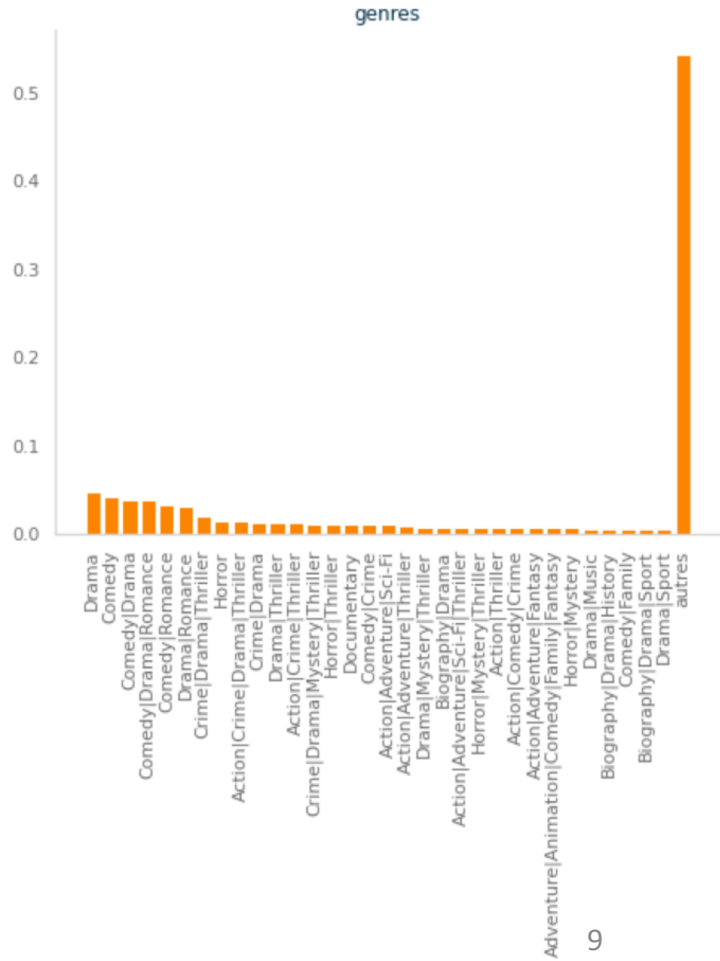








- Variables Qualitatives

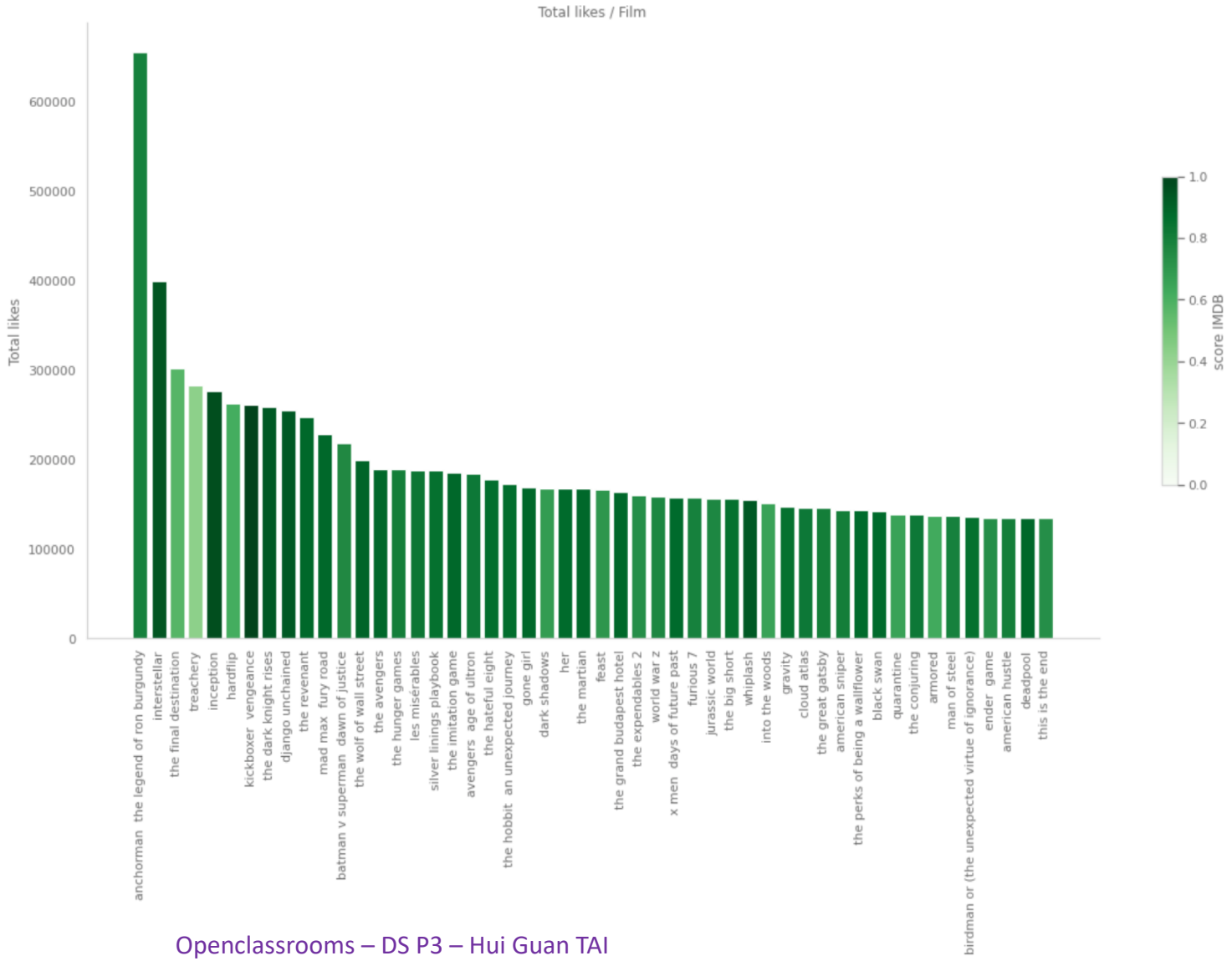


Exploration – Distributions (likes)

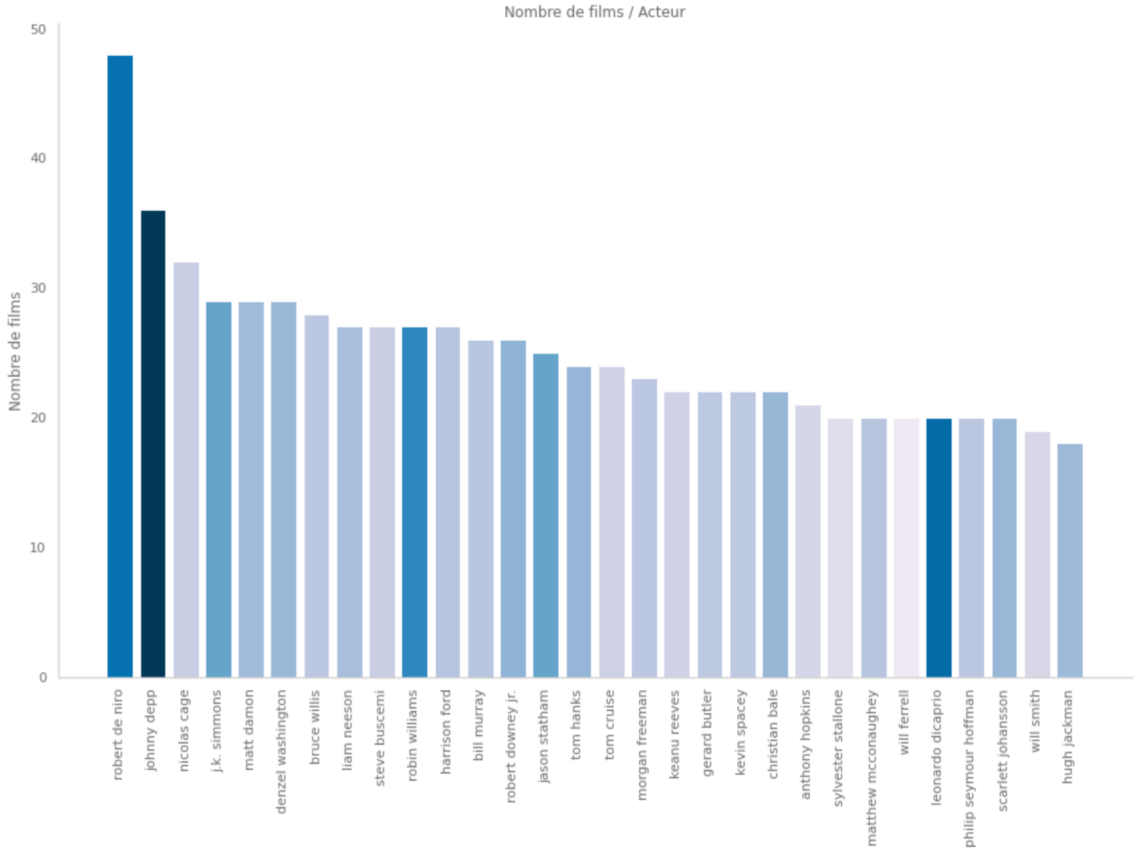
Variable créée

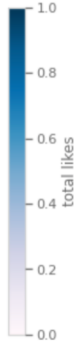
- Total facebook likes
  - director likes
  - actor1 likes
  - actor2 likes
  - actor3 likes
  - movie likes

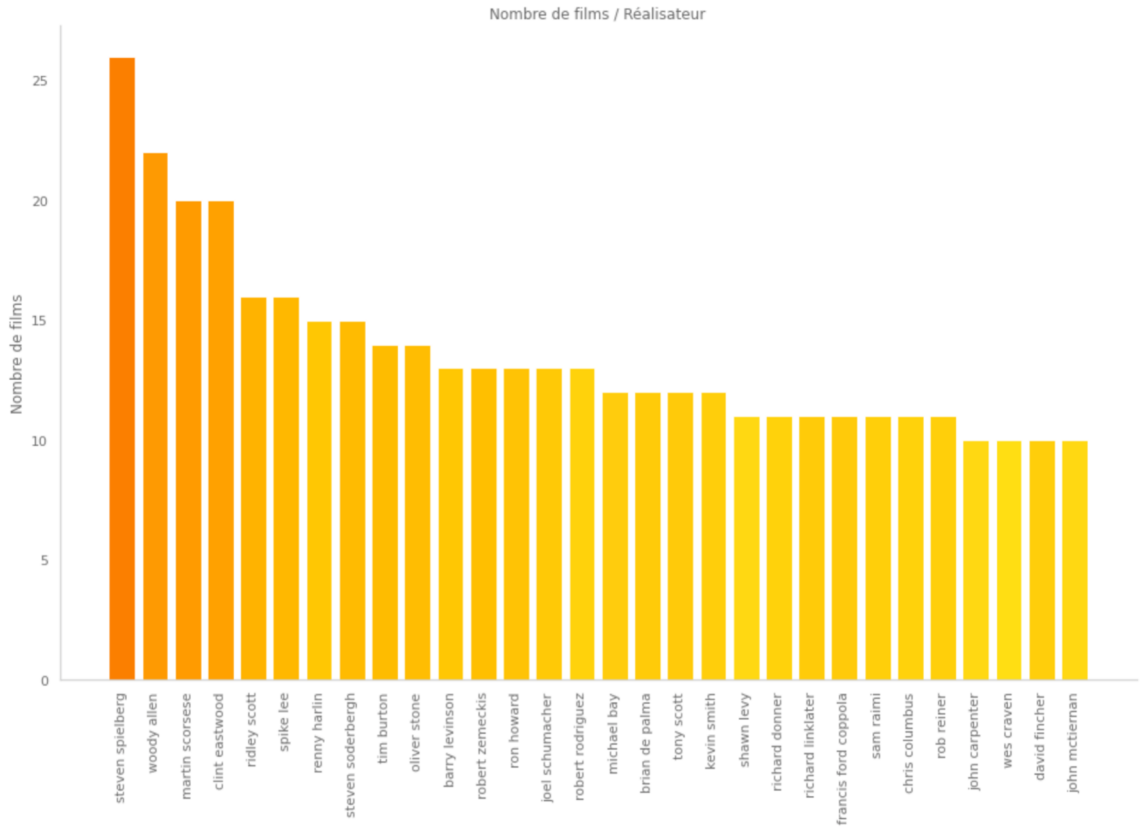
➔ Le score IMDB ne suit pas le nombre de likes

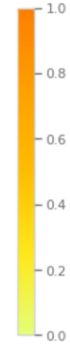


Exploration – Distributions (acteurs,directeurs)



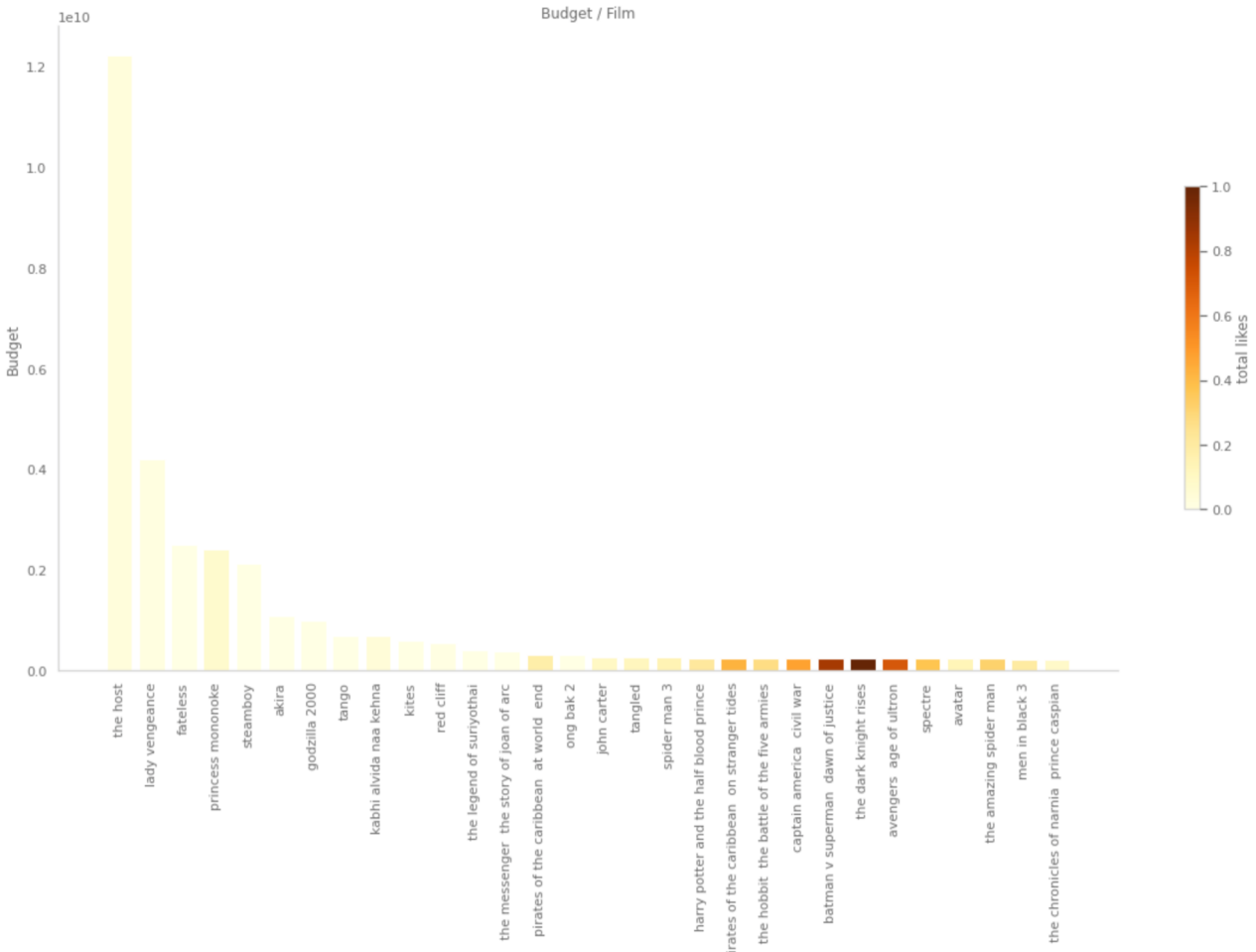


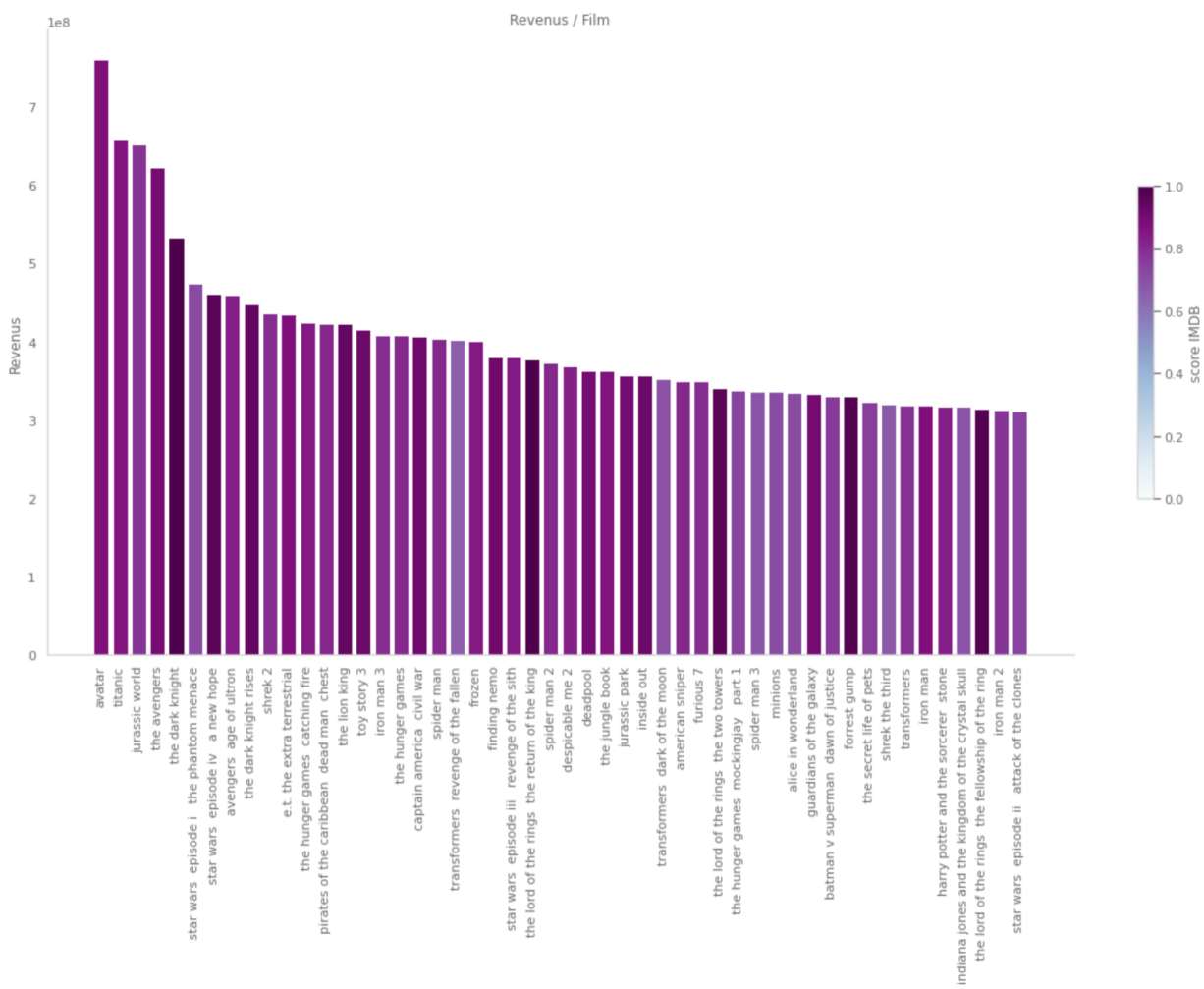




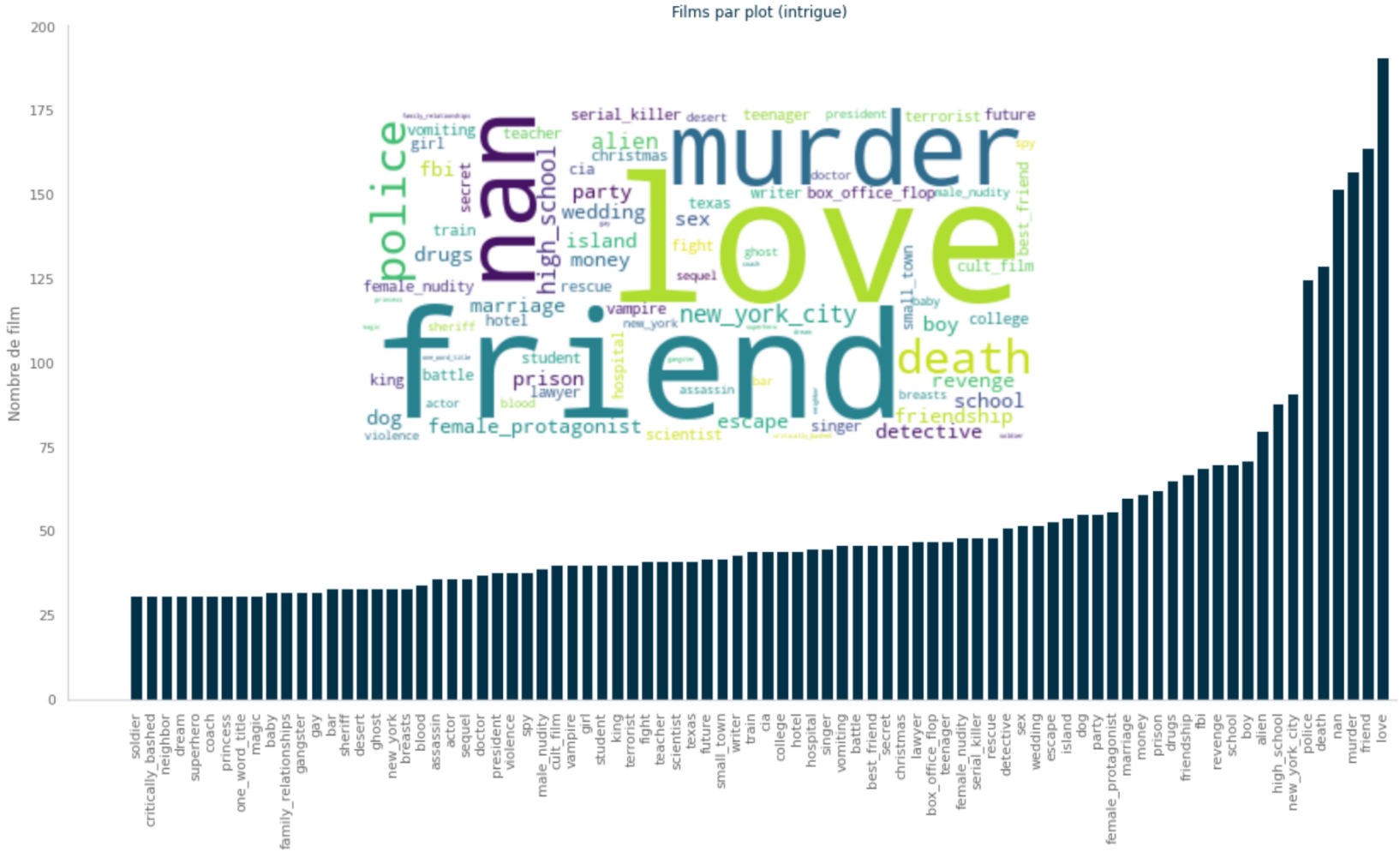
➔ Le score IMDB croit avec le nombre de films (par réalisateurs)
… contrairement au nombre de likes

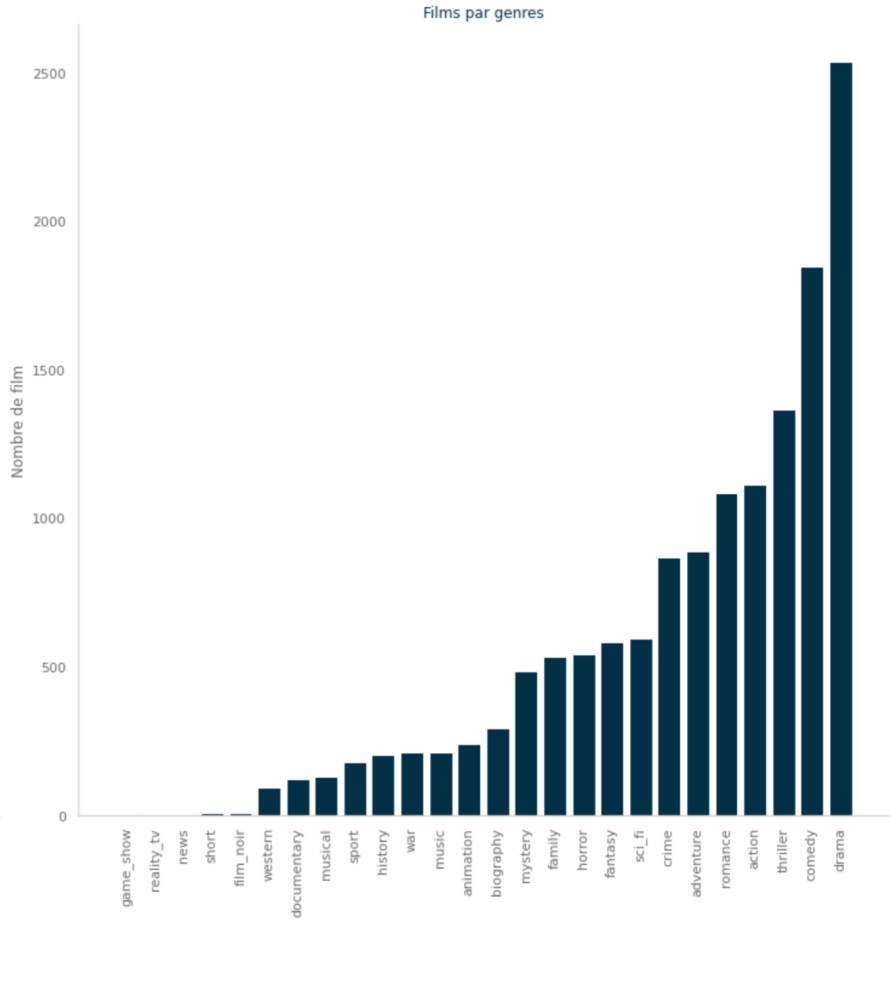
Exploration – Distributions (budget, revenus)





Exploration – Distributions (genre, plot)





➔ Les genres et plots sont les critères privilégiés pour découvrir des nouveaux films

Exploration – Règles d'association

- Genres

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
187	(comedy, animation, adventure)	(family)	0.024395	0.108559	0.022566	0.925000	8.520740	0.019917	11.885885
134	(comedy, animation)	(family)	0.034560	0.108559	0.030901	0.894118	8.236263	0.027149	8.419168
110	(animation, adventure)	(family)	0.033950	0.108559	0.030291	0.892216	8.218742	0.026605	8.270595
21	(animation)	(family)	0.048790	0.108559	0.041269	0.845833	7.791487	0.035972	5.782322

➔ Les films de comédie, d'animation ou d'aventure sont souvent des films pour la famille

- Likes

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
347	(director_facebook_likes, actor_1_facebook_likes, actor_3_facebook_likes, movie_facebook_likes)	(actor_2_facebook_likes)	0.029602	0.250156	0.027517	0.929577	3.715986	0.020112	10.64778
277	(actor_1_facebook_likes, actor_3_facebook_likes, movie_facebook_likes)	(actor_2_facebook_likes)	0.063790	0.250156	0.057953	0.908497	3.631716	0.041995	8.19472

➔ Les films très liké sur le réalisateur, l'acteur 1 et le film sont très souvent liké sur l'acteur 2 aussi

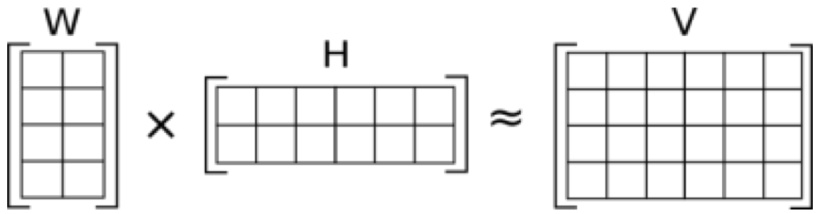
- Gross/Budget

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
4	(budget, imdb_score)	(gross)	0.06438	0.250132	0.048285	0.75	2.998418	0.032181	2.999472

➔ Les films a gros budget et avec un score imdb élevé ont des revenus élevés (dans le 3e quartile)

Exploration – Réduction dimensionnelle

- 8112 plot_keywords
- Comment synthétiser ces plots en moins d'une dizaine de facteurs?
  - NMF = trouver des vecteurs latents WxH qui factorisent le vecteur V de départ (contenant les plots/films)



movie_title	friend	police	death	alien	love	revenge	school	high_school	new_york_city	murder
avatar	0.000000	0.019288	0.000000	0.117070	0.000000	0.000000	0.000915	0.000000	0.000000	0.030754
spectre	0.000000	0.050056	0.034171	0.000000	0.000000	0.134406	0.000457	0.000000	0.042750	0.001287
star wars episode vii the force awakens	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
john carter	0.001754	0.000000	0.000000	1.185603	0.007573	0.000000	0.000000	0.000000	0.000000	0.000000
avengers age of ultron	0.000000	0.000000	0.000000	0.000502	0.000	0.020721	0.007862	0.000000	0.009127	0.000000
...	...	...	...	...	...	...	...	...	...	...
el mariachi	0.000000	0.004098	1.300508	0.0	00	0.004839	0.000000	0.027134	0.000000	0.000000
signed sealed delivered	0.000000	0.015257	0.000000	0.018193	0.000000	0.366216	0.000000	0.000054	0.000000	0.000000
the following	0.000000	0.033787	0.052642	0.023471	0.000000	0.047122	0.114815	0.000000	0.010481	0.132573
a plague so pleasant	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
shanghai calling	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

3030 rows × 10 columns

W

	soldier	critically_bashed	neighbor	dream	superhero	coach	princess	one_word_title	magic	baby	family_relationships	gangster	gay
0	0.002757	0.005070	0.002132	0.000000	0.000000	0.000181	0.002001	0.002461	0.001605	0.006960	0.001662	0.000000	0.005175
1	0.000000	0.005464	0.004358	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.020227	0.005314
2	0.004941	0.000000	0.005666	0.014261	0.000035	0.005248	0.000000	0.000000	0.006667	0.005442	0.000000	0.000000	0.000000
3	0.001997	0.014926	0.000000	0.000630	0.001241	0.000000	0.017863	0.000679	0.002177	0.032930	0.000000	0.000000	0.000000
4	0.000000	0.000000	0.018962	0.024319	0.000000	0.000000	0.012369	0.012001	0.011836	0.002529	0.012254	0.000000	0.003311
5	0.046125	0.019313	0.005122	0.001889	0.015969	0.006308	0.000000	0.001	0.00	0.000000	0.000000	0.050087	0.000300
6	0.005309	0.000000	0.012980	0.020452	0.009772	0.015203	0.007150	0.000	0.0	0.049858	0.020840	0.000000	0.031527
7	0.000000	0.000000	0.000000	0.000000	0.000374	0.088455	0.000000	0.001	0.00	0.000000	0.018060	0.000000	0.048935
8	0.000000	0.032580	0.029304	0.021413	0.023616	0.000000	0.000000	0.025925	0.000000	0.020448	0.001090	0.017216	0.000000
9	0.013852	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.008122	0.000000	0.000000	0.007814	0.001936	0.000000

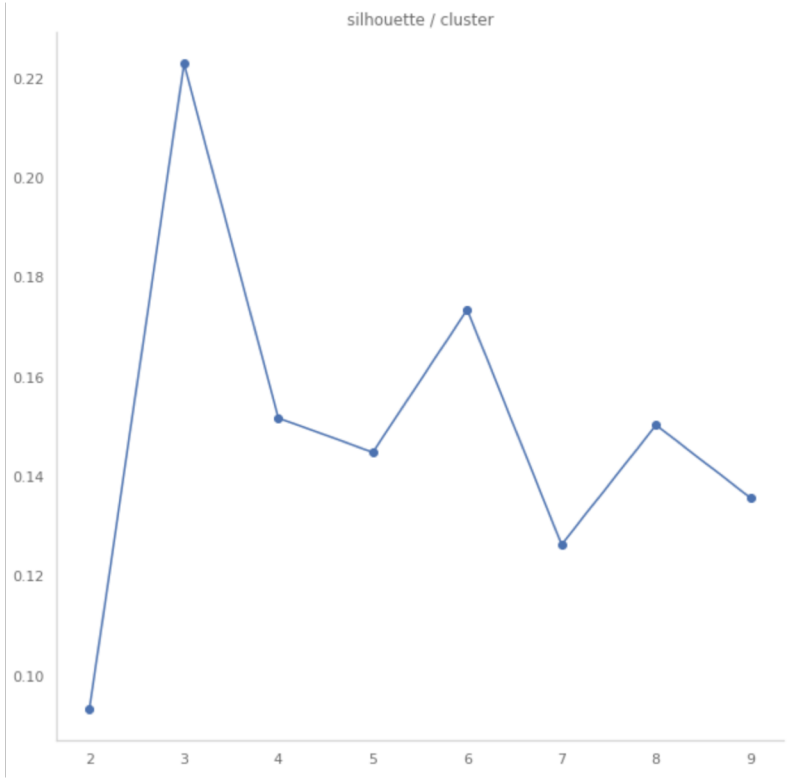
10 rows × 83 columns

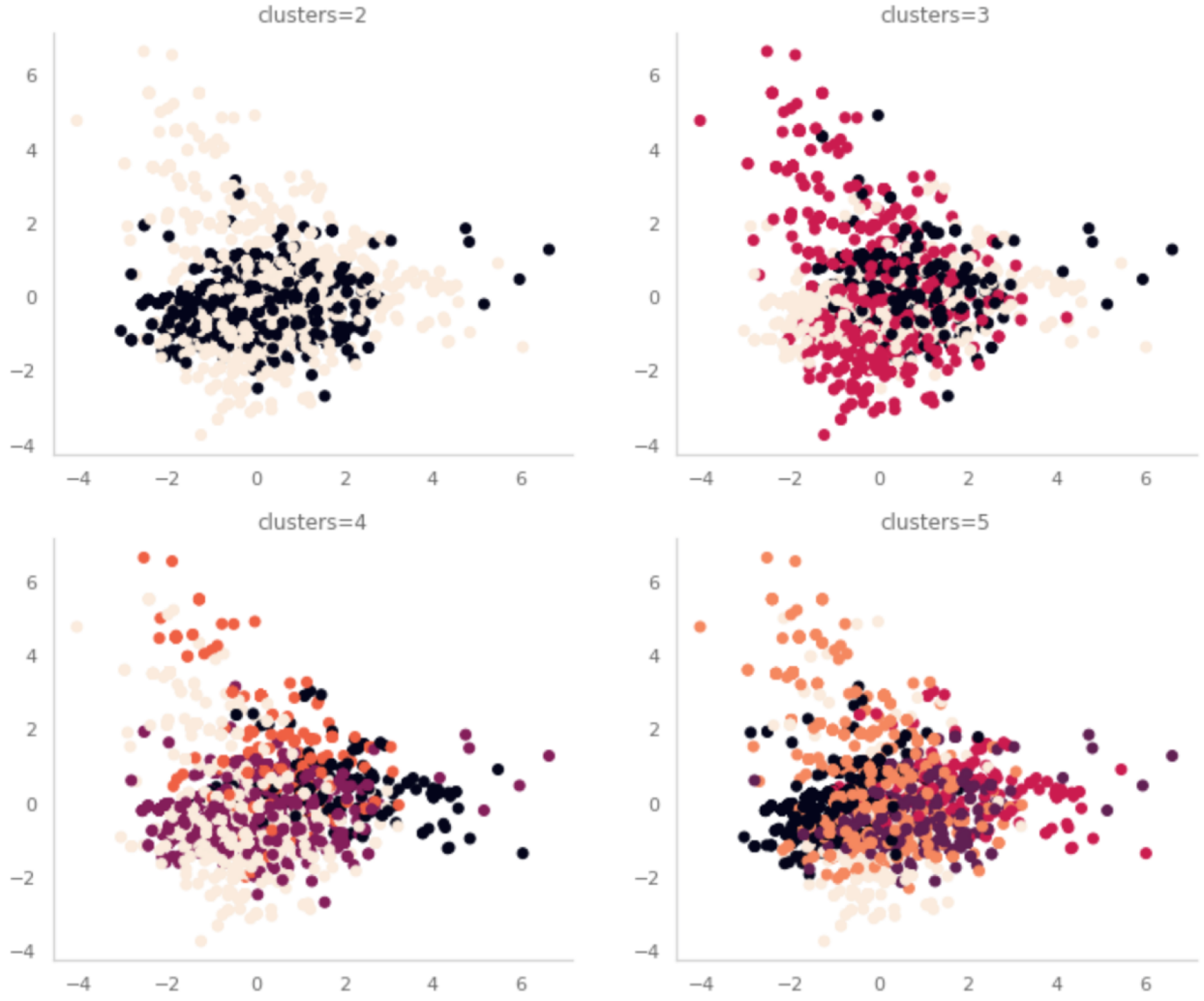
H

➔ 10 Facteurs latents qui expliquent l'ensemble des 8112 plots

Exploration – Clustering KMeans

- 23 genres de films

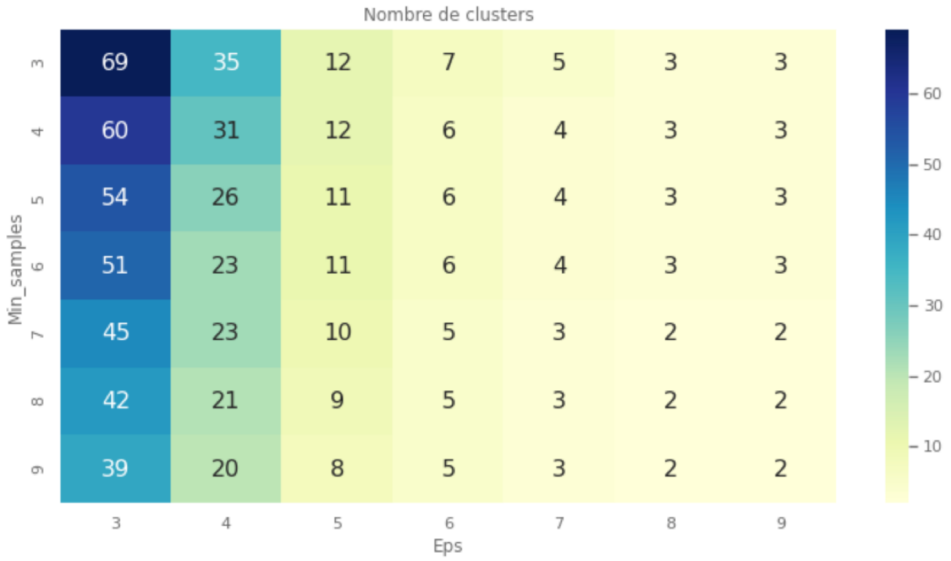




➔ Clustering Kmeans peu concluant

Exploration – Clustering DBSCAN

- Recherche de paramètres eps/nb points optimum




```
Cluster 0 -> thriller     319.802595
dtype: float64
Cluster 1 -> documentary     629.705428
dtype: float64
Cluster 2 -> musical     743.612248
dtype: float64
Cluster 3 -> western     587.487438
dtype: float64
Cluster 4 -> history     831.16204
dtype: float64
Cluster 5 -> sport     822.983761
dtype: float64
Cluster 6 -> sport     41.407988
dtype: float64
Cluster 7 -> documentary     56.673488
dtype: float64
Cluster 8 -> documentary     50.376434
dtype: float64
Cluster 9 -> film_noir     171.691584
dtype: float64
Cluster 10 -> action     0.0
dtype: float64
```

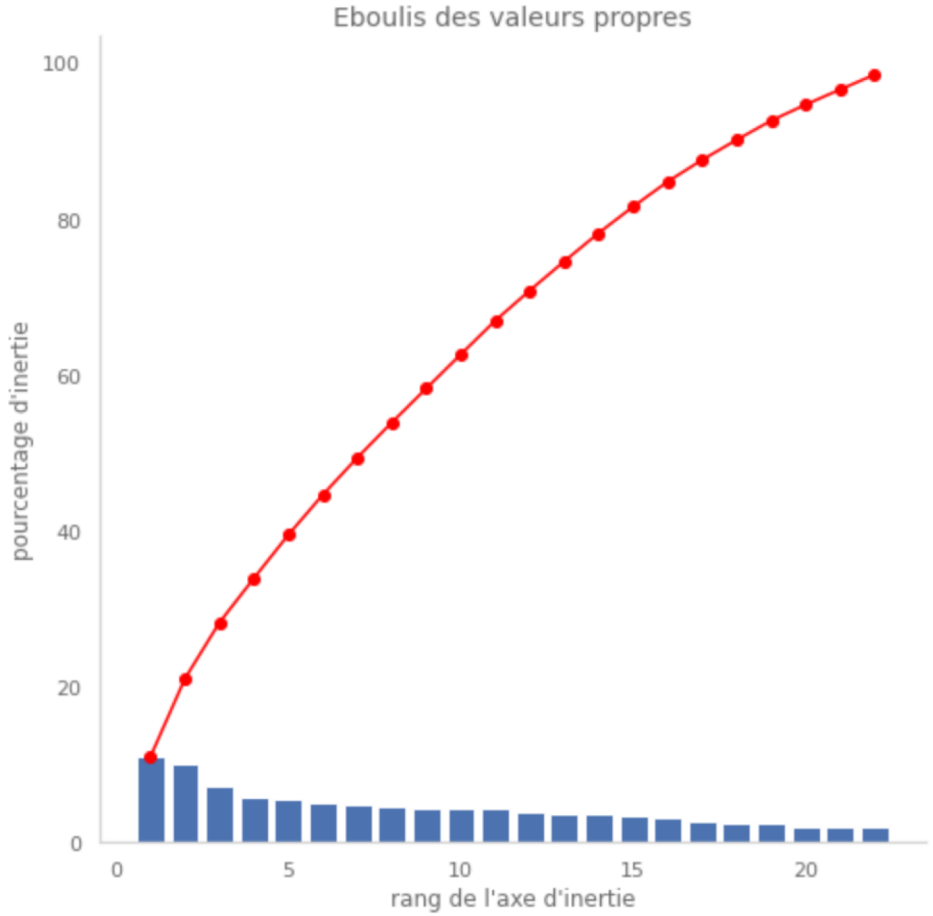
```
 0    4226
 4     172
 5     159
 2     123
 1     100
 3      82
-1      26
 7       9
 8       8
 6       8
 9       6
Name: Cluster, dtype: int64
```

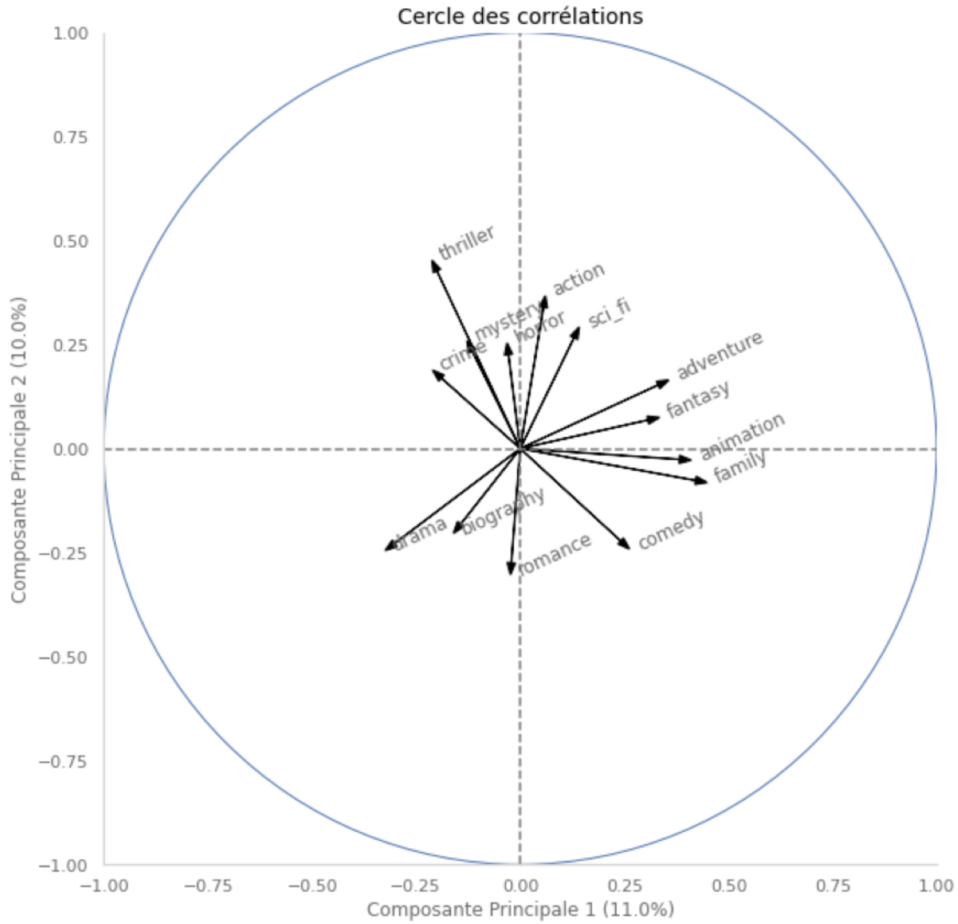
On a 10 clusters avec
une majorité dans le cluster 0

➔ 10 clusters de genre, dont 2 clusters redondants (documentary & sport)

Exploration – ACP

- 23 genres

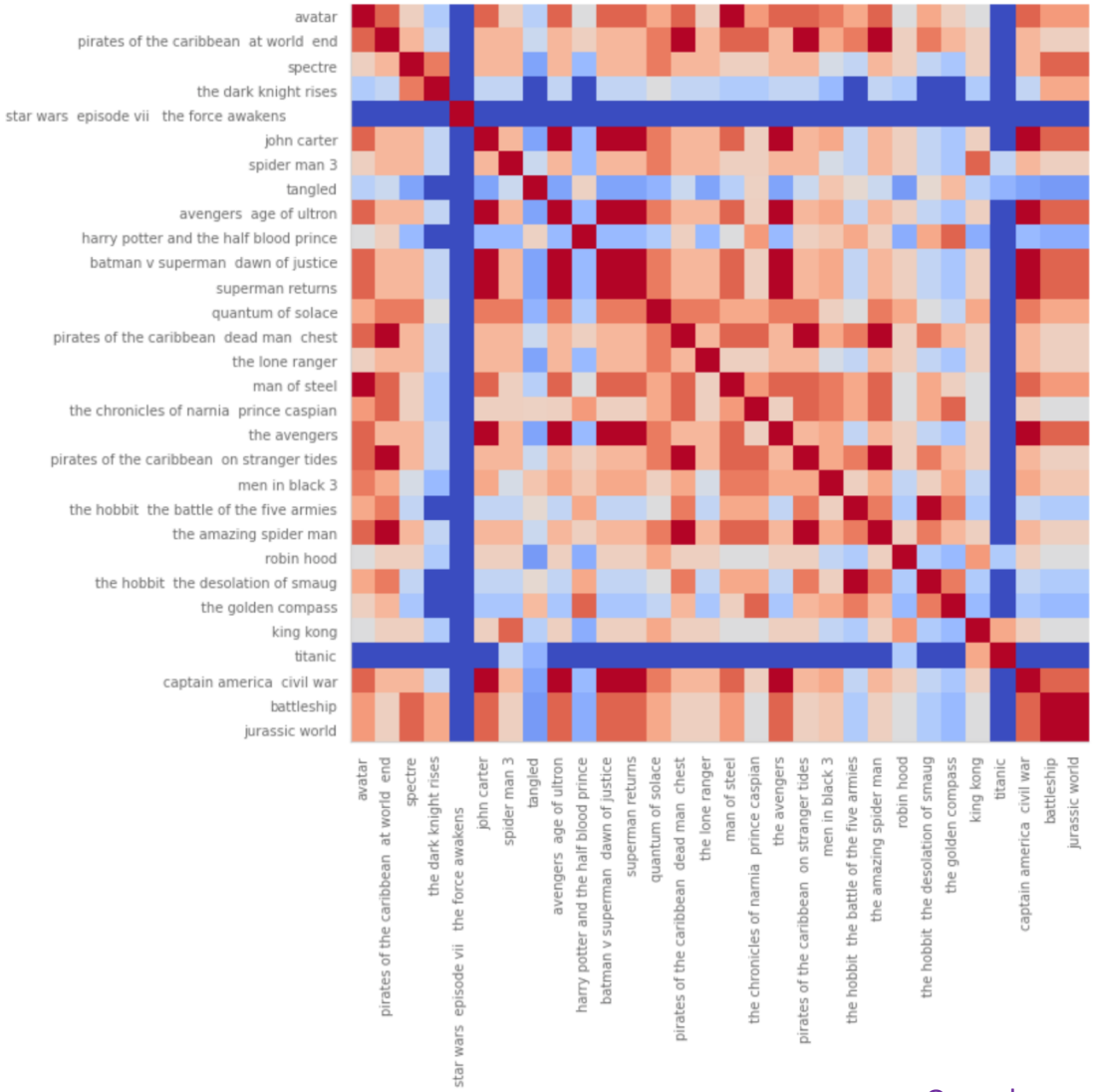




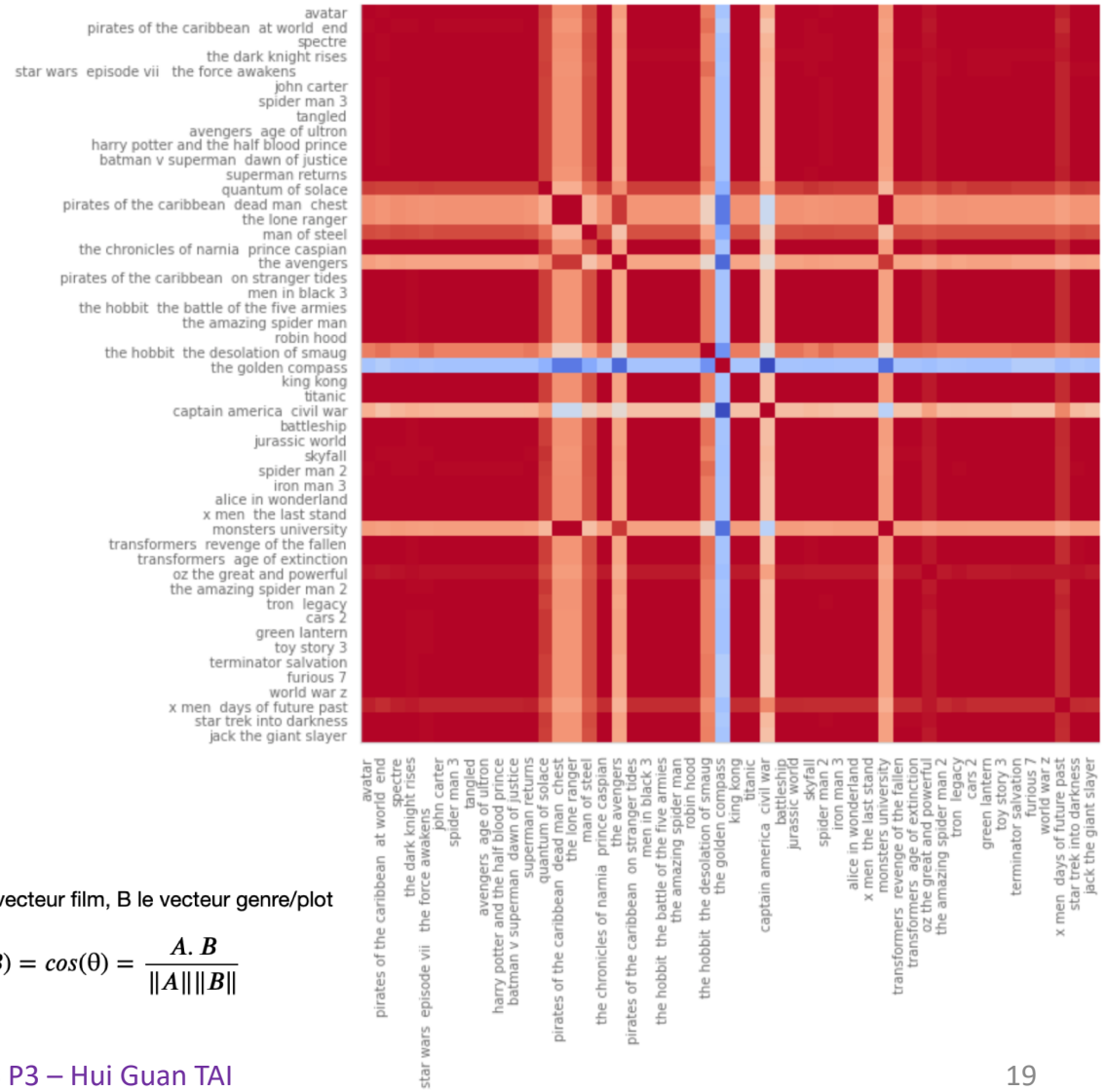
➔ 3 axes: famille (x>0), romance (y<0), thriller/action (y>0)

Exploration – Similarités cosinus

- Similarité sur les genres …



- … sur les plots (factorisés à 10 vecteurs)



Soit A le vecteur film, B le vecteur genre/plot

$$sim(A, B) = cos(\theta) = \frac{A.B}{\|A\|\|B\|}$$

Moteur – Algorithme

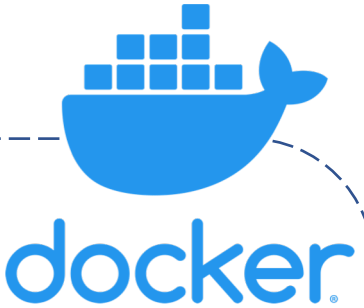
- Similarité sur les genres, sur les plots

```
titre='hollow man'
dfSimGenre=films_genre(titre,mGenres,dfMovies[["movie_title"]]).head(5)
dfSimPlot=films_plot(titre,mPlots,dfMovies[["movie_title"]]).head(5)
```

```
pd.options.display.min_rows=20
print(dfSimGenre)
print(dfSimPlot)
```

```
      simScore                          titre
1239       1.0  aliens vs. predator  requiem
1667       1.0                  ultraviolet
1406       1.0                      species
589        1.0        avp  alien vs. predator
976        1.0     resident evil  apocalypse
     simScore                                  titre
369       1.0           a.i. artificial intelligence
363       1.0                                ben hur
364       1.0              atlantis  the lost empire
365       1.0  alvin and the chipmunks  the road chip
366       1.0                                valkyrie
```

Moteur – Déploiement



FastAPI

- Asynchrone
- Validation de type
- Documentation automatique d'api

HEROKU

- Gratuit
- 1 app web + 1 app backend
- 512 Mb Ram

https://filmroulette.herokuapp.com/docs

Conclusion

Les –

- Une mesure de la pertinence très empirique
- Des plots qui décrivent les films un peu « à coté » et des facteurs latents vagues donnent des recos approximatives
- Les recommandations renvoient sur les même films au bout d'un moment

Les +

- Les films en tête de classement correspondent bien au genre
- Rapidité du moteur (temps de réponse quasi instantané)
- Légèreté (image docker 1,34 Gb)

Idées pour la suite

- Courbe ROC pour mesurer la pertinence des recommandations
- Formule de distance euclidienne, manhattan
- Distances moyennées de plusieurs variables
- Enrichir le dataset de données collaboratives