Présentation de la problématique

Objectif: créer un moteur de recommandations retournant 5 films similaires et intéressants pour le visiteur .

Comment recommander?

- **Grâce au contenu (*les films*) uniquement**:
  - Ce film est similaire à ces films.
- Via d'autres données (*utilisateurs*):
  - Les utilisateurs qui aiment ce film aiment aussi...
  - Les utilisateurs similaire à vous aiment aussi...

Difficultés:

- **Similarité**: trouver une mesure - bien définir ce qui est proche / loin.
- **Popularité**: trouver une mesure - à équilibrer avec la similarité

Présentation de la problématique

Base de données **IMDb**: 5000 films, 28 features

1. **Exploration, nettoyage et feature-engineering**:
   - Obtenir une base de données 'clean'
   - Utilisable pour la recommandation de film

2. **Pistes de modélisation**:
   - Choix de modéle:
     i. K-means Cluster
     ii. Unsupervised Nearest Neighbours
   - Métriques

3. **Modèle final**
   - Performances
   - Améliorations

IMDb

0	movie_id
1	movie_title
2	title_year
3	imdb_score
4	num_voted_users
5	color
6	content_rating
7	language
8	country
9	director_name
10	actor_1_name
11	actor_2_name
12	actor_3_name
13	genres
14	plot_keywords

Nettoyage des données

Gestion des NaN:

- Appel d'API (http://www.omdbapi.com): MAJ de données, **remplissage de champs manquants**
- **Color**: année charnière définie: généralisation de la couleur: 1955
  - Rempli à **'color'** après 1955
  - Rempli à 'black and white' avant 1955
- **Language**: lignes vides remplies à 'English' car 'country'='USA'
- **Remplacement** de NaN par **""** dans les champs *string*
  ['plot_keywords','director_name','actor_1_name','actor_2_name','actor_3_name','content_rating']
- **Suppression** de:
  - 2 lignes ayant 'country' & 'language' vides
  - 4 lignes ayant 'title_year' vide

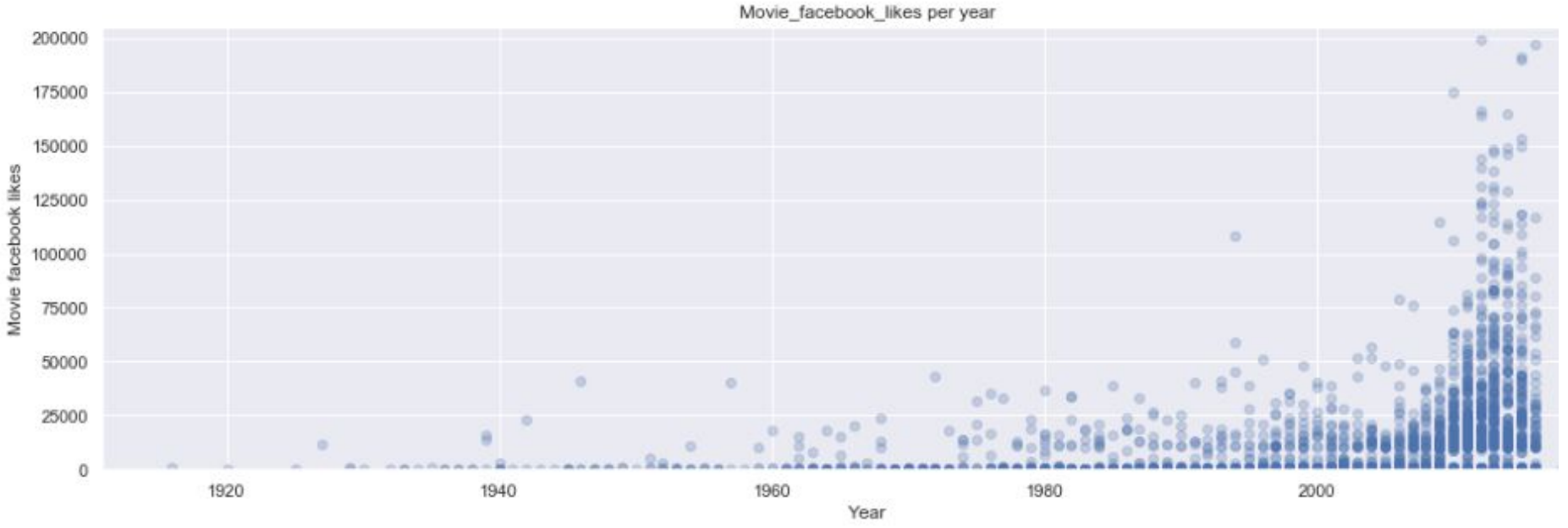
Gestion des valeurs erronées:

- **Suppression** duplicatas - s'appuyant sur l'ID du film
- **Homogénéisation** des catégories de 'content_rating':
  - 'Passed' → 'Approved'
  - 'Unrated' → 'Not Rated'
- 'movie_title': suppression de **caractères unicodes** `'Avatar\xa0'`
- **Correction typage** 'title_year': int → datetime: dt.year

Exploration des données

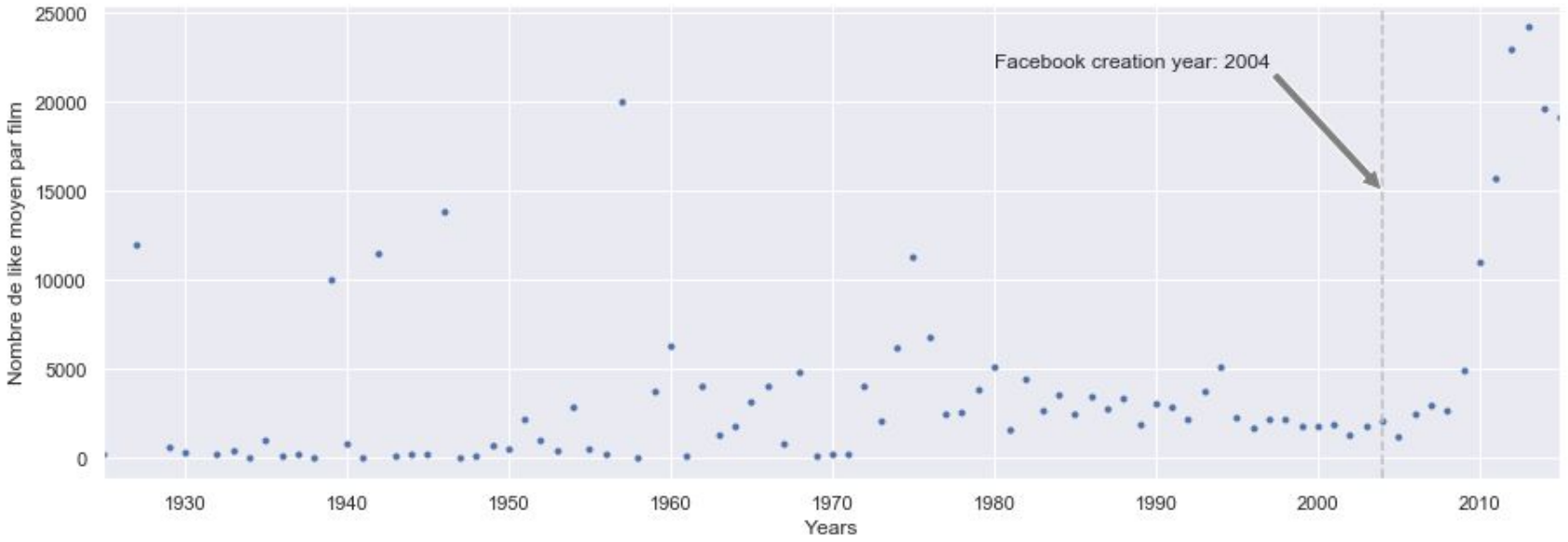
Histogramme des sorties de films:

Évolution des likes facebook:

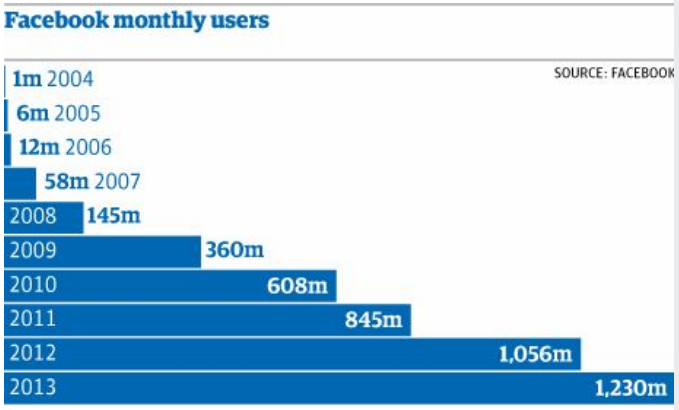


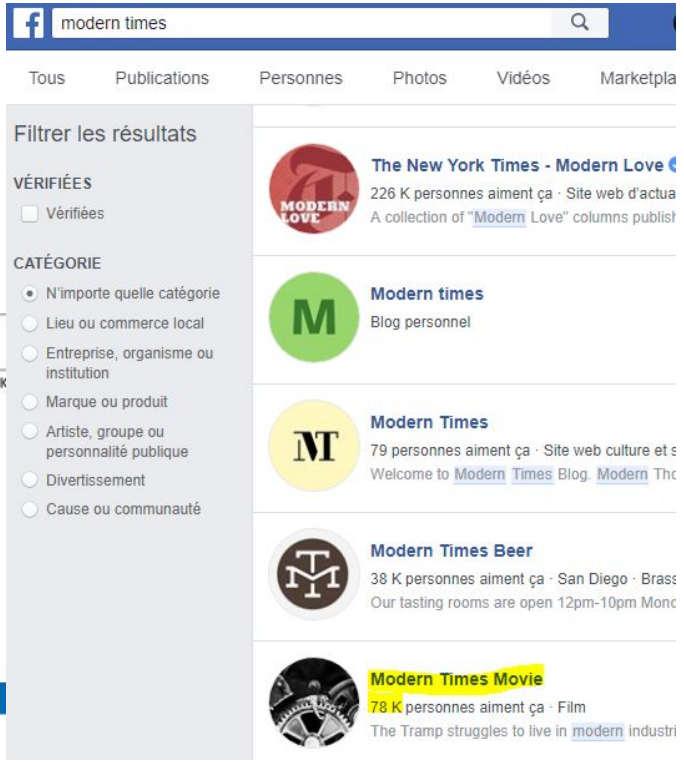
Exploration des données

Évolution du nombre moyen de like par film, par an:



- Vieux films moins likés que les nouveaux
  - à part quelques chef d'oeuvres: The Wizard of Oz, Seven Samurai, 12 angry men, etc.
- Valeurs inconsistantes dans le temps:
  - **Mesure subjective de succès**: 2 films à succès peuvent avoir un nombre de likes très différents selon l'année.
  - **Conditions de départ inégales**: avant/après internet, réseaux sociaux, campagne marketing, visibilité médiatique. Population 'connectée' augmente: les nouveaux films ont encore plus de likes.
  - **Fiabilité**: s'il y a plusieurs groupes facebook (differentes typographie, ou différentes langues)



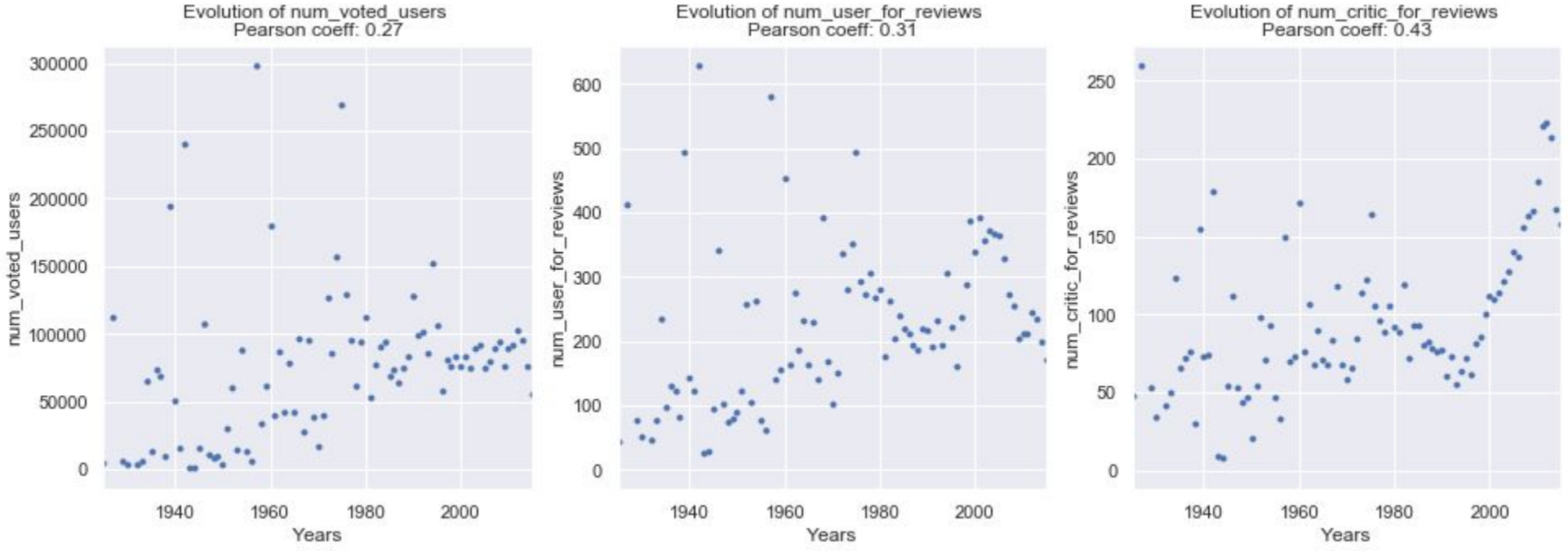


Nettoyage des données

Suppression colonnes:

- 'actor_1_facebook_likes'
- 'actor_2_facebook_likes'
- 'actor_3_facebook_likes'
- 'director_facebook_likes',
- 'movie_facebook_likes'
- 'cast_total_facebook_likes'

- 'num_user_for_reviews'
- 'num_critic_for_reviews'
- 'num_voted_users'



Exploration des données

Choix: **suppression** ‘budget’, ‘gross’ car:

1. Pour utiliser ‘budget’ et ‘gross’, il faut:
   - Corriger avec **l'inflation** chaque année (FMI data)
   - Corriger les **monnaies** (€, francs, ancien franc)
   - Homogénéiser les monnaies: choisir une monnaie unique ($ vu la DB)
   - Gérer les **NaN**: 389 NaN pour ‘budget et 761 pour ‘gross’ (régression / suppression)

2. On a déjà un indicateur de succès fiable: imdb_score ⇒

   ‘imdb_score’ pondéré par ‘num_voted_users’ $IMDB_{score} = \frac{v}{v+m} * R + \frac{m}{v+m} * C$

   $m = minimum\ num\ votes\ required = 100$
   $v = num\_voted\_users$
   $C = mean\ imdb_{score}\ on\ the\ whole\ database$
   $R = imdb\ score_{\ initial}$

title_year	movie_title	budget	imdb_score
1997	Speed 2: Cruise Control	160000000.0	3.7
2003	The Cat in the Hat	109000000.0	3.8
2010	The Last Airbender	150000000.0	4.2
2015	Fantastic Four	120000000.0	4.3
2003	Charlie's Angels: Full Throttle	120000000.0	4.8

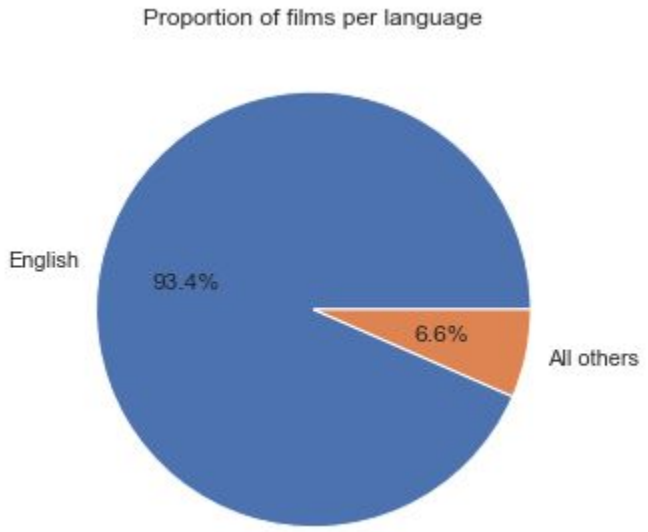
Top of the flop

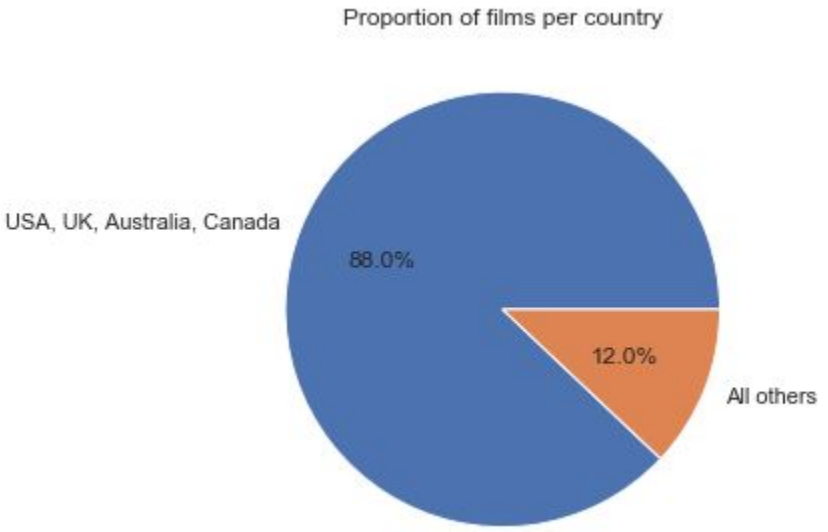
title_year	movie_title	budget	imdb_score
2004	Super Size Me	65000.0	7.3
1964	A Fistful of Dollars	200000.0	8.0
2007	The Man from Earth	200000.0	8.0
1981	The Evil Dead	375000.0	7.6
2013	UnDivided	250000.0	7.8

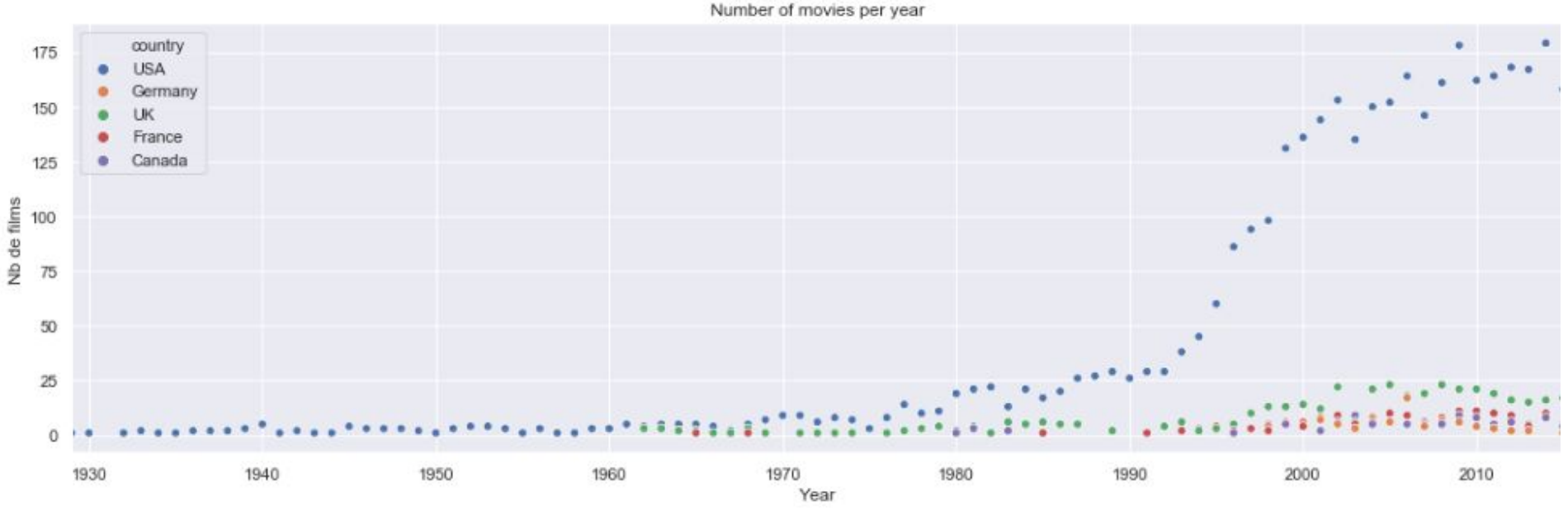
Cheap success

Exploration des données

- Database very English-oriented

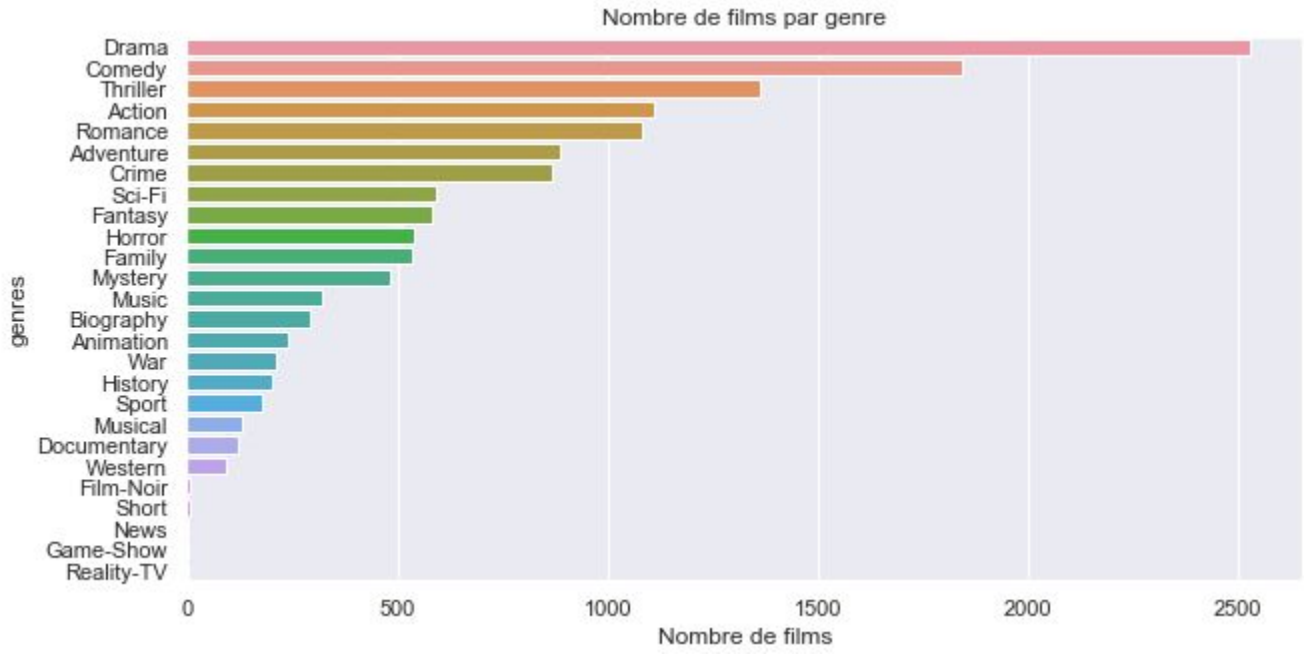






Exploration des données

- 26 genres de film
- Création d'un nouveau genre: 'Serie'
  Empty 'director_name'



- Mot-clefs: 7977

plot_keywords	|love|	|friend|	|murder|	||	|death|	|police|	|new york city|	|high school|	|alien|	|boy|	|school|	|revenge|	|fbi|	|friendship|	|drugs|
count	189	159	153	139	125	117	89	85	77	71	69	68	65	64	64

- Final table:
  - 132 films supprimés
  - 13 features supprimées

```
1  data.count()
```

```
movie_id          4911
movie_title       4911
title_year        4911
imdb_score        4911
num_voted_users   4911
color             4911
content_rating    4911
language          4911
country           4911
director_name     4911
actor_1_name      4911
actor_2_name      4911
actor_3_name      4911
genres            4911
plot_keywords     4911
```

Feature engineering

- **Encodage binaire** des champs 'string': One Hot Encoding
  - genres ⇒ 27 values 8 max per line
  - color ⇒ 2 values 1 max per line
  - content_rating ⇒ 16 values 1 max per line

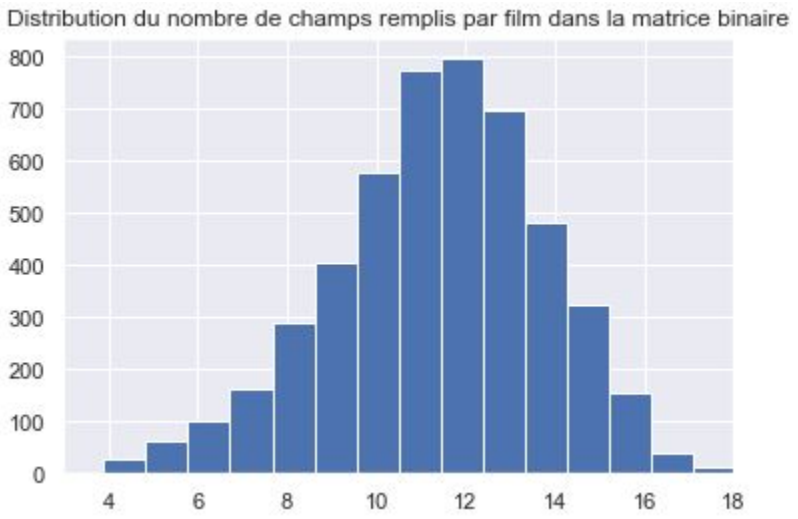
- Encodage et **seuils** minimums - Filtre sur les features:
  - Nb de film min > 2 ⇒ 512 'director_name' 1 max per line
  - Nb de film min > 3 ⇒ 968 'actor_names' 3 max per line
  - Nb de film min > 5 ⇒ 785 'plot_keywords' 5 max per line
  - Nb de film min > 2 ⇒ 29 'language' 1 max per line
  - Nb de film min > 2 ⇒ 37 'country' 1 max per line

- Taille finale:

```
1 X.shape
```

```
(4911, 2375)
```

  - 2 colonnes numériques: 'title_year' et 'imdb_score'
  - 2373 colonnes binaires



Pistes de modélisation

1. **Mesure de similarité:** Sélectionner les 15 films les plus proches du film en entrée
   Sur toute les features sauf le 'imdb_score'

   a. **K-means clustering**:

   - Séparer la base de données en clusters (idéalement des clusters: min 6 films et pas trop important)
   - Sélectionner des films parmi le cluster du film en entrée

   Similarité: films groupés par minimisation de variance intra-cluster
   = maximisation de l'homogénéité des clusters

	cluster1	cluster2	cluster3	cluster4	...	cluster26	cluster27	cluster28	cluster29
cluster size	368	351	336	328	...	24	20	20	10

1 rows × 29 columns

Critique:

- **clusters inégaux** - maximiser le nombre de cluster ⇒ clusters à 1 film
- **clusters trop grand** - 368 films
- **outliers** from 2 clusters might be closer together than the recommendations
- **réduire** ⇒ + de clusters et + homogènes mais **variance expliquée infime**

- *Améliorations possibles:*
  - *faire des clusters par groupe de variables: clusters sur les 'genres', puis clusters sur les 'acteurs', etc. Puis s'intéresser aux intersections de clusters. Modèle moins simple.*
  - *faire du clustering sur certaines variables priorisées (ex: les genres), puis enchaîner par du Nearest Neighbour. Mais autant faire directement du Nearest Neighbour sans clustering, et éviter de prioriser.*

Pistes de modélisation

b. Unsupervised Nearest Neighbours:

- Calcul de distance avec le film en entrée
- Sélectionner les films les plus proches

Aperçu: utilisant la distance euclidienne sur la matrice brute:

movie_title	title_year	language	country	director_name	actor_1_name	actor_2_name	actor_3_name	genres	plot_keywords
Capitalism: A Love Story	2009	English	USA	Michael Moore	Ronald Reagan	Michael Moore	Bernie Sanders	Crime|Documentary|News	capitalism|critique of capitalism|investment b...

movie_title	title_year	movie_id
Breaking Upwards	2009	1247644
I Want Your Money	2010	1560957
Impact Point	2008	1086841
The Final Destination	2009	1144884
Frat Party	2009	1414361

⇒ Nécessité de standardiser, ou séparer matrice binaire/numérique

Pistes de modélisation

Importance du choix de la métrique:

Comparaison de 3 films:

	Godfather	Mongol King	Scarface
actor_name_Al Pacino	1	0	1
actor_name_F. Murray Abraham	0	0	1
director_name_Francis Ford Coppola	1	0	0
genre_Crime	1	1	1
genre_Drama	1	1	1
color_Color	1	1	1
content_rating_PG-13	0	1	0
content_rating_R	1	0	1
country_USA	1	1	1
language_English	1	1	1
plot_keywords_1950s	1	0	0
plot_keywords_cocaine	0	0	1

Distance Euclidienne: $\sqrt{\sum(A_{x_i} - B_{x_i})^2}$

```
Distance Euclidienne [Godfather Mongol King] : 2.83
Distance Euclidienne [Godfather Scarface] : 3.0
```

Similarité Cosinus: $\frac{|A \cap B|}{||A||.||B||}$

```
Similarité Cosinus [Godfather Mongol King] : 0.59
Similarité Cosinus [Godfather Scarface] : 0.61
```

Similarité Jaccard: $\frac{|A \cap B|}{|A \cup B|}$

```
Similarité Jaccard [Godfather Mongol King] : 0.38
Similarité Jaccard [Godfather Scarface] : 0.44
```

- Les films ayant des champs vides vides sont avantagés
- La similarité Jaccard amoindri ce problème

Pistes de modélisation

Modèles initiaux testés:

- Matrice standardisée : knn avec distance euclidienne
- Matrice standardisée et réduite: knn avec distance euclidienne
- Séparation matrice binaire/numérique: knn avec similarité cosinus
- Séparation matrice binaire/numérique: knn avec similarité jaccard

Choix: **knn avec similarité jaccard** - à améliorer. pour trouver les films similaires

Puis utilisation du **imdb_score**: pour choisir les films populaires

Évaluation des performances

Exemple de recommandations: Amélie Poulain

	knn1	knn2	knn3	knn4	knn5	knn6	knn7	knn8	knn9
0	The Wash	A Very Long Engagement	When the Cat's Away	When the Cat's Away	When the Cat's Away	A Very Long Engagement	A Very Long Engagement	When the Cat's Away	The Artist
1	L'auberge espagnole	Barbecue	The Names of Love	The Names of Love	The Names of Love	Frances Ha	Paris, je t'aime	The Names of Love	A Very Long Engagement
2	Down and Out with the Dolls	Asterix at the Olympic Games	Micmacs	Micmacs	Micmacs	Paris, je t'aime	The Widow of Saint-Pierre	Micmacs	Molière
3	The Widow of Saint-Pierre	The Case of the Grinning Cat	Welcome to the Sticks	Welcome to the Sticks	Welcome to the Sticks	L'auberge espagnole	L'auberge espagnole	Welcome to the Sticks	Paris, je t'aime
4	Say It Isn't So	Molière	Barbecue	A Very Long Engagement	A Very Long Engagement	The Widow of Saint-Pierre	The Names of Love	A Very Long Engagement	L'auberge espagnole

Échantillon de 10 films: sélection variée

- Crime|Drama classic
- Sci-fy blockbuster
- Documentary
- French Comedy & Romance
- Mandarin & Adventure & History
- Japanese & Anime
- Italy, Western, 1966
- Movie with saga recommendation
- Quentin Tarantino, Bruce Willis
- Wes Anderson movie

	kNN1	kNN2	kNN3	kNN4	kNN5	kNN6	kNN7	kNN8	kNN9
Godfather	3.0	1.5	2.5	3.5	3.5	3.5	3.5	4.5	4.5
Star Wars: Episode I - The Phantom Menace	2.0	2.0	2.0	2.0	2.0	2.5	2.5	3.0	4.0
Capitalism: A Love Story	1.0	4.0	3.0	2.0	4.0	4.0	4.0	3.0	4.0
Amelie	1.0	2.5	3.5	4.5	4.5	4.0	4.0	4.5	5.0
Dragon Blade	1.0	4.0	5.0	5.0	5.0	4.0	4.0	5.0	5.0
Spirited Away	2.5	3.5	3.5	4.0	4.0	4.0	4.0	4.0	4.5
The Good, the Bad and the Ugly	2.0	2.5	3.0	3.0	4.5	5.0	4.5	4.5	4.5
Return to Oz	0.5	1.0	3.5	3.5	3.0	3.5	4.5	3.5	3.5
Pulp Fiction	1.0	1.0	1.5	2.5	3.0	4.0	4.0	4.0	5.0
The Grand Budapest Hotel	0.0	3.0	2.0	2.0	4.0	4.0	5.0	5.0	5.0

	kNN1	kNN2	kNN3	kNN4	kNN5	kNN6	kNN7	kNN8	kNN9
Moyenne	1.4	2.5	2.95	3.2	3.75	3.85	4.0	4.1	4.5

Evaluation personnelle, limitée, donc subjective: il faudrait

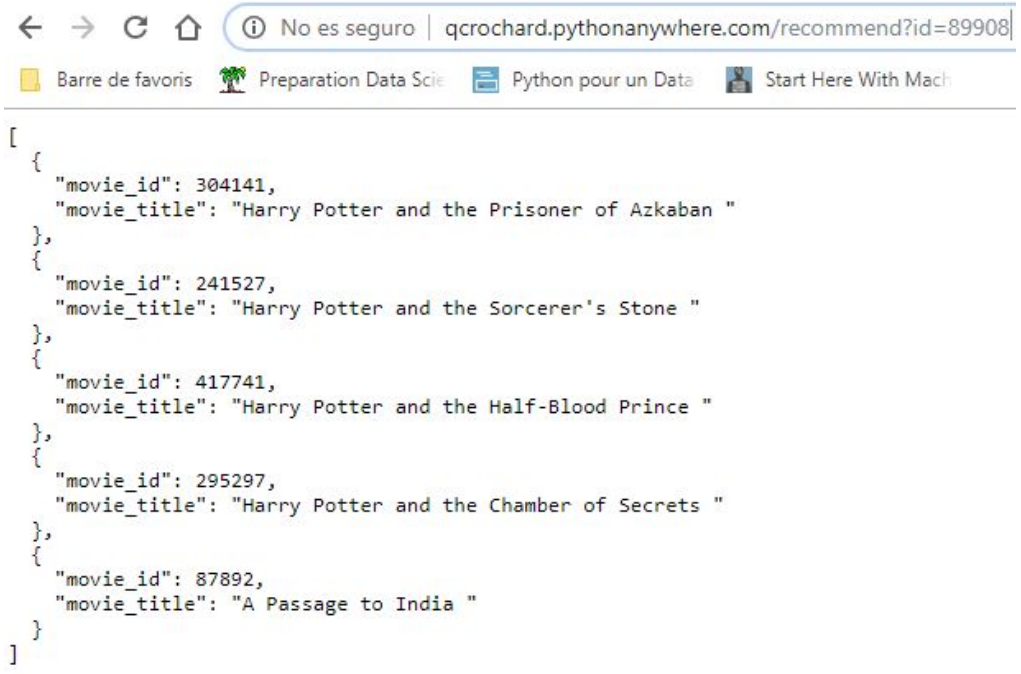
- un panel de notes d'utilisateurs
- comparer les résultats à d'autres moteurs de recommandation

Améliorations

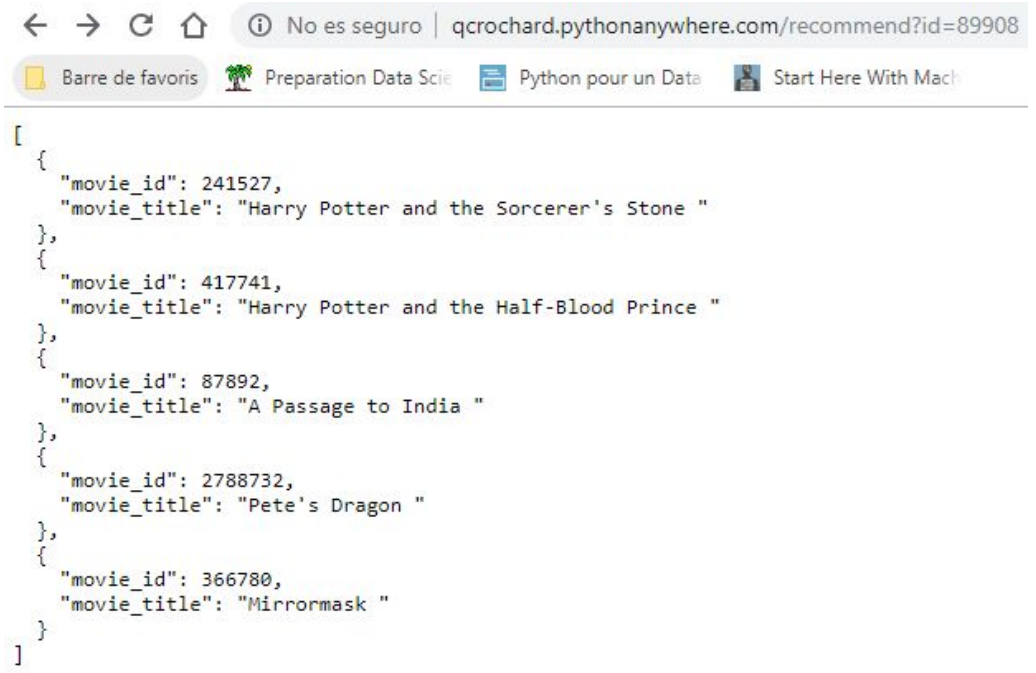
1. **Filtre** sur les **vecteurs quasi-vides**: moins de 6 features renseignées dans la matrice binaire (149 films)

movie_title	title_year	content_rating	country	genres	plot_keywords	director_name	actor_1_name	actor_2_name	actor_3_name
The Case of the Grinning Cat	2004		France	Documentary		Chris Marker	Marina Vlady	Bertrand Cantat	Léon Schwartzenberg
The Little Ponderosa Zoo	2016		USA	Family		Luke Dye	Mike Stanley	Jeff Delaney	Jamison Stalsworth
Mi America	2015	R	USA	Crime|Drama		Robert Fontaine	Michael Derek	Arturo Castro	Brad Lee Wind

2. **Filtrer les sagas/trilogies**: via FuzzyWuzzy: ne garder que 2 recommandations
   - Testé: avec le nom du film sélectionné ⇒ Problème: 'Return to Oz'
   - Amélioré: boucler sur la sélection, avec chaque voisin en entrée




```
[
  {
    "movie_id": 304141,
    "movie_title": "Harry Potter and the Prisoner of Azkaban "
  },
  {
    "movie_id": 241527,
    "movie_title": "Harry Potter and the Sorcerer's Stone "
  },
  {
    "movie_id": 417741,
    "movie_title": "Harry Potter and the Half-Blood Prince "
  },
  {
    "movie_id": 295297,
    "movie_title": "Harry Potter and the Chamber of Secrets "
  },
  {
    "movie_id": 87892,
    "movie_title": "A Passage to India "
  }
]
```


```
[
  {
    "movie_id": 241527,
    "movie_title": "Harry Potter and the Sorcerer's Stone "
  },
  {
    "movie_id": 417741,
    "movie_title": "Harry Potter and the Half-Blood Prince "
  },
  {
    "movie_id": 87892,
    "movie_title": "A Passage to India "
  },
  {
    "movie_id": 2788732,
    "movie_title": "Pete's Dragon "
  },
  {
    "movie_id": 366780,
    "movie_title": "Mirrormask "
  }
]
```

Améliorations

3. Amélioration du kNN:

   - *Testé:* ***priorisation*** *de features lors du kNN:*
     - kNN sur 'genres' puis kNN sur toutes features des 15 films sélectionnés
     - kNN sur 'genres' +'plot_keywords' puis kNN sur toutes features des 15 films sélectionnés
     - Suppression des features: 'content_rating' : à utiliser en filtre utilisateur, pas en similarité.

   Final: avec **unique kNN** sur [**genres, plot_keyword, country, language, director_name, actor_names, color**]
   Et **sélection** des **25 films les + proches**

4. Pondération: entre similarité jaccard (matrice binaire) et distance euclidienne (matrice numérique: title_year)
   Distance finale pondérée = $\alpha * Numerical\ distance + (1 - \alpha) * Categorical\ distance$
   Sélection des **15 films les + proches**

Numerical distance	Categorical distance	Final_dist	movie_title
2.0	0.428571	0.431714	The Godfather: Part II
7.0	0.562500	0.575375	Apocalypse Now
18.0	0.562500	0.597375	The Godfather: Part III
29.0	0.545455	0.602364	L.I.E.
33.0	0.545455	0.610364	The Mongol King
11.0	0.600000	0.620800	Scarface
1.0	0.625000	0.625750	The French Connection
20.0	0.600000	0.638800	Hoffa
35.0	0.571429	0.640286	Brooklyn Rules
11.0	0.625000	0.645750	The Outsiders
25.0	0.611111	0.659889	Donnie Brasco
20.0	0.625000	0.663750	Glengarry Glen Ross
18.0	0.631579	0.666316	Goodfellas
35.0	0.600000	0.668800	We Own the Night

Améliorations

5. Sélectionner les **films à succès**:
   - *Testé:* *Filtre les films avec imdb_score< 6*
     Problème: *parfois pas assez de recommandations*
   - Amélioration: Tri décroissant du imdb_score puis sélection
     Sélection finale: les **5 meilleurs imdb_score**

Final_dist	movie_title
0.431714	The Godfather: Part II
0.575375	Apocalypse Now
0.597375	The Godfather: Part III
0.602364	L.I.E.
0.610364	The Mongol King

⇒

movie_title	weighted_rating
The Godfather: Part II	9.0
Goodfellas	8.7
Apocalypse Now	8.5
Scarface	8.3
The French Connection	7.8

6. **Affinage / calibration**: .
   - *Modifier la sélection de features du kNN (ex: supprimer content_rating)*
   - *Modifier les seuils lors de l'encodage*
   - *Modifier la quantité de films sélectionné à chaque étape de filtre: 25 puis 15 puis 5*
   - *Modifier α dans 'distance finale pondérée*

Autres améliorations - non réalisées

- Faire des filtres sur **genres** opposés:
  Ex le film sélectionné est une 'Comedy', ou 'Family', on peut filtrer 'Horror'
  Pour éviter valeurs erronées

- Travailler sur les **mot-clefs**:
  - Homogénéiser les **189 groupes de mots** (sur 785 mot-clefs retenus)
  - Chercher des groupes principaux de mot-clefs (via fuzzywuzzy levenshtein similarity)

movie_title	plot_keywords	imdb_score
Catwoman	based on cult comic book|bechdel test passed|c...	3.3

movie_title	plot_keywords	imdb_score
The Dark Knight	based on comic book|dc comics|psychopath|star ...	9.0
The Wolverine	healing power|marvel comics|mecha|regeneration...	6.7
Batman Returns	box office hit|dc comics|gotham|mayor|penguin	7.0
Constantine	based on comic|based on comic book|dc arrowver...	7.5
Spawn	based on comic|based on comic book|dark hero|i...	5.2
Batman	city|dc comics|gotham|pantyhose|police	7.6

  - Problème: ⇒ travail manuel

Row	Column	Similarité
new york city	new york	100
	manhattan new york city	100
high school	high school student	100
	high school senior	100
based on novel	based on young adult novel	100
african american	african american protagonist	100
martial arts	mixed martial arts	100
world war two	world war one	85
stand up comedian	stand up comedy	88
hit in the crotch	kicked in the crotch	87
female nudity	male nudity	92
	female frontal nudity	100
	female rear nudity	100
	topless female nudity	100

Row	Column	Similarité
19th century	18th century	92
	17th century	92
	16th century	92
box office flop	box office hit	83
one word title	three word title	83
	two word title	86
	four word title	83
	six word title	83
urban setting	rural setting	85
claim in title	animal in title	83

Modèle final

Pré traitement:

1. Nettoyage base de donnée et **sélection de features**
2. **Encodage** binaire des features string + filtre **seuils** minimums
3. **Split**: matrice binaire / numérique
4. **Filtre** vecteurs quasi-vide matrice binaire

Modèle final:

1. **kNN** sur matrice binaire (metric=Jaccard) + **sélection** des **25 films les + proches**.
2. **Filtre** des **sagas/trilogies**.
3. **Distance finale pondérée** (via title_year) et **sélection** des **15 films les + proches**
4. **Tri** via le **imdb_score** et **sélection** des **5 'meilleurs' films** à succès

Limites du modèle

- Valeurs erronées:
  - Genres:
    - 'Destiny', 'Battlefield Earth' : jeux videos
    - 'Kevin Hart' : in documentary
    - 'Aliens' not : in horror
  - Mot-clefs:
    - Nécessite conventions de nommage (ex: 'Catwoman' not linked to other comics)
    - Star Wars I: mot-clefs pas assez générique:
      - |martial arts|
      - |murdered before giving protagonist information|
      - |part computer animation|
      - |prequel|
      - |prequel to cult film|

- Filtre sur les sagas basé sur ‘movie_title’: ne fonctionne pas toujours.
  Ex: les James Bond

- La base de donnée est limitée à 5000 films:
  - 'Modern Times' il manque les autres films de Chaplin
  - ‘Seven Samurai’ il manque les autres films de Kurosawa

- Erreurs acceptables:
  - certains films sont uniques et indescriptibles:
    - ‘The Fountain’
    - ‘Human Traffic’
  - Les acteurs et directeurs tournant des films hétéroclites pourraient fausser la recommandation
    - Jean-Jacques Annaud (‘L’ours’, ‘Coup de tête’, ‘Le nom de la rose’, etc.)
    - Darren Aronofsky

API

URL: http://qcrochard.pythonanywhere.com/recommend

Query: id

Exemple: *The Godfather*

http://qcrochard.pythonanywhere.com/recommend?id=68646


```
[
  {
    "movie_id": 71562.0,
    "movie_title": "The Godfather: Part II "
  },
  {
    "movie_id": 99685.0,
    "movie_title": "Goodfellas "
  },
  {
    "movie_id": 78788.0,
    "movie_title": "Apocalypse Now "
  },
  {
    "movie_id": 86250.0,
    "movie_title": "Scarface "
  },
  {
    "movie_id": 67116.0,
    "movie_title": "The French Connection "
  }
]
```