Moteur de recommandations de films

Thomas Weber

Introduction

- Un site sur le cinema souhaite réaliser un moteur de recommandations de films.
- Source des données: IMDB (5 000 films)
- Objectif: tester plusieurs modèles de recommandations et mettre en ligne une API utilisant celui avec les meilleurs résultats.

Quel type de problème ?

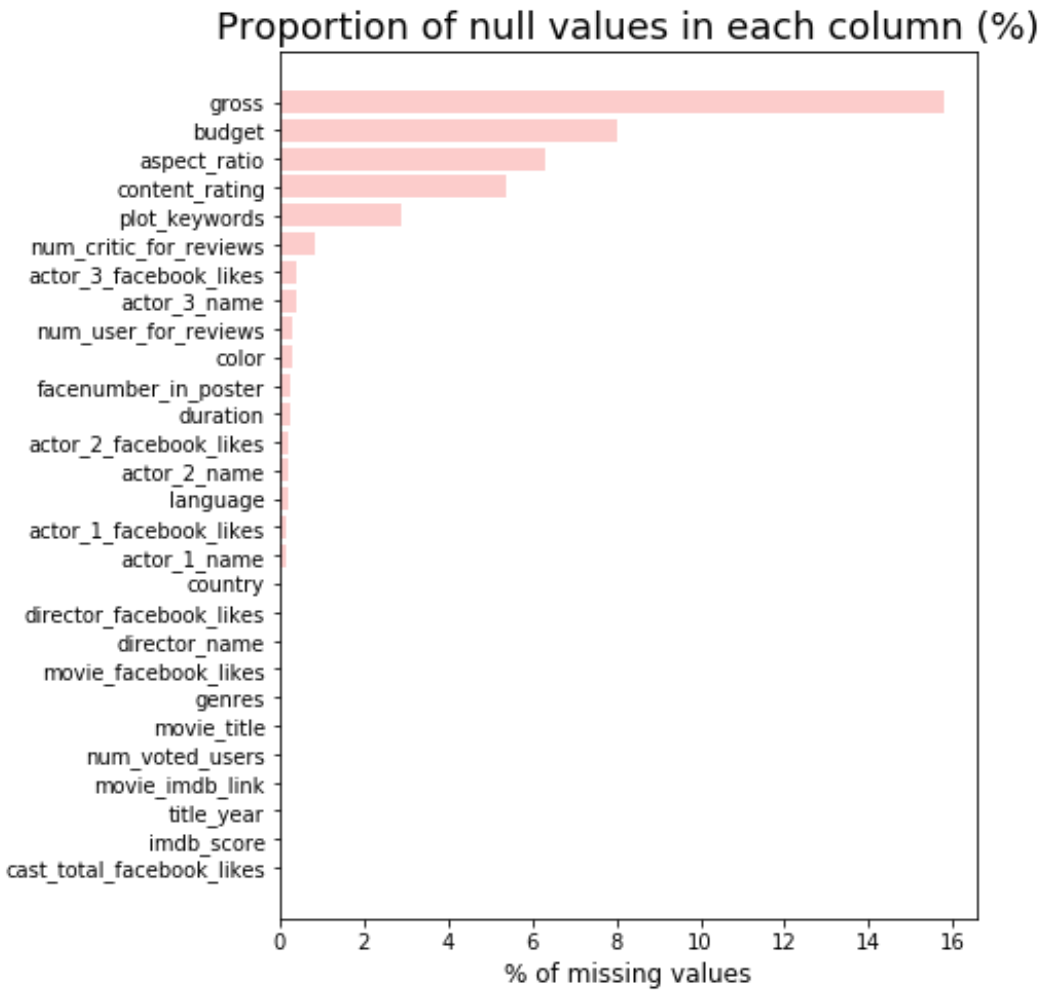
- Content-based filtering
- Non-supervisé
- Solutions de type k-NN

Qu’est-ce que 2 films similaires ?

Variables choisies pour définir la similarité	**Variables non choisies**
Genre (ex: Action/Drame/Comédie) Mots-clés Langue Pays Note IMDB (score entre et 1 et 10) Année Durée Popularité (nombre de votes pour le score IMDB)	Facebook likes (film, réalisateur, acteurs) Couleur ou Noir&Blanc Format d’image Restriction d’âge (ex: films interdits aux – de 16 ans) Nom des acteurs, du réalisateur Budget Recettes Nombre de critiques Titre du film Nombre de visages sur l’affiche

Nettoyage

- Au départ: 5043 lignes
- Elimination des doublons:
  - 124 lignes
- Elimination des titres sans année
  - Séries ou émissions de TV
  - 106 lignes
- A la fin: 4813 lignes



Traitement des valeurs manquantes

- Variables numériques:
  - Durée du film: imputation par la moyenne
  - Autres données numériques: 0
- Variables textuelles: chaîne de caractères vide
- Variables catégorielles:
  - Imputation à la main

Variables numériques: recette et budget

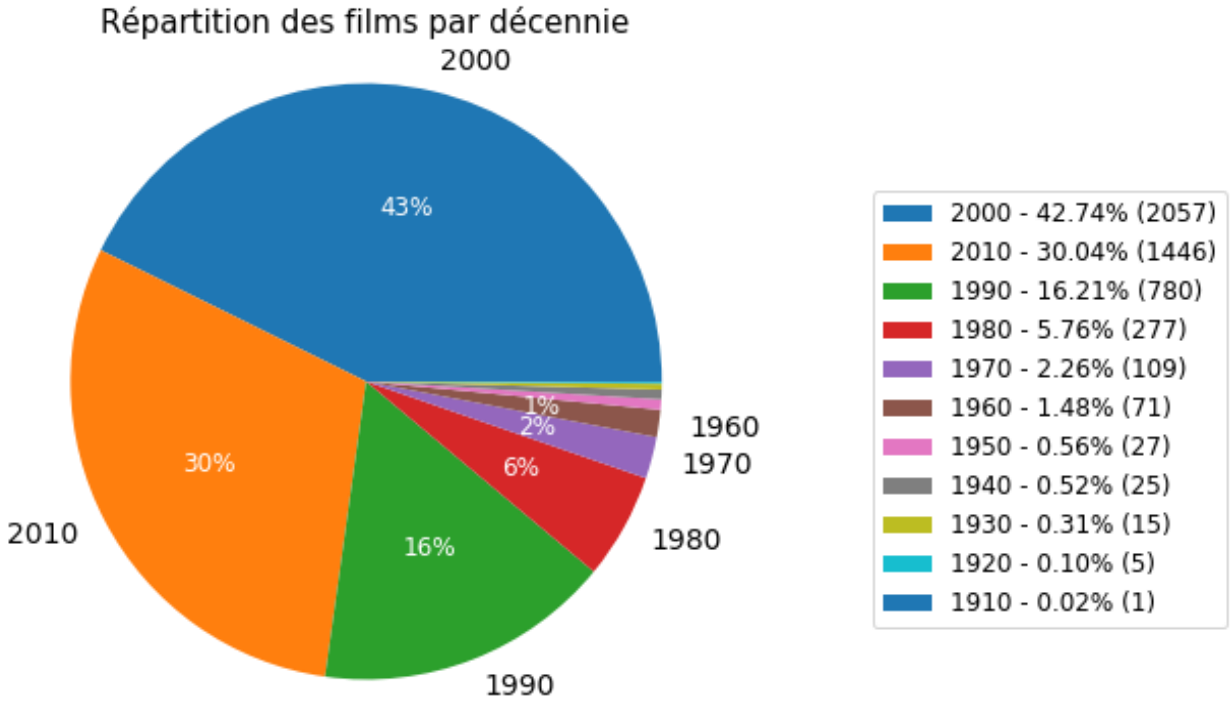
- Recettes:
  - Uniquement aux Etats-Unis
  - Beaucoup de valeurs manquantes
- Budget:
  - Des résultats en devise étrangère
  - Beaucoup de valeurs manquantes

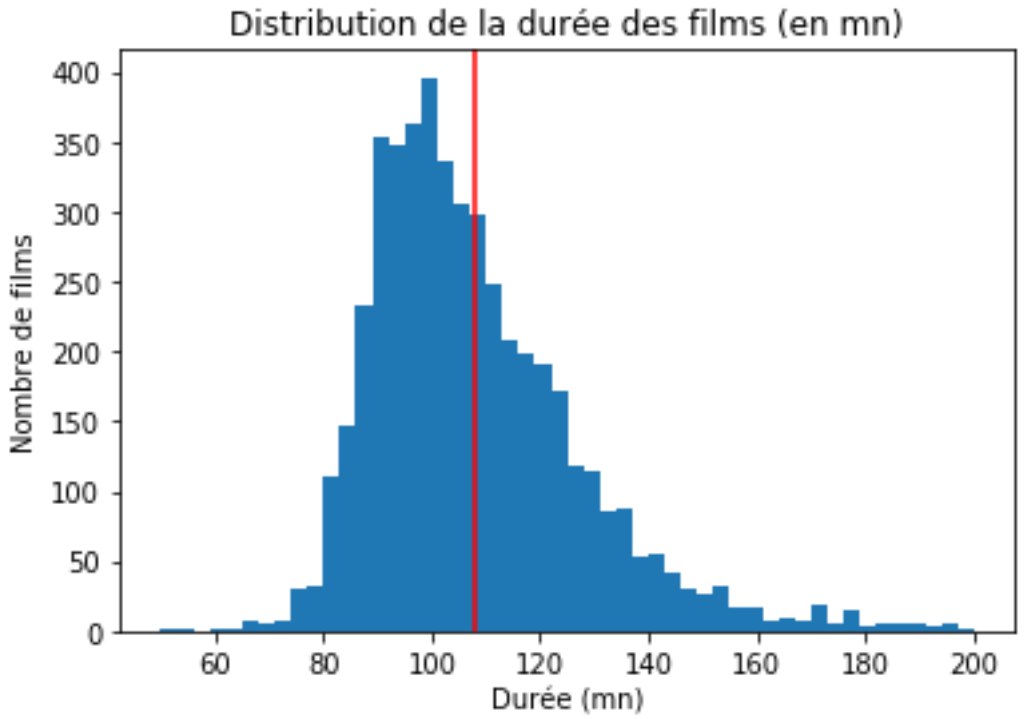
```
In [12]: df.nlargest(10, 'budget')[['budget', 'movie_imdb_link', 'movie_title']]
```

Out[12]:

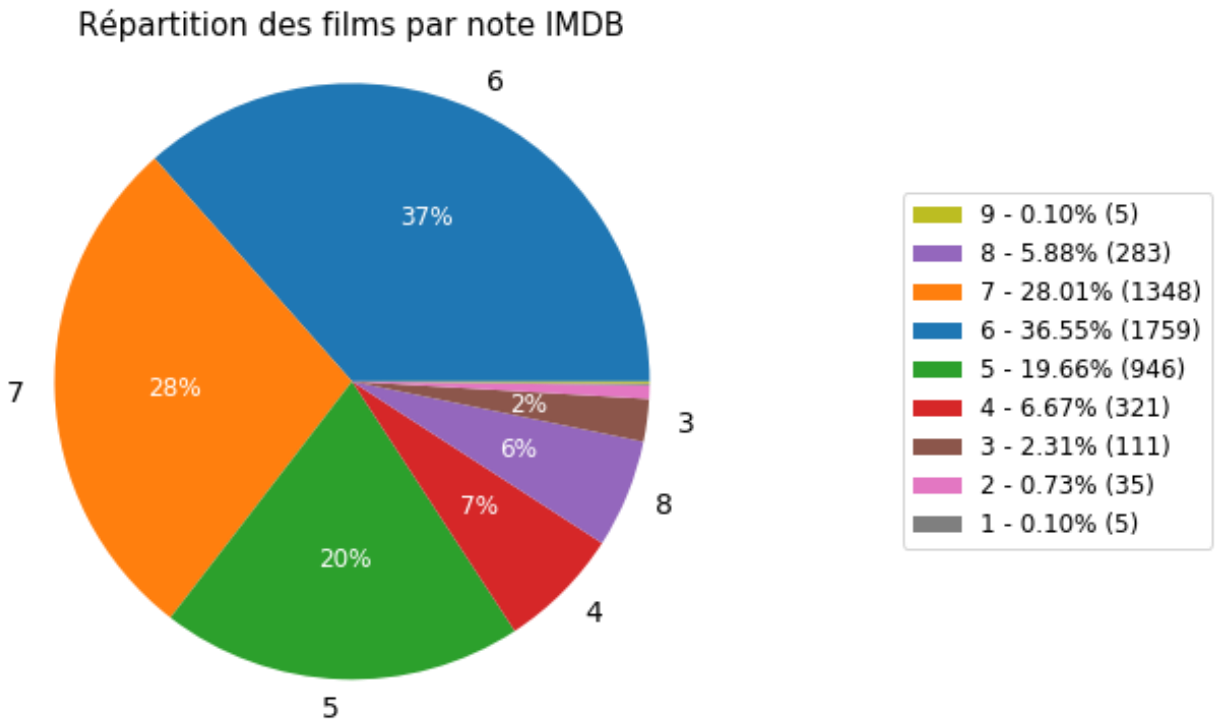
	budget	movie_imdb_link	movie_title
2820	1.221550e+10	http://www.imdb.com/title/tt0468492/?ref_=fn_t...	The Host
3660	4.200000e+09	http://www.imdb.com/title/tt0451094/?ref_=fn_t...	Lady Vengeance
2837	2.500000e+09	http://www.imdb.com/title/tt0367082/?ref_=fn_t...	Fateless
2189	2.400000e+09	http://www.imdb.com/title/tt0119698/?ref_=fn_t...	Princess Mononoke

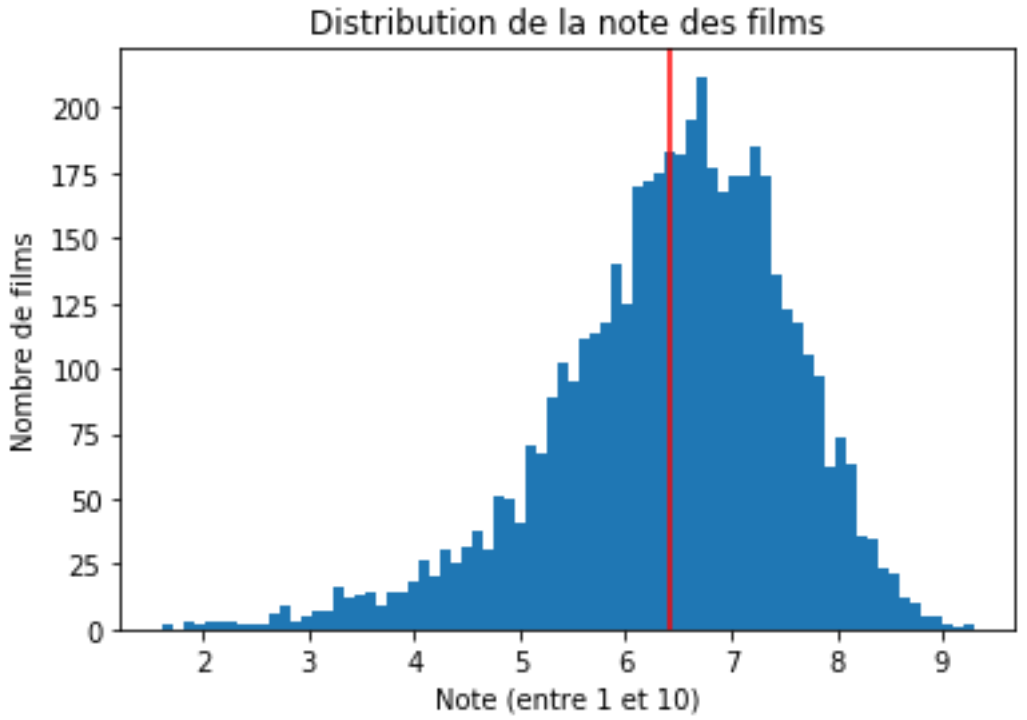
Variables numériques: année et durée





Variables numériques: note IMDB





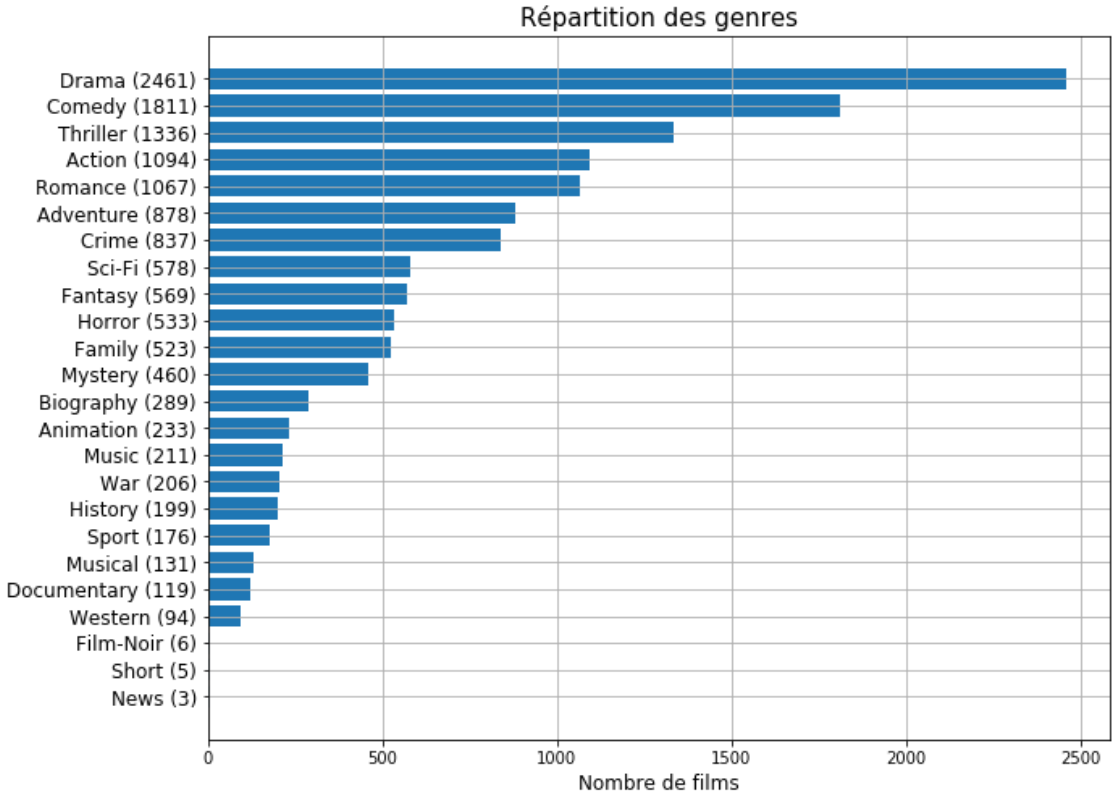
Variables catégorielles: genre

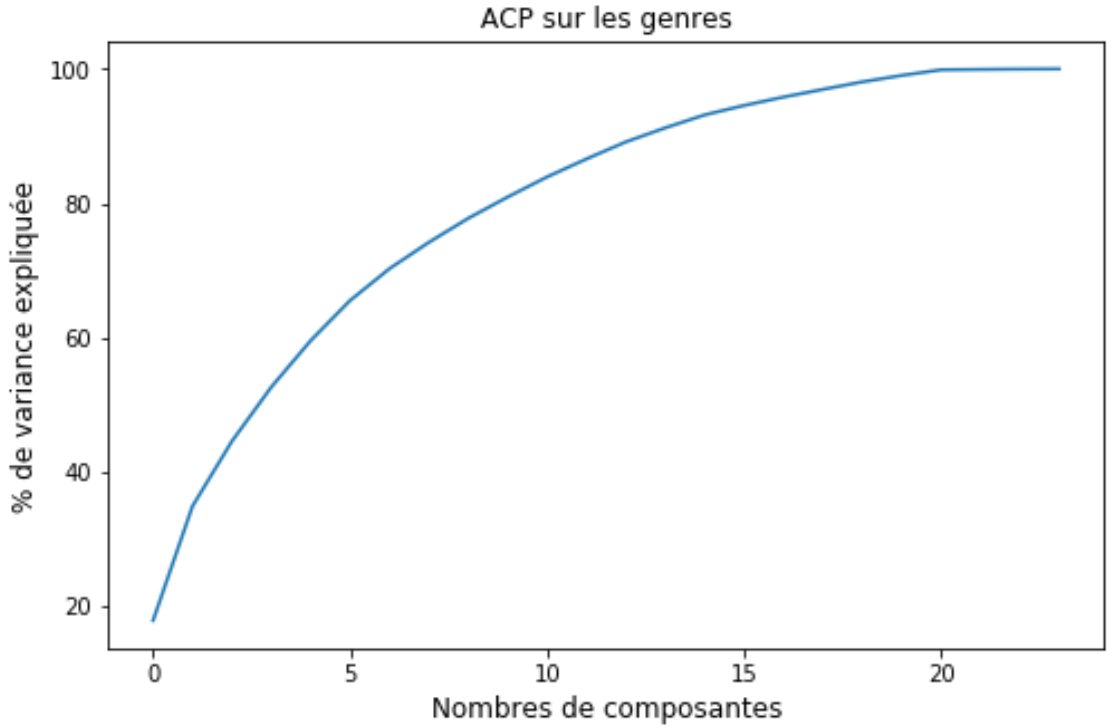
One-hot encoding

Nom du film	Genres
Film X	Drama
Film Y	Comedy|Romance
Film Z	Drama|Thriller|Western

Nom du film	Drama	Comedy	Romance	Thriller	Western
Film X	1	0	0	0	0
Film Y	0	1	1	0	0
Film Z	1	0	0	1	1

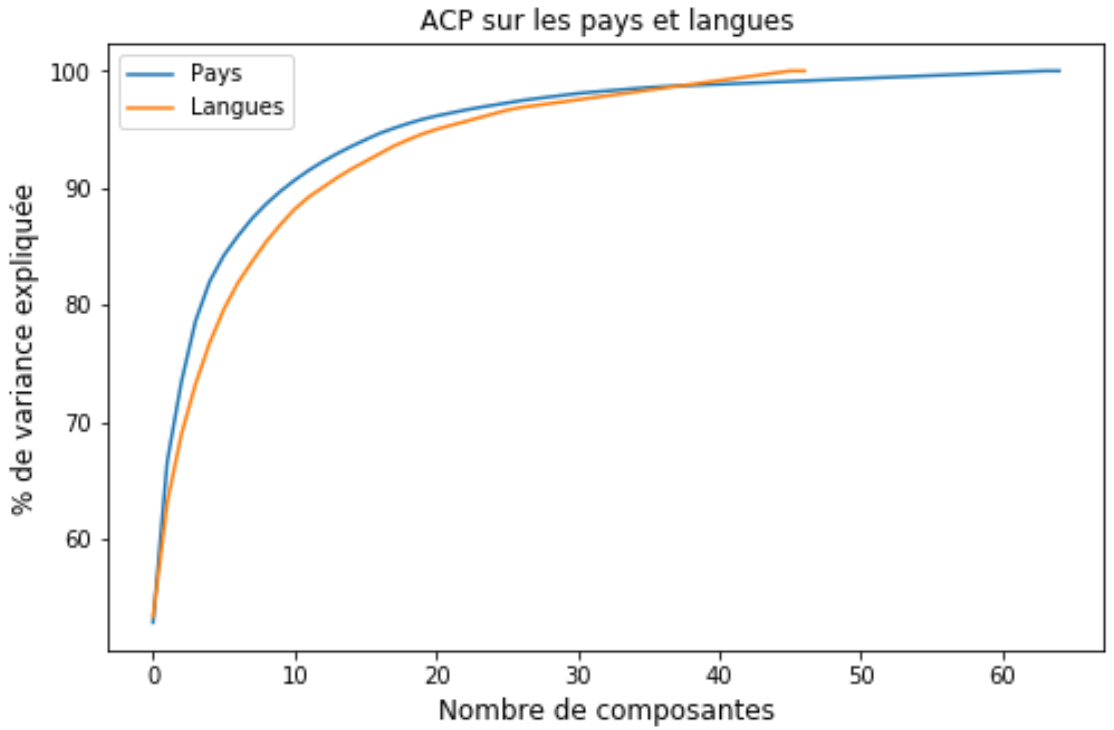
Variables catégorielles: genre





Variables catégorielles: langue et pays

- 47 langues / 65 pays
- 3 options pour les intégrer:
  - One-hot encoding (comme pour le genre)
  - ACP en ne gardant qu'une vingtaine de composantes
  - Transformer les variables:
    - Pour la langue, ne garder que les 19 les plus représentées et mettre les autres sous « Other »
    - Pour le pays, faire des regroupements par région géographique pour limiter le nombre de groupes



Variables textuelles

- 5 variables textuelles: réalisateur + 3 acteurs + mots-clés

- Nombre de valeurs uniques:
  - Réalisateur: 2395
  - Acteurs: 6120
  - Mots-clés: 7978

- ACP non recommandée

Méthode n°1: Principe

- K-NN simple:
  - Nombre de voisins k=16
- Suppression des suites de films
- Plusieurs datasets testés:
  - Minimal: score IMDB, année, durée et nombre de votes
  - Genres: minimal + genres
  - Pays: minimal + pays
  - Langues: minimal + langue
  - All: minimal + genres + pays + langue

Méthode n°1: Résultats

- Meilleur dataset: ‘Genres’

	movie_title	movie_imdb_link	imdb_score	title_year	duration	num_voted_users	genres	country	language
0	Avatar	http://www.imdb.com/title/tt0499549/?ref_=fn_t...	7.9	2009.0	178.0	886204	Action|Adventure|Fantasy|Sci-Fi	USA	English
13	Man of Steel	http://www.imdb.com/title/tt0770828/?ref_=fn_t...	7.2	2013.0	143.0	548573	Action|Adventure|Fantasy|Sci-Fi	USA	English
213	Star Wars: Episode III - Revenge of the Sith	http://www.imdb.com/title/tt0121766/?ref_=fn_t...	7.6	2005.0	140.0	520104	Action|Adventure|Fantasy|Sci-Fi	USA	English
738	The Avengers	http://www.imdb.com/title/tt0848228/?ref_=fn_t...	8.1	2012.0	173.0	995415	Action|Adventure|Sci-Fi	USA	English
185	Pirates of the Caribbean: The Curse of the Bla...	http://www.imdb.com/title/tt0325980/?ref_=fn_t...	8.1	2003.0	143.0	809474	Action|Adventure|Fantasy	USA	English
39	X-Men: Days of Future Past	http://www.imdb.com/title/tt1877832/?ref_=fn_t...	8.0	2014.0	149.0	514125	Action|Adventure|Fantasy|Sci-Fi|Thriller	USA	English

Méthode n°1: Limites

	movie_title	movie_imdb_link	imdb_score	title_year	duration	num_voted_users	genres	country	language
1218	Amélie	http://www.imdb.com/title/tt0211915/?ref_=fn_t...	8.4	2001.0	122.0	534262	Comedy|Romance	France	French
1723	The 40-Year-Old Virgin	http://www.imdb.com/title/tt0405422/?ref_=fn_t...	7.1	2005.0	133.0	313797	Comedy|Romance	USA	English
1688	Knocked Up	http://www.imdb.com/title/tt0478311/?ref_=fn_t...	7.0	2007.0	133.0	298590	Comedy|Romance	USA	English
2010	Silver Linings Playbook	http://www.imdb.com/title/tt1045658/?ref_=fn_t...	7.8	2012.0	122.0	533607	Comedy|Drama|Romance	USA	English
1956	There's Something About Mary	http://www.imdb.com/title/tt0129387/?ref_=fn_t...	7.1	1998.0	107.0	247289	Comedy|Romance	USA	English
2276	Birdman or (The Unexpected Virtue of Ignorance)	http://www.imdb.com/title/tt2562232/?ref_=fn_t...	7.8	2014.0	119.0	395087	Comedy|Drama|Romance	USA	English

Méthode n°2: Principe

- Création d'un dataset 'à la volée' par-rapport au film recherché:

	imdb_score	title_year	duration	num_voted_users	language	country	kw_0	kw_1	kw_2	kw_3	kw_4	genre_0	genre_1	genre_2	genre_3
0	7.9	2009.0	178.0	886204	True	True	True	True	True	True	True	True	True	True	True
126	4.2	2010.0	103.0	118951	True	True	True	False	False	False	False	True	True	True	False
4791	3.0	2011.0	143.0	125	True	True	True	False	False	False	False	False	False	False	False
551	5.8	2011.0	116.0	154955	True	True	False	False	True	False	False	True	False	False	True
705	6.2	2014.0	105.0	99035	True	True	False	False	True	False	False	True	False	False	False
1087	6.3	2002.0	115.0	30077	True	True	False	False	True	False	False	False	False	False	False

- Avantages:
  - On évite le one-hot encoding et de donner trop de poids au genre
  - On intègre les mots-clés pour plus de précision

Méthode n°2: Résultats et limites

	movie_title	movie_imdb_link	imdb_score	title_year	duration	num_voted_users	genres	country	language	plot_keywords
0	Avatar	http://www.imdb.com/title/tt0499549/?ref_=fn_t...	7.9	2009.0	178.0	886204	Action|Adventure|Fantasy|Sci-Fi	USA	English	avatar|future|marine|native|paraplegic
126	The Last Airbender	http://www.imdb.com/title/tt0938283/?ref_=fn_t...	4.2	2010.0	103.0	118951	Action|Adventure|Family|Fantasy	USA	English	avatar|fire|kingdom|tribe|water
4791	The Ridges	http://www.imdb.com/title/tt1781935/?ref_=fn_t...	3.0	2011.0	143.0	125	Drama|Horror|Thriller	USA	English	avatar|college|death|tron|university
551	Battle Los Angeles	http://www.imdb.com/title/tt1217613/?ref_=fn_t...	5.8	2011.0	116.0	154955	Action|Sci-Fi	USA	English	alien|extraterrestrial|invasion|marine|mission
705	Jack Ryan: Shadow Recruit	http://www.imdb.com/title/tt1205537/?ref_=fn_t...	6.2	2014.0	105.0	99035	Action|Drama|Thriller	USA	English	covert analyst|marine|russian|spy|stock market
1087	High Crimes	http://www.imdb.com/title/tt0257756/?ref_=fn_t...	6.3	2002.0	115.0	30077	Crime|Drama|Mystery|Thriller	USA	English	defense lawyer|lawyer|marine|murder|villager

	imdb_score	title_year	duration	num_voted_users	language	country	kw_0	kw_1	kw_2	kw_3	kw_4	genre_0	genre_1	genre_2	genre_3
0	1.330645	0.526411	3.101105	5.756094	0.265083	0.567993	40.041645	10.788431	28.304888	69.368581	69.368581	1.84376	2.117021	2.731062	2.706841
126	-1.981109	0.606718	-0.224209	0.252222	0.265083	0.567993	40.041645	-0.092692	-0.035330	-0.014416	-0.014416	1.84376	2.117021	2.731062	-0.369434
4791	-3.055191	0.687025	1.549292	-0.600173	0.265083	0.567993	40.041645	-0.092692	-0.035330	-0.014416	-0.014416	-0.54237	-0.472362	-0.366158	-0.369434
551	-0.548999	0.687025	0.352179	0.510496	0.265083	0.567993	-0.024974	-0.092692	28.304888	-0.014416	-0.014416	1.84376	-0.472362	-0.366158	2.706841
705	-0.190972	0.927947	-0.135534	0.109355	0.265083	0.567993	-0.024974	-0.092692	28.304888	-0.014416	-0.014416	1.84376	-0.472362	-0.366158	-0.369434
1087	-0.101465	-0.035740	0.307841	-0.385313	0.265083	0.567993	-0.024974	-0.092692	28.304888	-0.014416	-0.014416	-0.54237	-0.472362	-0.366158	-0.369434

Méthode n°3: Principe

1. Création du dataset ‘à la volée’ (comme méthode n°2)
2. Premier k-NN sur genres/mots-clés/langue/pays, sans standardisation
3. Deuxième k-NN sur durée/année/nombre de votes
4. Pondération avec un coefficient qui dépend de la note IMDB:

$$\text{nouvelle distance} = \text{ancienne distance} * \frac{10}{\text{note IMDB}}$$

Méthode n°3: Résultats

	movie_title	movie_imdb_link	imdb_score	title_year	duration	num_voted_users	genres	country	language	plot_keywords
0	Avatar	http://www.imdb.com/title/tt0499549/?ref_=fn_t...	7.9	2009.0	178.0	886204	Action|Adventure|Fantasy|Sci-Fi	USA	English	avatar|future|marine|native|paraplegic
13	Man of Steel	http://www.imdb.com/title/tt0770828/?ref_=fn_t...	7.2	2013.0	143.0	548573	Action|Adventure|Fantasy|Sci-Fi	USA	English	based on comic book|british actor playing amer...
213	Star Wars: Episode III - Revenge of the Sith	http://www.imdb.com/title/tt0121766/?ref_=fn_t...	7.6	2005.0	140.0	520104	Action|Adventure|Fantasy|Sci-Fi	USA	English	elongated cry of no|friends become enemies|kic...
109	X-Men Origins: Wolverine	http://www.imdb.com/title/tt0458525/?ref_=fn_t...	6.7	2009.0	119.0	361924	Action|Adventure|Fantasy|Sci-Fi|Thriller	USA	English	army|civil war|claw fight|commando|wolverine t...
16	Men in Black 3	http://www.imdb.com/title/tt1409024/?ref_=fn_t...	6.8	2012.0	106.0	268154	Action|Adventure|Comedy|Family|Fantasy|Sci-Fi	USA	English	alien|criminal|m.i.b.|maximum security prison|...
389	Hellboy II: The Golden Army	http://www.imdb.com/title/tt0411477/?ref_=fn_t...	7.0	2008.0	120.0	208422	Action|Adventure|Fantasy|Horror|Sci-Fi	USA	English	creature|elf|prince|rebellion|superhero

Méthode n°3: Limites

	movie_title	movie_imdb_link	imdb_score	title_year	duration	num_voted_users	genres	country	language	plot_keywords
2132	Capitalism: A Love Story	http://www.imdb.com/title/tt1232207/?ref_=fn_t...	7.4	2009.0	105.000000	35137	Crime|Documentary|News	USA	English	capitalism|critique of capitalism|investment b...
4059	Inside Job	http://www.imdb.com/title/tt1645089/?ref_=fn_t...	8.3	2010.0	105.000000	55382	Crime|Documentary	USA	English	florida|iceland|interview|new york city new yo...
2333	Madea Goes to Jail	http://www.imdb.com/title/tt1142800/?ref_=fn_t...	4.1	2009.0	103.000000	9544	Comedy|Crime|Drama	USA	English	adaptation directed by original author|cross d...
1416	Black Water Transit	http://www.imdb.com/title/tt0490087/?ref_=fn_t...	7.2	2009.0	108.056863	219	Crime|Drama	USA	English	based on novel
2362	ATL	http://www.imdb.com/title/tt0466856/?ref_=fn_t...	6.0	2006.0	105.000000	8522	Comedy|Crime|Drama|Music|Romance	USA	English	high school|rollerskating rink|spelman college...
2218	The Frozen Ground	http://www.imdb.com/title/tt2005374/?ref_=fn_t...	6.4	2013.0	105.000000	43879	Crime|Drama|Mystery|Thriller	USA	English	anchorage alaska|based on true story|pole danc...

Méthode n°4: Principe

- Pour le genre, utiliser le one-hot encoding:
  - Sélection des films avec le moins de genres différents en priorité
  - Poids important sur le genre (24 variables sur une trentaine)

- Sinon même principe que la méthode n°3 avec 2 k-NNs successives sur des variables différentes.

Méthode n°4: Résultats

	movie_title	movie_imdb_link	imdb_score	title_year	duration	num_voted_users	genres	country	language	plot_keywords
0	Avatar	http://www.imdb.com/title/tt0499549/?ref_=fn_t...	7.9	2009.0	178.0	886204	Action|Adventure|Fantasy|Sci-Fi	USA	English	avatar|future|marine|native|paraplegic
185	Pirates of the Caribbean: The Curse of the Bla...	http://www.imdb.com/title/tt0325980/?ref_=fn_t...	8.1	2003.0	143.0	809474	Action|Adventure|Fantasy	USA	English	caribbean|curse|governor|pirate|undead
83	Guardians of the Galaxy	http://www.imdb.com/title/tt2015381/?ref_=fn_t...	8.1	2014.0	121.0	682155	Action|Adventure|Sci-Fi	USA	English	bounty hunter|outer space|raccoon|talking anim...
1927	Star Wars: Episode V - The Empire Strikes Back	http://www.imdb.com/title/tt0080684/?ref_=fn_t...	8.8	1980.0	127.0	837759	Action|Adventure|Fantasy|Sci-Fi	USA	English	duel|famous twist|rebel|rescue|snowy landscape
13	Man of Steel	http://www.imdb.com/title/tt0770828/?ref_=fn_t...	7.2	2013.0	143.0	548573	Action|Adventure|Fantasy|Sci-Fi	USA	English	based on comic book|british actor playing amer...
2343	Aliens	http://www.imdb.com/title/tt0090605/?ref_=fn_t...	8.4	1986.0	154.0	488537	Action|Adventure|Sci-Fi	USA	English	alien|human versus alien|monster|rescue missio...

Méthode n°4: Résultats

	movie_title	movie_imdb_link	imdb_score	title_year	duration	num_voted_users	genres	country	language	plot_keywords
2132	Capitalism: A Love Story	http://www.imdb.com/title/tt1232207/?ref_=fn_t...	7.4	2009.0	105.0	35137	Crime|Documentary|News	USA	English	capitalism|critique of capitalism|investment b...
4059	Inside Job	http://www.imdb.com/title/tt1645089/?ref_=fn_t...	8.3	2010.0	105.0	55382	Crime|Documentary	USA	English	florida|iceland|interview|new york city new yo...
4687	The Trials of Darryl Hunt	http://www.imdb.com/title/tt0446055/?ref_=fn_t...	7.7	2006.0	106.0	771	Crime|Documentary	USA	English	false accusation|murder|north carolina|trial|w...
4119	Slacker Uprising	http://www.imdb.com/title/tt0850669/?ref_=fn_t...	5.3	2007.0	102.0	2242	Documentary	USA	English	character name in title|election campaign|pres...
4331	Food, Inc.	http://www.imdb.com/title/tt1286537/?ref_=fn_t...	7.9	2008.0	94.0	42389	Documentary	USA	English	farming|flesh eating|food|food industry|gluttony
4325	An Inconvenient Truth	http://www.imdb.com/title/tt0497116/?ref_=fn_t...	7.5	2006.0	96.0	67654	Documentary	USA	English	climate|earth|global warming|science|truth

Choix du modèle final

- Méthodologie:
  - Liste de 10 films:
    - Avatar (2009)
    - Spectre (2015)
    - Toy Story 3 (2010)
    - Waterworld (1995)
    - Destination finale 2 (2003)
    - Cloud Atlas (2012)
    - Amélie Poulain (2001)
    - Bruce Tout-Puissant (2003)
    - A Beautiful Mind (2001)
    - Les Temps Modernes (1936)
  - Pour chaque film, on classe les 4 méthodes

- Conclusions:
  - Dans l'ordre de la moins bonne à la meilleure:
    - 2
    - 1
    - 3
    - 4
  - Les méthodes 1, 3 et 4 donnent au moins un résultat satisfaisant à chaque fois
  - La méthode 4 est la plus consistante dans ses résultats: elle sera donc mise en place avec l'API

Quelques résultats

	movie_title	movie_imdb_link
0	Avatar	http://www.imdb.com/title/tt0499549/?ref_=fn_t...
185	Pirates of the Caribbean: The Curse of the Bla...	http://www.imdb.com/title/tt0325980/?ref_=fn_t...
83	Guardians of the Galaxy	http://www.imdb.com/title/tt2015381/?ref_=fn_t...
1927	Star Wars: Episode V - The Empire Strikes Back	http://www.imdb.com/title/tt0080684/?ref_=fn_t...
13	Man of Steel	http://www.imdb.com/title/tt0770828/?ref_=fn_t...
2343	Aliens	http://www.imdb.com/title/tt0090605/?ref_=fn_t...

	movie_title	movie_imdb_link
2	Spectre	http://www.imdb.com/title/tt2379713/?ref_=fn_t...
139	Mission: Impossible - Ghost Protocol	http://www.imdb.com/title/tt1229238/?ref_=fn_t...
243	Live Free or Die Hard	http://www.imdb.com/title/tt0337978/?ref_=fn_t...
134	Die Another Day	http://www.imdb.com/title/tt0246460/?ref_=fn_t...
2777	Casino Royale	http://www.imdb.com/title/tt0381061/?ref_=fn_t...
155	The World Is Not Enough	http://www.imdb.com/title/tt0143145/?ref_=fn_t...

	movie_title	movie_imdb_link
35	Toy Story 3	http://www.imdb.com/title/tt0435761/?ref_=fn_t...
57	Up	http://www.imdb.com/title/tt1049413/?ref_=fn_t...
215	Monsters, Inc.	http://www.imdb.com/title/tt0198781/?ref_=fn_t...
864	Shrek	http://www.imdb.com/title/tt0126029/?ref_=fn_t...
68	Inside Out	http://www.imdb.com/title/tt2096673/?ref_=fn_t...
47	Brave	http://www.imdb.com/title/tt1217209/?ref_=fn_t...