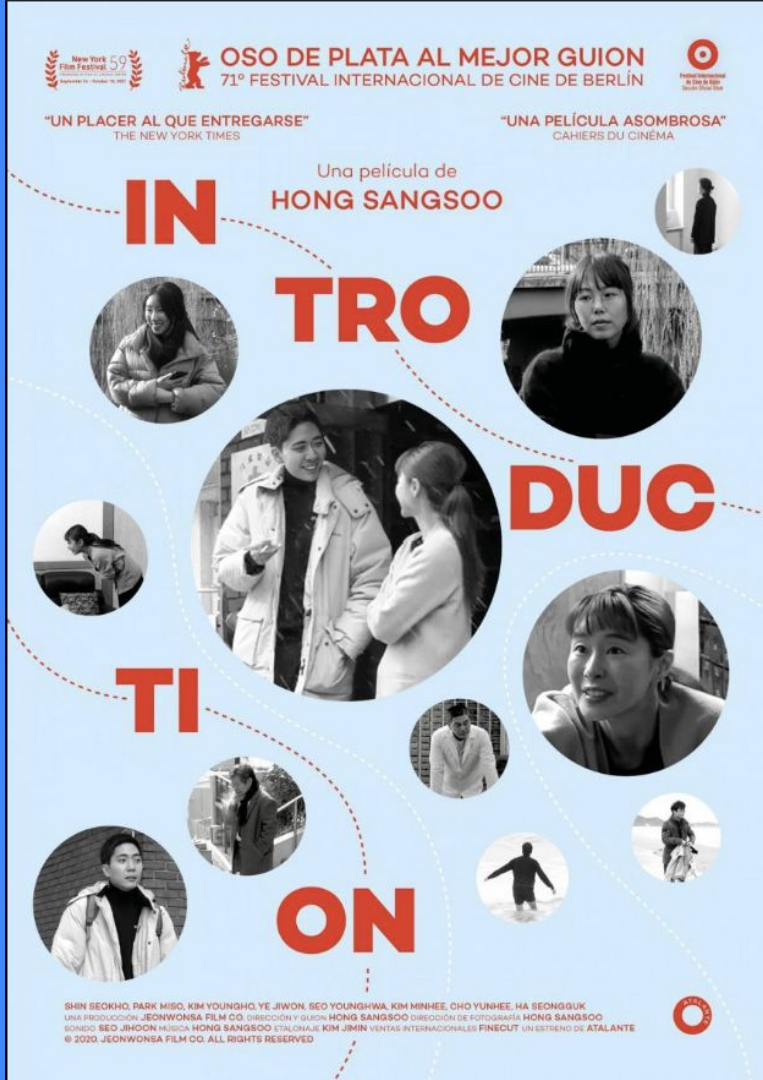
Projet DS

“Système de Recommandation de Films”

Soutenance - 10 décembre 2024

Mentor: Antoine Fradin

Intro


OSO DE PLATA AL MEJOR GUION
71º FESTIVAL INTERNACIONAL DE CINE DE BERLÍN
"UN PLACER AL QUE ENTREGARSE"
THE NEW YORK TIMES
"UNA PELÍCULA ASOMBROSA"
CAHIERS DU CINÉMA
Una película de
HONG SANGSOO
IN
TRO
DUC
TI
ON
SHIN SEOKHO, PARK MISO, KIM YOUNGHO, YE JIWON, SEO YOUNGHWA, KIM MINHEE, CHO YUNHEE, HA SEONGGUK
UNA PRODUCCIÓN JEONWONSA FILM CO. DIRECCIÓN Y GUION HONG SANGSOO DIRECCIÓN DE FOTOGRAFÍA HONG SANGSOO
SONIDO SEO JIHOON MÚSICA HONG SANGSOO ETALONAJE KIM JIMIN VENTAS INTERNACIONALES FINECUT UN ESTRENO DE ATALANTE
© 2020. JEONWONSA FILM CO. ALL RIGHTS RESERVED

Equipe

Ana
Actuaire, future Data Scientist

Expertise en analyse statistique et modélisation et fort intérêt pour l'analyse comportementale.

Lam
Product Manager, futur Ingénieur ML

Expertise en gestion de produit et connaissance approfondie des systèmes de recommandation en tant qu'utilisateur.

Ariel
Ing. Génie Civil, futur Ingénieur ML

Expertise en optimisation de processus, techniques et architecture de systèmes orienté solution.

Charles
Maître d'Ouvrage IT, futur Data Scientist

Expérience en IT et compréhension des systèmes de recommandation dans un contexte bancaire.

Contexte

Définition d’un système de recommandation de films

Une application logicielle ou un algorithme conçu pour suggérer des films pertinents aux utilisateurs en fonction de leurs préférences, de leurs comportements passés ou des tendances globales.

Objectifs

Proposer de nouveaux films à découvrir

Améliorer l’engagement sur une plateforme

Proposer rapidement des choix personnalisés

Exploration

Exploration des Données

Jeux de données

+ IMDb

movielens

MovieLens

- 138k utilisateurs,
- 27k **films**,
- 20M ratings
- 6 fichiers CSV pour structurer les données
- Période couverte : jan 1995 - mars 2015

IMDb

- 5M films et séries,
- 1.5M votes,
- Variables descriptives par **film et série**:
  - Réalisateurs
  - Personnel lié aux films (y compris acteurs)
  - Films associées au personnel
  - Langues
  - Pays
  - Etc
- 7 fichiers TSV.gz pour structurer les données

Rapport d'exploration de données complet disponible ici : Rapport d'exploration de données

Traitement des données

Prise de décision ➜ **Focus sur un système de recommandation des films exclusivement**

Traitement des données divisé en deux étapes :

Filtrage collaboratif :

- Utilisateur : userid
- Film : MovieId & title_name
- Evaluation : Rating

Variables IMDb incluses (voie du peuple):

- Votes Imdb par film
- Nombre de votes IMDb par film

➜ Grâce à ces variables, la création d'un score de pertinence a été créé (explication à posteriori)

Filtrage basé sur le contenu :

Récupération des caractéristiques des éléments (basées sur les informations IMDb) :

- Réalisateurs
- Acteurs
- Films associés aux réalisateurs et aux acteurs

Pre-Processing et Feature Engineering

Les points clés du traitement des données :

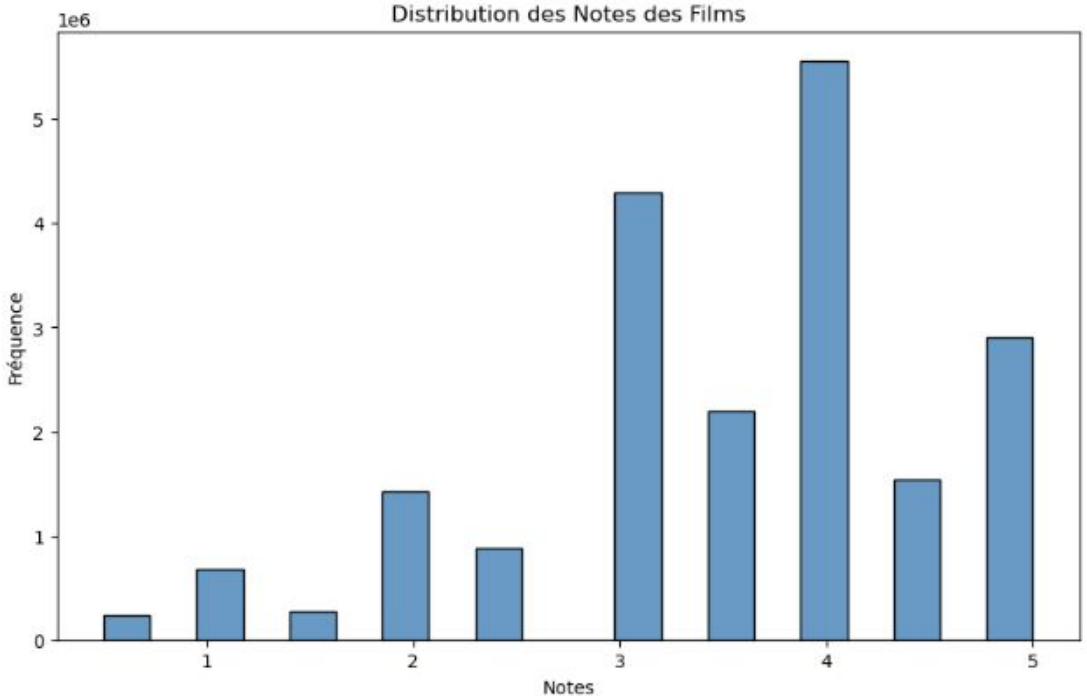
- **Suppression des données** dans les informations de title.basics.tsv :
  630 lignes (sur 11M) ont été supprimées de title.basics.tsv en raison de décalages entraînant des incohérences. Ce fichier fournit des informations de base sur les films (titre, genre).

- **Transformation des variables** :
  Recherche dans IMDb pour identifier réalisateurs, acteurs et films associés, afin de créer des descriptions de films pour tester le **filtrage basé sur le contenu.**

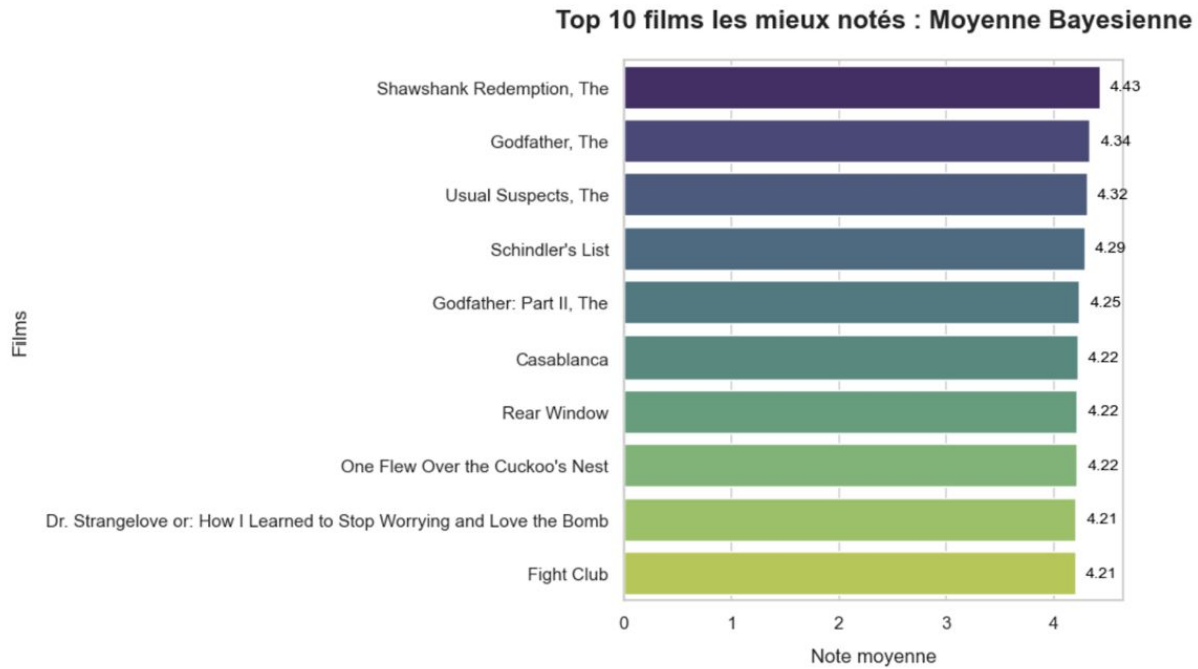
- **Filtre principale appliqué** :
  - **Utilisateurs :** filtrage des utilisateurs trop ou peu actifs pour limiter les biais et réduire la taille des données (500 et 700 évaluations).

Statistiques descriptives Principaux Graphiques

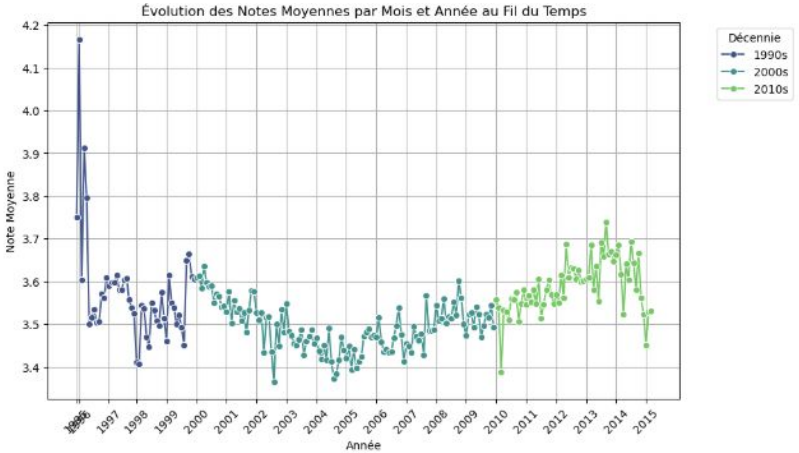
Distribution des notes



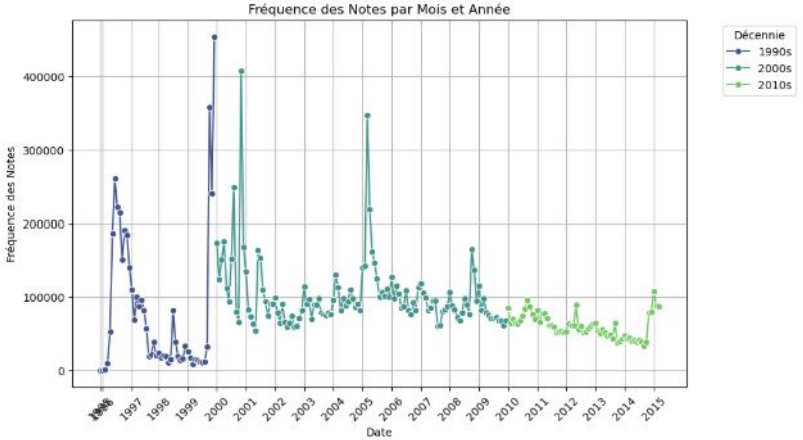
Top 10 des films avec les meilleures notes dans la base de données MovieLens compte tenu de la moyenne bayésienne



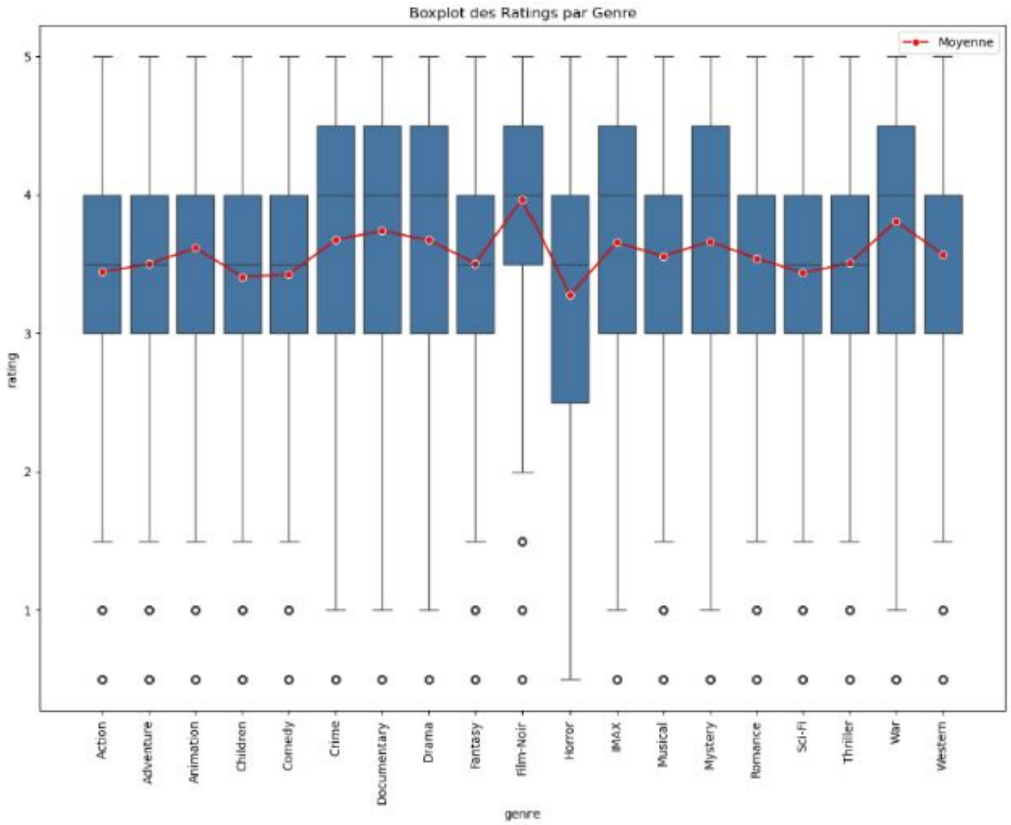
Évolution des notes dans le temps (moyenne)



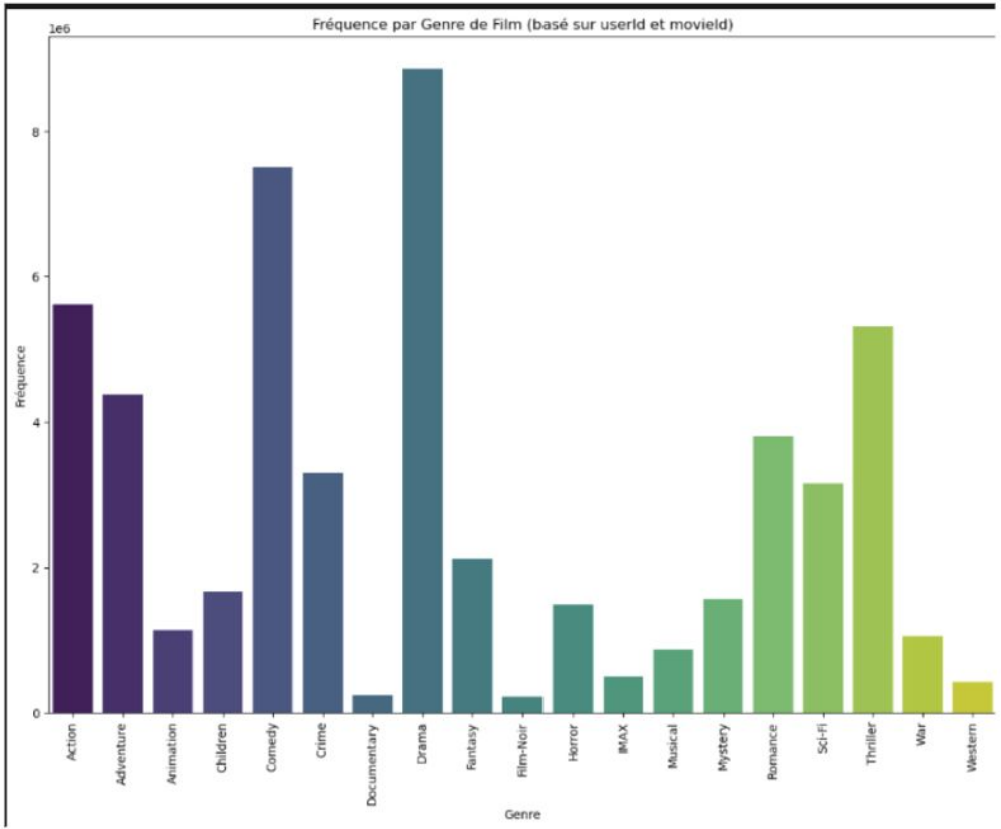
Évolution des notes dans le temps (fréquence)



Distribution des genres



Fréquence par genre des films



Modèles testés

- **Filtrage Collaboratif : approche mémoire**
  - **User-based**
  - **Item-based**
- **Filtrage Collaboratif : approche modèle**
  - **Item-based + SVD**
  - **Surprise (SVD)**
- **Filtrage basé sur le contenu**

Métriques utilisées

- RMSE Root Mean Squared Error
- MAE Mean Absolute Error

Modélisation

Filtrage basé sur le Contenu

Méthode :

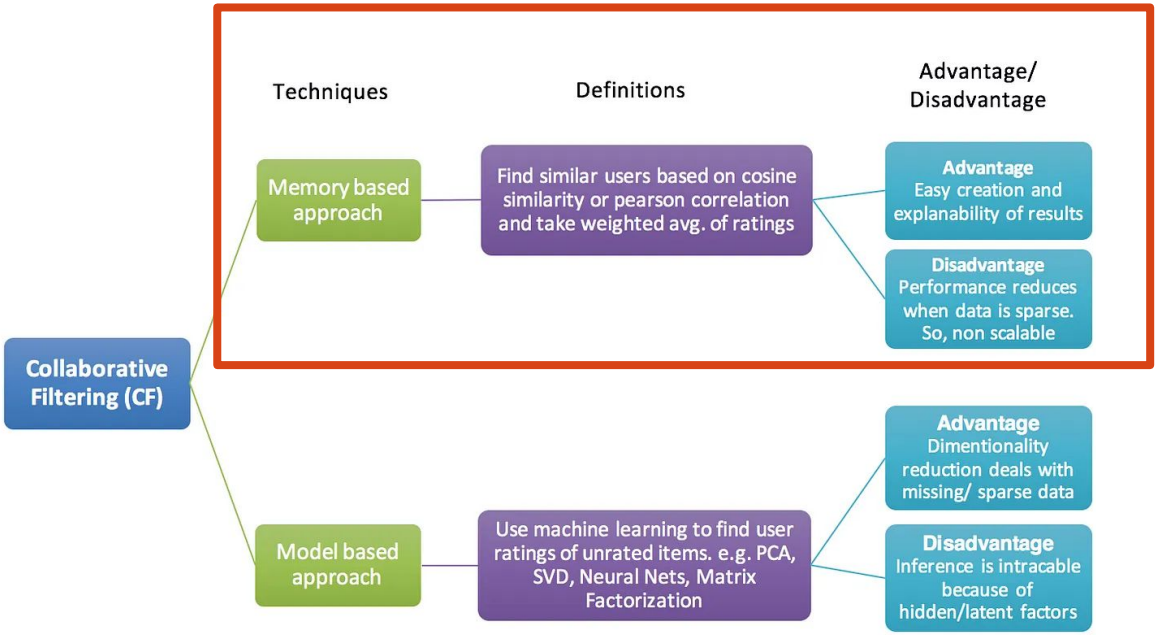
- Création d'une nouvelle variable texte **"description"** (Titre, Réalisateur, Acteurs, Genre, Films connus)
- **Tokenisation** + **Vectorisation**
- **Calcul de similarité** sur les vecteurs (Cosinus et Distance Euclidienne)

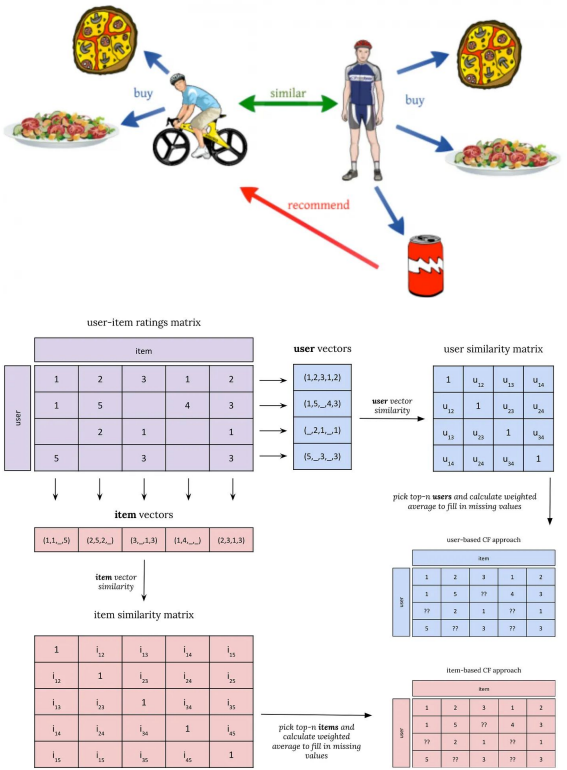
Résultats :

- **Précision faible** (score de similarité, pas de prédictions)
- **Temps de calcul relativement rapide** mais problème de taille de la base de données
- **Interprétabilité mitigée:** évidente pour les premiers résultats, plus ténue ensuite

- **Possibilité d'amélioration** avec une **description plus robuste** (synopsis, tags, etc.)
- Mais **risque d'alourdissement** de la base de donnée

Filtrage Collaboratif (Approche Mémoire)





Filtrage Collaboratif (Approche Modèle)

Factorisation matricielle e.g. SVD (Single Value Decomposition):
le but du SVD est d'apprendre les matrices réduites telles que leur produit est une bonne approximation de la matrice de notation complète.

Modèle Sélectionné : FC Surprise + SVD

Création d'un **score de pertinence (normalisé)**

- Prise en compte de **"la voix du Peuple"** (note moyenne et nombre de votes IMDb))
- **Modèle plus discriminant** pour un meilleur classement des recommandations

$$P = \left((0.4 * N) + (0.6 * M)\right) * log_{10}(\sqrt[3]{V})$$

	userId	imdbId	rating	title_name	imdb_averageRating	imdb_numVotes
960	11	114709	4.5	Toy Story	8.3	1088953
961	11	113189	2.5	GoldenEye	7.2	273041
962	11	112281	3.5	Ace Ventura: When Nature Calls	6.4	237008

Ajout de **recommandations hors des sentier battus** (aléatoire parmi les meilleurs films)

Résultats:

- La **meilleure précision** ET la **meilleure scalabilité**
- L'assurance de **plus de diversité** et la **prise en compte du cold start**
- Une **bonne interprétabilité** (matrice de notation factorisée plus "sagesse populaire")

Benchmark

	Modèle	RMSE	MAE	Temps de calcul de l'evaluation
0	SVD (Surprise)	15.2936	12.3635	Quelques secondes
1	Item-based + SVD (modele)	15.5542	12.2443	Plusieurs heures
2	User-based (approche mémoire)	16.5321	14.0516	30 min
3	Item-based (approche mémoire)	15.8114	13.1578	30 min
4	Filtrage basé sur le contenu	Non applicable	Non applicable	RAS

Meilleure précision:
=> **Surprise + SVD**, possibilité d'amélioration à la marge i.e. **optimisation des hyperparamètres**

Possibilité d'une précision encore meilleure avec une **modélisation plus avancée** (e.g. Deep Learning) MAIS:

=> Les **modèles marchent** déjà bien
=> **Pas besoin de précision "chirurgicale"** pour un système de recommandation
=> Les **gains de performance** sont **incertains**

Démo !


DEMONS
TRATION
UN FILM BALLET DE VICTOR KOSSAKOVSKY Y 32 ESTUDIANTES DEL MÁSTER
EN DOCUMENTAL DE CREACIÓN DE LA UNIVERSITAT POMPEU FABRA
DIRECTED BY: IGNACIO ACCONCIA, CRISTINA ALONSO, GARBIÑE ARMENTIA, IRENE BARTOLOMÉ, DANIEL CALDERÓN, GABRIEL CASANOVA, NEREA CASTRO, YOOJIN CHOI, ANGÉLICA CLAVIJO, CRESPO, GAETANO CRIVARO, DEBORA DE SÁ, SANTIAGO D. RISCO, DARIO FERRARO, SERGI FERRÉS, VÍCTOR FORNÉS, MARTA GIL, NATALIA GUALY, TANIA HERNÁNDEZ, MARINA LAMEIRO, MARGALEF, RAUL MARTÍNEZ, PABLO MARTÍNEZ, PERE MARZO, LLUÍS NADAL, CRISTÓFOL OLIVER, HELGA RAFNSDÓTTIR, POL ROIG, DAVID ROJAS, CRISTINA RUBIO, ANDRÉS SANTAMARÍA, SORRIBAS, DANIEL TOUATI, AINARA VERA, VICTOR KOSSAKOVSKY · SOUND DESIGNER: JUAN SÁNCHEZ 'CUIT' · POST-PRODUCTION: FEDERICO DELPERO
DELEGATE PRODUCTION UPF: JORDI BALLÓ · EXCUTIVE PRODUCTION: EVA VILA
An initiative of the Master in Creative Documentary Pompeu Fabra University. Producer by Kossakovsky Film Production / IDEC-UP

Conclusion


FEATURED IN THE COLLECTIVE VOLUME: 3 FROM JABB PICTURES
STEVE DASH
KEVIN VAN HENTENRYCK
"ONCE YOU CHECK IN...
...THERE'S ONLY ONE WAY TO CHECK OUT"
CONCLUSION
JABB PICTURES PRESENTS A MYSTERIOUS MANSION PRODUCTIONS IN ASSOCIATION WITH BACKSEAT CONCEPTIONS
A FILM BY VANESSA ROMANELLI STARRING STEVE DASH KEVIN VAN HENTENRYCK GOLDIE ZWIEBEL ERICA L. FROST
VANESSA ROMANELLI PRODUCED BY VANESSA ROMANELLI ERICA L. FROST STORY BY VANESSA ROMANELLI
WRITTEN BY VANESSA ROMANELLI ERICA L FROST CAMERON SCOTT DIRECTED BY VANESSA ROMANELLI
WWW.MYSTERIOUSMANSIONPRODUCTIONS.COM

Défis et difficultés rencontrées

1. Gestion des bases de données

- **Défi :** Taille massive des bases.
- **Solution :** Utilisation de Dask.
- **Limite :** Insuffisant pour traiter les informations de filtrage basé sur le contenu.

2. Limitations matérielles

Problèmes :

- RAM insuffisante pour certaines données essentielles (acteurs, réalisateurs).
- Suppression de variables importantes (langues, régions).

3. Contraintes de temps et planification

- **Temps limité :** Études et examens parallèles.
- **Planification limitée :** Apprentissage progressif des modules.
- **Modules clés tardifs :** Réduction de dimension, text mining

4. Limitations techniques de Streamlit

- **Taille des bases :** Restriction à 200 MB.
- **Incompatibilités :** Bibliothèques non supportées (e.g., Surprise).
- **Adaptation :** Réécriture partielle du code Python.

Impact global :

- Réduction de la richesse des analyses
- Ajustements nécessaires pour garantir la faisabilité

Perspectives et Axes d'Amélioration

Enrichissement de la base de données

- **Web scraping** pour enrichir la **description des films** (pour filtrage basé contenu et NLP)
- Intégration des **films** les **plus récents** i.e. après 2016 (et pourquoi pas, ajout des séries)
- Incorporation de sources supplémentaires représentant la **"voix du Peuple"** (e.g. Rotten Tomatoes, AlloCiné, etc.)

Ajustement et **réglage fin du score de pertinence** si besoin (selon feedback utilisateur)

- Notre **formule** est très **flexible**
- Nous avons privilégié la **qualité des films**
- On pourrait aussi décider de favoriser plus de **diversité** ou une **personnalisation** plus poussée

Modélisation plus avancée i.e. **Deep Learning**

- **NLP**
- Prédiction de rating utilisant des **réseaux de neurones** (meilleure précision potentielle)

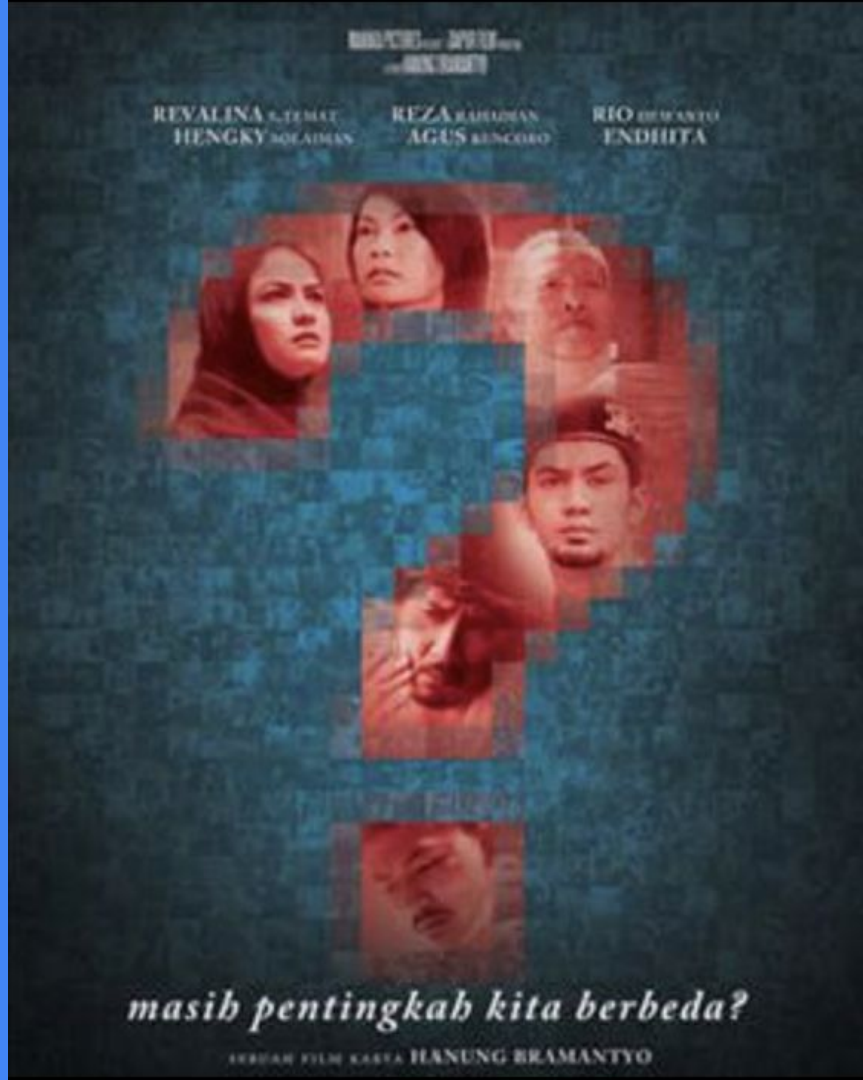
Interprétabilité

- Mise en oeuvre de SHAP pour identifier les variables les plus influentes dans notre modèle

Merci !


DIANNE WIEST
JANE BIRKIN
SIMON CALLOW
JERRY HALL
VANESSA REDGRAVE
BULLE OGIER
and introducing
STANISLAS MERHAR
MERCHANT IVORY PRODUCTIONS
PRESENTS
MERCI
DOCTEUR
REY
MERCIDOCTEURREY.COM

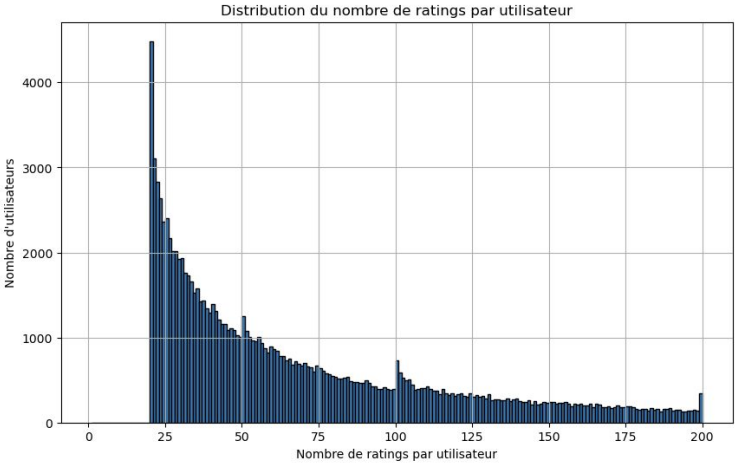
Questions

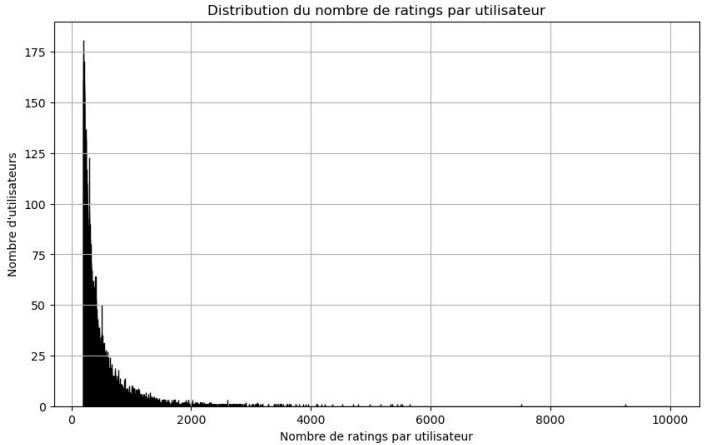

REVALINA S. TEMAT
HENGKY SOLAIMAN
REZA RAHADIAN
AGUS KUNCORO
RIO DEWANTO
ENDHITA
masih pentingkah kita berbeda?
SEBUAH FILM KARYA HANUNG BRAMANTYO

Annexes

Réduction de dimension par rapport aux utilisateurs

Distribution du nombre de notes par utilisateur





Dispersion des moyennes des ratings selon le nombre de ratings des utilisateurs

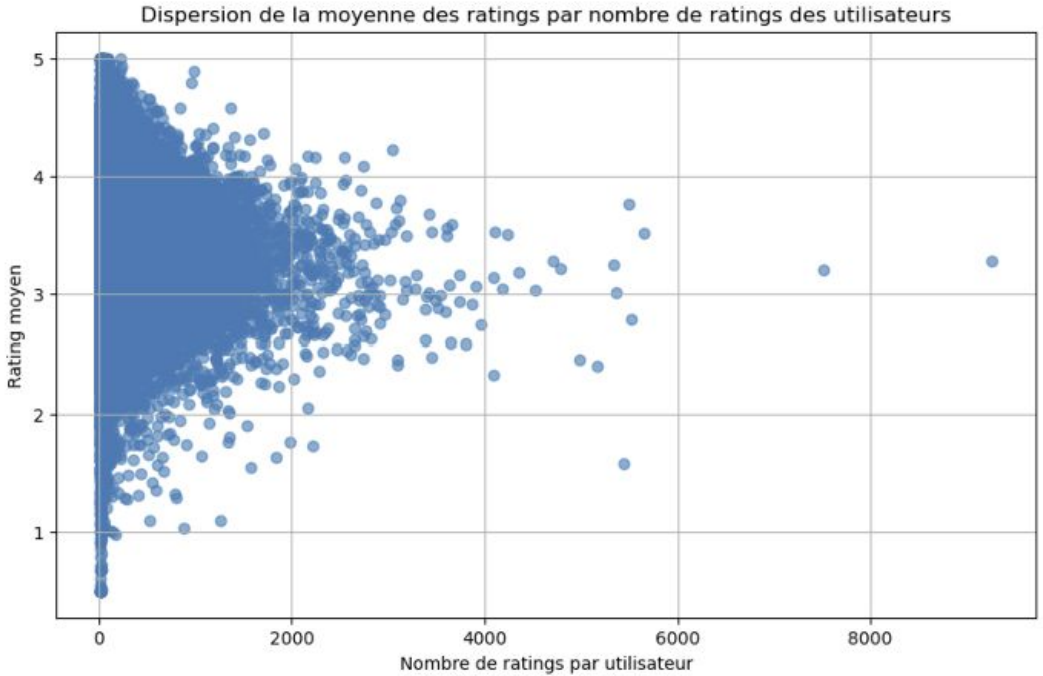


Tableau récapitulatif des réductions de dimensions possibles (en cours d'évaluation, pour nous donner des ordres de grandeur) :

Obs	1	2	3	4
Min number ratings	500	1,000	...	...
Max number ratings	700	2,000		

Matrice originale serait de :

Userid (matrix rows)	138,493	138,493		
Films (matrix columns)	26,744	26,744		

Matrice envisagée

Userid (matrix rows)	3,502	1,639		
Films (matrix columns)	16,346	21,080		

% d'information possible à garder

Userids %	2.53%	1.18%		
Films %	61.12%	78.82%		

Différence Item-based avec SVD vs Surprise avec SVD

Aspect	Modèle item-based avec SVD	Modèle Surprise avec SVD
Principe de base	Calcule les similarités entre les items après factorisation avec SVD.	Optimise directement les biais et les vecteurs latents p_u, q_i.
Biais utilisateur et item	Intégrés manuellement ou non pris en compte explicitement.	Pris en compte directement dans le modèle via b_u et b_i.
Optimisation	Pas d'optimisation explicite des paramètres latents (post-SVD).	Optimise les paramètres pour minimiser une fonction de perte.
Performance	Plus dépendant de la qualité et de la densité des données initiales.	Généralement plus robuste grâce à l'optimisation globale.
Similitudes	Repose sur les similarités item-item calculées sur V^T.	Ne calcule pas explicitement de similarités.
Implémentation	Approche manuelle nécessitant des étapes distinctes (factorisation, similarité, etc.).	Implémentation directe avec les bibliothèques comme Surprise.
Utilisation des hyperparamètres	Peut utiliser k pour les similarités (voisins proches).	Hyperparamètres comme le nombre de facteurs latents et le taux d'apprentissage.