Recommandation Musicale: Filtrage

1) Introduction

L'objectif de ce projet est de faire une système de recommandation musicale en se basant sur des données très hétérogènes liées au contexte de leur génération:

-Playlist (Spotify)
-Commentaires sur réseaux sociaux (Twitter)
-Achat de musique en ligne (Amazon, Fnac , 7digital...)

On y retrouve potentiellement les variables identifiants les utilisateur, les morceaux , les noms de titre , d'albums et d'artistes, score , senti , les caractéristiques musicales intrinsèques , genre , date …

<u>Les difficultés rencontrées sont diverses et variées</u>

-Les datasets sont un peu étranges après réflexion , plusieurs problèmes on été constatés tels que:

Les déséquilibres par rapport à leur cible: beaucoup de notes positives , très peu négatives par exemple

Les contradictions de notes: certains utilisateurs ont un comportement qui relève de la schizophrénie. Cela nécessite de traiter en amont les variable par une moyenne par exemple

Les duplicatas par rapport aux variables considérées doivent être nettoyés

-La taille des datasets et leurs traitements nécessitant beaucoup de temps et de mémoire , nous sommes obligés de tronquer considérablement les données d'entrée en prenant le risque de fausser les résultats.

-Les données des datatsets sont quelques fois difficiles à utiliser.

-valeurs non lisibles par l'utilisateur telles que identifiants de morceaux sans accès aux noms des artistes et autres variables contextuelles liées aux morceaux
-incompatibilité de format de valeurs avec les identifiants de morceaux, ce qui empêche le regroupement de données séparées sur plusieurs datasets.

Une approche mal réfléchie peut aboutir à tourner en rond sans résultats probants:
Dans une recommandation , la valeur la plus importante est l 'utilisateur ou un groupe d'utilisateurs. Il est évident que tout le monde n'a pas les même goûts, ce qui peut aboutir à quelques incohérences lors du traitement, si l'on n'en tient pas compte.

Nous allons donc utiliser quatre types d'approche:

1-Le filtrage par contenu basé sur les caractéristiques musicales, plus approprié pour une vérification subjective des résultats obtenus. En utilisant des datasets comportant des noms d'artiste , noms de chanson et caractéristiques intrinsèques musicales, on peut vérifier à l'oreille si les calculs faits s'accordent avec notre perception en écoutant les morceaux les plus proches obtenus par le modèle

2-Le filtrage collaboratif objet, approprié lorsque l'utilisateur a une interaction avec la source des morceaux: appartenance à un genre apprécié d'un utilisateur ; note attribuée like/dislike à un artiste, achat en ligne
Ce filtrage repose sur le fait que deux morceaux traités de la même manière par un utilisateur sont susceptibles d'être proches en terme de recommandations.
Si un morceau d'un utilisateur appartient à un groupe (genre , artiste) et que cet utilisateur aime ce groupe , alors il est susceptible d'aimer ce morceau.

3-Le filtrage collaboratif utilisateur, adapté pour une comparaison avec d'autres utilisateurs dont les goûts se rapprochent de l'utilisateur concerné: playlist par exemple. Les utilisateurs qui ont noté les mêmes morceaux que vous ont aussi noté d'autres morceaux tel que vous le feriez
Donc si un autre utilisateur a aimé un morceau, on peut en conclure qu'il va apprécier un autre morceau plaisant à un utilisateur similaire :on peut se baser sur la présence de morceaux en playlist ou sur une note.

4-Le filtrage collaboratif hybride qui permet l'usage de modèles annexes , entraînables , et plus performants que les filtrages user et item classiques

Ces quatre types de filtrages s'adaptent à des usages différents qui dépendent des variables à disposition.

Content Filtering: Nécessité d'avoir les **caractéristiques intrinsèques** des morceaux.
et un **identifiant** de morceau

User , Item et hybride Filtering: Nécessité d'avoir un **identifiant d'utilisateur**, de **morceau**, et une **note**

Pour vérifier nos résultats ilest important d'avoir un retour visuel sur les morceaux musicaux issus de nos prédictions: l'idéal est d'avoir pour un morceau donné , le nom de l'artiste , le titre du morceau , le genre (utile pour se faire une idée lorsque l'on ne connaît rien du titre) , le nom de l'album, etc.…

Contourner l'absence de certaines informations nous utilisons:

-Avec les données Twitter, le nom d'un hashtag peut être utilisé pour obtenir un pseudo genre: En effet nous avons à faire à des noms tels que 'hardrock', 'punk', 'romanticlove' ...qui ne laissent pas trop le doute planer quant au genre du morceau commenté.
Si des données inhérentes aux morceaux (issues de twitter par exemple) ne nous renseignent que sur un id et des données intrinsèques musicales, l 'usage du nom de hashtag est un bon moyen d'avoir un retour humainement intelligible

-Avec les données de playlist , rajouter un score de 1 peut palier à l'absence de note: une personne rationnelle ne sélectionne que des titres qu'il aime pour se constituer une playlist.

.

2) Filtrage basé sur le Contenu

Distance euclidienne des caractéristiques musicales nécessitent les valeurs

loudness, tempo, energy, danceability, accousticness ,key , liveness, mode, speechness, valence, instrumentalness.

Ces dernières peuvent être utilisées pour un calcul de distance euclidienne comme ou pourrait le faire avec les caractéristiques d'onde mfcc issus de la transformée de Fourrier calculés à partir d'un fichier wav par exemple, à l'aide de la libriarie python librosa .
(Les **MFCC** ou **Mel-Frequency Cepstral Coefficients** sont des coefficients cepstraux calculés par une transformée en cosinus discrète appliquée au spectre de puissance d'un signal)

Avec le nom de l'artiste , et du morceau une vérification à l'oreille peut être effectuée

Dataset utilisé **content_context_feature.csv , dataset.csv**

3) Filtrage collaboratif

Approche mémoire

Il nous faut pour cela les valeurs identifiant les utilisateurs et les morceaux ainsi qu'une note.
Dans les cas de datasets sans note, par exemple les playlist , on considère juste la présence d'un morceau dans la playlist de l'utilisateur.
Dans le cas d'un dataset sans nom d 'artiste et de morceau on peut contourner l 'identifiant en combinant l'id de track et son appartenance à un hashtag ou un genre pour un retour visuel plus parlant qu' un id illisible...

Item

Similarité de morceaux en passant par une matrice pivot (tracks / items / notes)

User

Similarité entre goûts d'utilisateurs en passant par une matrice pivot (tracks / items / notes)

Approche hybride et approche modèle

Usage de model de prédictions NormalPredictor et SVD

4) Préprocessing

Pour procéder aux filtrages par contenu il faut les données intrinsèques des morceaux et leurs ids
Pour procéder aux filtrages mémoire/modèle user et item il faut des identifiants d' utilisateurs , de morceaux et des notes.

Variables nécessaires:

-Lorsque les données sont issues d'une playlist il manque potentiellement une note, on contourne le problème en ajoutant un score à 1 à toutes les lignes (nécessaire pour les filtrages mémoire et modèle). C'est le cas de certains datasets spotify

-Il est parfois nécessaire de faire des jointures de datasets . Cela repose sur la cohérence des ids. C'est la cas de données twitter

Variables de présentations

-Lorsqu'il manque manque les noms d 'artistes, chansons …, on créer une variable visuelle basée sur le nom du hashtag ou de playlist (ce n'est pas nécessaire pour le calcul , mais il est important pour une vérification visuelle).
C'est la cas de certains datasets Twitter. L'idée est de présenter des informations adressées à l'oeil humain pour se faire une idée des morceaux prédits. Cette variable visuelle peut réunir pas mal d'informations si elle peuvent être obtenues dans le dataset:

*nom d'artiste
*nom de morceaux
*nom d'album
*data de sortie
*genre
*nom de playlist
*nom de hashtag
*durée

Données

-Incohérentes en terme de notes d'appréciations: On procède alors à la moyenne des notes pour le couple user/item

-déséquilibrées: présence de certaines valeurs en surabondance . On procède alors à l'élimination ou un rééquilibrage à la main du genre on garde le premier ou n au hasard.

-La standardisation des données de caractéristiques musicales est nécessaire pour le calcul content filtering basé sur une distance euclidienne .

5) Tests

Filtrage de contenu

Le datasets utilisés sont

data.csv:

parce qu'il possède les variables de caractéristiques musicales intrinsèques , ainsi que les noms d 'artiste et de chanson, ce qui facilite considérablement les vérifications subjectives.

Target:['liveness', 'speechiness','danceability', 'valence', 'loudness', 'tempo', 'acousticness', 'energy', 'mode', 'key', 'instrumentalness']
user_id: 'user_id'
item_id: 'id'
track_name: 'artists name'

final_all.csv:

Target:['liveness', 'speechiness','danceability', 'valence', 'loudness', 'tempo', 'acousticness', 'energy', 'mode', 'key', 'instrumentalness']
user_id: 'user_id'
item_id: 'track_id'
track_name: 'hashtag'

Présentation des résultats:

Résultats obtenus avec data.csv:

Le premier morceau de la liste de sortie est le morceau choisi, ce qui explique sa distance euclidienne nulle. La colonne artists contient le nom des artistes suivis des noms de chanson .Si d autre morceaux apparaissent avec cette distance null , il s agit de doublon .

********** **Test numéro1 (funk rock)**

	song_id	artists	song_name	distance
0	13822	Red Hot Chili Peppers	Give It Away	0.0
1	32863	Red Hot Chili Peppers	Give It Away	0.0
2	11318	Genesis	Follow You Follow Me - 2007 Remaster	0.0038580746667449617
3	88617	Lazlo Bane	Superman	0.004185917004751411
4	85151	Los Bukis	Más Feliz Que Tú	0.004402593358431681

5	145325	Jethro Tull	A Song for Jeffrey - 2001 Remaster	0.0044841628430626265
6	35294	Kindred The Family Soul	Stars	0.004695312927529833
7	66688	Buddy Rich	Birdland	0.004759700026543967
8	117710	P-Funk All Stars	Hydraulic Pump	0.00479978087890877
9	44571	Johnny Horton	I'm Ready If You're Willing	0.005047515402768715
10	165193	Shaquille O'Neal", 'Def Jef	(I Know I Got) Skillz (feat. Def Jef)	0.005132203101161798
11	165800	Dr. Octagon	Bear Witness	0.005141061456471332
12	35255	Zoé	Love	0.005302234992143041
13	50123	Los Bukis	Mi Najayita	0.005339800356362599
14	98896	Uriah Heep	Lady In Black	0.0054385314510413205
15	51723	Dr. Dre	Lyrical Gangbang	0.005451002423796584
16	12696	Elton John	Nikita	0.0055424895764423976
17	100218	Donna Summer	Love's Unkind	0.005604757700376372
18	165520	Lighthouse Family	Ocean Drive	0.0056808885697180785
19	152928	Lil Wayne	Kush	0.005691532415872202

Les ‘Red Hot Chilli Peppers’ peuvent être classés dans plusieurs style musicaux , rap ,rock , soul , funk… tout dépend du morceau. Avec ‘Give it Away’ on est plutôt dans le funk: On constate des morceaux de sortie assez variés, de ‘Uriah Heep’ , ‘Jethrow Tull’ à ‘Donna Summer’ en passant par ‘Elton John’. Avec plus ou moins de mauvaise foi on peut dire que ces morceaux sont proches , mais la ressemblance ‘Elton John’ et ‘Red Hot Chilli Peppers’ est un peu difficile à concevoir pour certaines oreilles….

********** **Test numéro2 (blues rcok psyché)**

	song_id	artists	song_name	distance
0	9413	Led Zeppelin	Whole Lotta Love - 1990 Remaster	0.0
1	122303	Led Zeppelin	Whole Lotta Love - Remaster	0.002260060209813591
2	9547	Led Zeppelin	Whole Lotta Love - Remaster	0.004140747916220957
3	28441	Led Zeppelin	Whole Lotta Love - Remaster	0.004140747916220957
4	70388	Enigma	The Child In Us	0.00549570191872322
5	50571	Gang Green	Alcohol	0.005815590651173384
6	165683	Cypress Hill	Throw Your Set In the Air	0.006495250687649826
7	150234	Phish	The Wedge	0.006504985980638525
8	101804	Anthrax	Gung-Ho	0.006882455269672286
9	149114	Morris Day	Fishnet	0.007086844100319224
10	14977	"Weird Al" Yankovic	Amish Paradise (Parody of "Gangsta's Paradise" by Coolio)	0.00736669342513667
11	14607	Coolio', 'L.V.	Gangsta's Paradise	0.00742980371375396
12	102543	Rod Stewart	Stay with Me	0.007635192310199364
13	31828	Boston	Cool The Engines	0.007755760463918818
14	165049	Journey	Only the Young	0.007770126690745467
15	81718	Traffic	Heaven Is In Your Mind	0.007796414040002611
16	48306	Ronnie Laws	Tidal Wave	0.007885598175201288
17	147595	Blackfoot	Baby Blue	0.007943888089967903
18	115124	Johnny Paycheck	Friend, Lover, Wife	0.008018082012647462
19	148772	The Dead Milkmen	Filet of Sole	0.008058386486523808

Nous constatons sur cet exemple que 3 des morceaux les plus proches correspondent morceau d’entrée , avec une distance euclidienne différente , ce qui est explicable par des paramètres d’encodage différents. Ce la met en évidence la capacité du modèle à reconnaître les similarités entre morceaux

********** **Test numéro3 (psyché)**

	song_id	artists	song_name	distance
0	10634	Pink Floyd	Shine On You Crazy Diamond (Pts. 1-5)	0.0
1	166538	Jesse Cook	Virtue	0.003111032161948985
2	121379	The Red Army Choir	Kalinka	0.00344857123040123
3	162710	Pink Floyd	Shine On You Crazy Diamond (Pts. 1-4)	0.0034673171930259218

4	25454	Jackie Gleason	Yesterdays	0.003944821783749464
5	33318	Michael Nyman Sarita Devi', 'Master	The Heart Asks Pleasure First / The Promise - Edit	0.004055729269058515
6	22402	Mohammad	Jhanda Uncha Rahe Hamara	0.004314130797375705
7	23565	Kaushalya	Maa Pyari Maa	0.004430683097879538
8	127645	Jimmy Yancey	At The Window	0.00460986330700917
9	95408	Stan Kenton	I Got It Bad (And That Ain't Good)	0.00476206268646879
10	141177	Kostas Roukounas	H mana sou tha ta plirosei	0.004773691903919848
11	160197	Bert Kaempfert	Mister Sandman	0.004847475053116486
12	77252	Orchestra Studio 7	Zappatore - Musical base Version	0.005056591423506526
13	151159	Spiritualized Johann Sebastian Bach', 'Isaac Stern', 'London	Broken Heart	0.005077004770696306
14	81832	Symphony Orchestra Antônio Carlos Jobim',	Violin Concerto in A Minor, BWV 1041: I. Allegro	0.005289941582465515
15	64522	'Frank Sinatra	Once I Loved (O Amor en Paz)	0.005332337625094473
16	70357	Fiona Apple	Slow Like Honey	0.005527100078173297
17	7145	Chet Baker	There Is No Greater Love	0.005813405536528247
18	73415	Random Rab	Apparently	0.005988226651169064
19	43183	Sarah Vaughan	My Reverie	0.006123385866158589

Le choix d’un morceau comme ‘Shine on crazy Diamond’ de ‘PinkFloyd’ est intéressant car il met en évidence la pluralité des morceaux les plus proches calculés par le modèle: C’est une musique qui réunit plusieurs styles

********** **Test numéro4 (trash metal)**

	song_id	artists	song_name	distance
0	12874	Slayer	Angel Of Death	0.0
1	118667	Sepultura	Slaves of Pain	0.0031643979200016758
2	151576	Misfits	The Crawling Eye	0.0034137370633807425
3	16003	Slipknot	The Heretic Anthem	0.0034870360962809873
4	121049	Slipknot	The Heretic Anthem	0.0034870360962809873
5	169777	Martin Garrix	Animals - Radio Edit	0.003915286608748716
6	70912	Misfits	Forbidden Zone	0.004105909026244939
7	104273	KISS	Psycho Circus	0.004179484644716137
8	88759	Thirty Seconds To Mars	Edge Of The Earth	0.004228269124582365
9	118863	Slayer	Expendable Youth	0.004422360024598018
10	72162	Jack's Mannequin	The Mixed Tape - 2015 Remastered	0.0045022557252041425
11	70486	Deftones	Lotion	0.004502933059816112
12	55075	August Burns Red	Mariana's Trench	0.004599068657697258
13	137699	311	Flowing	0.004643298319397586
14	73385	NERO	Doomsday	0.00468343940804959
15	151860	Slayer	God Send Death	0.004684515769892857
16	15593	Misfits	Saturday Night	0.004823945463979259
17	69215	Slayer	Expendable Youth	0.004835206297210696
18	151664	Electrasy	Cosmic Castaway - From the Film "Titan A.E."	0.004838679822564591
19	120972	...Bender	Isolate	0.004842509436050353

Le choix du morceau ‘Angel of Death’ de ‘Slayer’ permet de mettre en évidence le bon calcul du modèle concernant les musiques extrêmes, en l’occurence du trash/metal: La plupart des morceaux les plus proches correspondent à des musiques que l’on peut qualifier d’energisante.

********** **Test numéro5 (classique)**

	song_id	artists	song_name	distance
0	62775	Gioachino Rossini', 'Fritz Reiner', 'Chicago	Guillaume Tell: Overture - Remastered	0.0

		Symphony Orchestra		
1	102848	Chet Baker	Deep In A Dream	0.004668945188914143
2	129239	Andy Williams	Stranger on the Shore - Single Version	0.004766244847517284
3	24168	Johannes Brahms', 'Isaac Stern', 'Sir Thomas Beecham', 'Royal Philharmonic Orchestra	Violin Concerto in D Major, Op. 77: II. Adagio	0.004818416308974489
4	35477	Ilene Woods', 'Mike Douglas	So This Is Love - From "Cinderella"	0.004824014511783818
5	48804	Carpenters	Christ Is Born	0.00513367191680089
6	129144	Dean Martin	Let Me Know - Alternate Take	0.005280845687295824
7	40883	Wolfgang Amadeus Mozart', 'Szymon Goldberg', 'Lili Kraus	Sonate pour violon et piano No. 17 in C Major, K. 296: II. Andante sostenuto	0.005327265480364045
8	40708	Umm Kulthum	Nasheid El Gamea	0.005347396157527674
9	93259	Ludwig van Beethoven', 'NBC Symphony Orchestra', 'Arturo Toscanini	Symphony No. 1 in C Major, Op. 21: II. Andante cantabile con moto	0.0053679066663567645
10	123932	Bon Iver	29 #Strafford APTS	0.005376263905229226
11	58557	Johannes Brahms', 'Jascha Heifetz', 'Arturo Toscanini', 'New York Philharmonic	Concerto for Violin in D Major, Op. 77: II. Adagio - Live	0.005451151931217302
12	112284	Richard Rodgers', 'Julie Andrews', 'Alfredo Antonini	Cinderella - Original Broadway Cast: In My Own Little Corner (Reprise)	0.005700215440387374
13	78762	Johannes Brahms', 'Isaac Stern', 'Sir Thomas Beecham', 'Royal Philharmonic Orchestra	Violin Concerto in D Major, Op. 77: II. Adagio	0.005764998467090089
14	63513	Ludwig van Beethoven', 'George Szell', 'Cleveland Orchestra	Symphony No. 6 in F Major, Op. 68 "Pastoral": I. Erwachen heiterer Empfindungen bei der Ankunft auf dem Lande. Allegro ma non troppo	0.005785875242880433
15	46618	Elvis Presley	Where Could I Go But to the Lord	0.005825007842693796
16	62716	Frank Sinatra', 'Larry Walsh	Willow Weep For Me - 2018 Stereo Mix	0.005841262622373029
17	125662	Claude Debussy', 'Fritz Reiner', 'Pittsburgh Symphony Orchestra	Danse "Tarantelle styrienne", L. 69 - Remastered	0.0060744236661826176
18	160097	Jackie Gleason	Let It Snow! Let It Snow! Let It Snow!	0.006079287802795744
19	156187	Giacomo Puccini', 'Kiri Te Kanawa', 'Sir John Pritchard', 'London Philharmonic Orchestra	La Rondine: Chi il bel sogno di Doretta	0.0060943027922148495

Avec le choix de l’ouverture de guillaume Tell’ de ‘Rossini’, nous voyons des résultats avec un style plutôt classique . Le modèle semble fonctionner pour des musiques marquées.

**********Test numéro6 (classique)

	song_id	artists	song_name	distance
0	7718	Pyotr Ilyich Tchaikovsky', 'Leonard Bernstein', 'New York Philharmonic	The Nutcracker Suite, Op. 71a, TH 35: IIb. Danses caractéristiques. Danse de la fée-dragée	0.0
1	63216	Pyotr Ilyich Tchaikovsky', 'Leonard Bernstein', 'New York Philharmonic	Nutcracker Suite, Op. 71a: Danse Russe - Trépak (Russian Dance). Tempo di trepak, molto vivace	0.001194022110637909
2	117434	George Frideric Handel', 'English Baroque Soloists', 'John Eliot Gardiner	Messiah, HWV 56 / Pt. 1: Symphony	0.003972651973981233
3	97448	Johann Sebastian Bach', 'Glenn Gould	Prelude in E-Flat Minor & Fugue in D-Sharp Minor No. 8, BWV 853: Prelude	0.005628070068567267
4	125669	Franz Liszt', 'Vladimir Horowitz	Valse oubliée in F-Sharp Major, S. 215/1	0.0058829816396174395
5	83868	Pat Metheny	Oasis	0.005926516364288213

6	148250	Johann Sebastian Bach', 'Yo-Yo Ma', 'Ton Koopman', 'Amsterdam Baroque Orchestra	Ich ruf' zu dir, Herr Jesu Christ, BWV 639	0.006082502953388261
7	88038	Felix Mendelssohn', 'Murray Perahia	Lieder ohne Worte, Op. 19, No. 1	0.008281191549774882
8	60310	Frédéric Chopin', 'Arthur Rubinstein	24 Preludes, Op. 28: Prelude No. 2 in A minor	0.008372815468212448
9	43059	Maurice Ravel', 'Philippe Entremont', 'Régine Crespin	Histoires naturelles, M. 50: III. Le Cygne	0.008664147709188203
10	153476	Nicholas Hooper	Ginny	0.008681119149067185
11	41402	Richard Wagner', 'Arturo Toscanini	Act III: Siegfried's Death and Funeral Music	0.008688838004144118
12	81842	Ennio Morricone	Marcia Without Hope	0.008960356013042866
13	85229	Maurice Ravel', 'Ivo Pogorelich	Gaspard de la nuit, M. 55: 1. Ondine	0.008973289136369682
14	5516	Mel Tormé	The Christmas Song (Chestnuts Roasting on an Open Fire)	0.009090599353129074
15	134564	Philip Glass	Metamorphosis: One	0.00912758897687392
16	4651	Doris Day', 'Harry James	Too Marvelous For Words	0.009194326228794505
17	156944	Wolfgang Amadeus Mozart', 'Eugene Ormandy', 'Philadelphia Orchestra	Piano Concerto No. 20 in D Minor, K.466: I. Allegro	0.009261717587322376
18	111029	Sergei Prokofiev', 'Jascha Heifetz', 'Serge Koussevitzky	Violin Concerto No. 2, Op. 63 in G Minor: Allegro moderato	0.009374856875005727
19	42884	Ludwig van Beethoven', 'Jascha Heifetz	Sonata No. 6, Op. 30, No. 1 in A: Variation V	0.009507198648155805

Un autre test de musique classique qui semble fonctionner

**********Test numéro6 (jazz)**

	song_id	artists	song_name	distance
0	8632	Nina Simone	I Put A Spell On You	0.0
1	160014	Carmela Y Rafael', 'Rondalla Mexicana Del Chato Franco	Caminemos	0.0031697879757019752
2	2356	Billie Holiday	What Is This Going to Get Us?	0.0033831626224165214
3	22537	Billie Holiday	What Is This Going to Get Us?	0.0033831626224165214
4	115272	Elvis Presley	I, John	0.0035759554858520314
5	25238	Charlie Parker	Almost Like Being In Love	0.004060217770450573
6	60918	Lata Mangeshkar	Jo Dil Mein Khushi Ban Kar Aai	0.00417976191165332
7	81260	Javier Solís	Sombras	0.004215217977191121
8	41621	Prem Adib	Man Aisa Geet Na Gana	0.0042517237901593145
9	42657	Δημήτρης Γκόγκος	Όμορφη Σμυρνιά	0.004268300386552473
10	40321	Fréhel	Où Sont Tous Mes Amants	0.004276435235043861
11	6941	Miles Davis Quartet	Tune-Up	0.004323115457424004
12	80727	Mohammed Rafi	Aye Gulbadan	0.004364623961255546
13	143534	June Christy	I Want To Be Happy	0.0045095162042543616
14	5440	Burl Ives	Mr. Froggie Went A-Courtin'	0.004645148854273247
15	62152	Lord Invader	Barbados	0.0047038115608047425
16	95636	Mohammed Rafi	Mujhe Apni Sharan Men Lelo Ram	0.004735146450105675
17	65694	Grateful Dead	Mississippi Half-Step Uptown Toodeloo - 2013 Remaster	0.004748537043286343
18	159704	謝雷	啊咿奧	0.004844439909659385
19	45514	Talat Mahmood', 'Lata Mangeshkar	Itna Na Mujhse Tu Pyar Badha - Duet	0.004868636584972355

Le style Jazz avec ‘I Put a Spell on You’ de ‘Nina Simone‘ semble avoir de bon résultats de similarité

Remarque globale
On reste en générale dans le style . La liste de sortie semble mieux fonctionner avec les styles marqués, tel que la musique classique ou le métal. Ce qui semble évident , lorsque les styles sont mélangés dans un morceau il est peu probable que la distance euclidienne des caractéristiques musicales puissent sortir un résultat s'accordant avec la perception de l'utilisateur.

L approche basée sur le contenu semble donc donner d'assez bons résultats avec des musiques marquées.
Nous avons choisi quelques morceaux connus pour certains , et marqués pour d'autres afin de se faire idée.
Bien entendu ce test s 'adresse à l appréciation subjective de l 'utilisateur.

Filtrage mémoire item et user (collaboratif objet et user)

Les variables importantes sont les identifiants de morceaux et d'utilisateurs ainsi que leur sentiment. Mais il faut distinguer la provenance des informations: une liste de notes d'un utilisateur ou sa playlist n'ont pas la même signification . La note donnée renseigne directement l appréciation positive ou négative d'un morceau par un utilisateur , alors que la présence d'un morceau dans la playlist d'un utilisateur ne renseigne que le fait d'une appréciation positive, en effet on n'imagine pas un quelqu'un créant un playlist avec des morceaux qu'il déteste…

L'idée du test est de repérer un utilisateur aléatoire puis de calculer à l'aide d'un matrice pivot , son utilisateur le plus proche. Ensuite on compare leurs préférences , prédictions item et user qui devraient être proches....

Cela dit la prédiction item sans note attribuée pourrait être un peu bancale, puisque limitée à 0 ou 1.

Présentation des résultats:

La qualité des datsets est très discutable , des doublons de variables sont cachés , il faut passer beaucoup de temps à les comprendre , les détecter , les contourner. De plus les résultats ne sont pas faciles à analyser.
C'est pour cela que nous allons travailler sur un dataset de simulation , cela nous apportera l'avantage d'une rapidité d'exécution, une limitation d'usage mémoire , une lisibilité accrue pour les vérifications.
Nous passerons ensuite à des datasets issus de spotify et twitter , avec leurs défauts et difficultés d'interprétation des résultats

Test 1 dataset de simulation

Nous allons donc commencer par un data frame de simulation comportant 5 utilisateurs 'user0' , ...'user4' et 10 morceaux 'track0' … 'track9', remplis aléatoirement par des notes allant de 0 à 10 :

user_id	track0	track1	track2	track3	track4	track5	track6	track7	track8	track9
user0				3.0				4.0		2.0
user1					4.0		5.0	0.0		
user2	5.0	1.0	0.0		0.0		1.0	0.0		1.0
user3					5.0					
user4					0.0		1.0			

On choisi un l'utilisateur on affiche ses préférences , prédictions item et user:
L'utilisateur choisi est user0:

Ses Préférences avec un seuil de 0:

	track_id	sentiment_score
5	track7	4.0

9	track3	3.0
21	track9	2.0

Sa prédiction user:

track_id	score moyen
track0	4.248522721414388
track1	1.4161742404714628
track2	0.7080871202357314
track4	2.1676515190570744
track6	3.167651519057074

Sa prédiction Item:

track_id	score moyen
track4	0.7930207033741828
track0	0.6513878188659973
track1	0.6513878188659973
track2	0.6513878188659973
track6	0.0

Les tracks de préférences n'appaissent pas dans les prédictions
Les tracks 8 et 5 n apparaissent ni dans les préférences ni dans les prédictions , puisqu'ils n'ont jamais été notés

Le deuxième utilisateur choisi est user2:

Ses Préférences avec un seuil de 0:

	track_id	sentiment_score
8	track0	5.0
16	track6	1.0
24	track9	1.0
35	track1	1.0
1	track4	0.0
19	track7	0.0
45	track2	0.0

Sa prédiction user:

track_id	score moyen
track3	1.0238304570537966

Sa prédiction item:

track_id	score moyen
track3	1.4637184137552184

Les tracks de prédictions n'apparaissent pas dans la liste des préférences, c'est bon signe
Les tracks 8 et 5 n apparaissent ni dans les préférences ni dans les prédictions , puisqu'ils n'ont jamais été notés
Seul le track 3 est dans les prédictions , c'est en effet le seul qui reçu une note mais qu'i n'a pas été noté par le user2

La similarité entre les deux utilisateurs est 0.21783245945486832

Test 2 spotify_user_track_reduced10.01.csv:

est une playlist spotify dont les variables nom d 'artiste et de chanson sont présentes , ainsi que les identifiants des utilisateurs. Un score global est présent , mais nous considérerons que le simple fait qu'un utilisateur possède le morceau dans sa playlist implique qu'il l'apprécie. Nous oublierons la popularité qui ne reflète qu'un appréciation globale.

target: présence dans la playlist ou popularité
user_id: ‘user_id’
item_id: ‘id’
track_name: ’artists name’

Présentation des tableaux:

La colonne ‘name’ indique ’Nom d’artiste’:‘titre de morceau’

on choisi un l’utilisateur le plus présent et on fait des prédictions user pour celui ci , ensuite on prend on utilisateur le plus proche pour effectuer une prédiction item et user:

L’utilisateur choisi au hasard est 4398de6902abde3351347b048fcdc287
Préférence: Nous les avons mis tous les scores à 1 , la variable de la note étant absente : tous les morceaux de sa playlist ont le même poids

Sa prédiction user:

	name	sentiment_score
0	Regina Spektor:Regina Spektor : Lacrimosa	1
1	SI SEÑOR !!:Los Miticos Del Ritmo : Willy's Merengue	1
2	Shooter Jennings – Black Ribbons:Shooter Jennings : Fuck You (I'm Famous)	1
3	Fancy Dancy:Tiësto feat. Nelly Furtado : Who Wants To Be Alone - Phillip D Remix	1
4	16/05/13_31/12/13:Red Fang : Blood Like Cream	1
5	2014 Jan Touring:Pure Bathing Culture : Ever Greener	1
6	Upbeat:Sara Bareilles : Uncharted	1
7	July2013:Shalamar : I Can Make You Feel Good	1
8	Indie:Florence + The Machine : Dog Days Are Over	1
9	Barrett Strong - Hitsville USA - The Motown Singles Collection 1959-1971:Marvin Gaye : Mercy Mercy Me (The Ecology)	1

Sa prédiction Item:

	name	sentiment_score
0	Regina Spektor:Regina Spektor : Lacrimosa	1
1	SI SEÑOR !!:Los Miticos Del Ritmo : Willy's Merengue	1
2	Shooter Jennings – Black Ribbons:Shooter Jennings : Fuck You (I'm Famous)	1
3	Fancy Dancy:Tiësto feat. Nelly Furtado : Who Wants To Be Alone - Phillip D Remix	1
4	16/05/13_31/12/13:Red Fang : Blood Like Cream	1
5	2014 Jan Touring:Pure Bathing Culture : Ever Greener	1
6	Upbeat:Sara Bareilles : Uncharted	1
7	July2013:Shalamar : I Can Make You Feel Good	1
8	Indie:Florence + The Machine : Dog Days Are Over	1
9	Barrett Strong - Hitsville USA - The Motown Singles Collection 1959-1971:Marvin Gaye : Mercy Mercy Me (The Ecology)	1

L’utilisateur le plus proche est 0007f3dd09c91198371454c608d47f22

Ses 10 préferences sont:

	name	sentiment_score
352931	Fav songs:Dead by April : Beautiful Nightmare	1

Une seule valeur car l’utisateur n’a qu’une seule ligne dans le dataset (problème lié aux réduction de lignes)

Sa prédiction user:

	name	sentiment_score
0	Regina Spektor:Regina Spektor : Lacrimosa	1

1	SI SEÑOR !!:Los Miticos Del Ritmo : Willy's Merengue	1
2	Shooter Jennings – Black Ribbons:Shooter Jennings : Fuck You (I'm Famous)	1
3	Starred:Le Castle Vania : Nobody Gets Out Alive Part II	1
4	Fancy Dancy:Tiësto feat. Nelly Furtado : Who Wants To Be Alone - Phillip D Remix	1
5	16/05/13_31/12/13:Red Fang : Blood Like Cream	1
6	2014 Jan Touring:Pure Bathing Culture : Ever Greener	1
7	Upbeat:Sara Bareilles : Uncharted	1
8	July2013:Shalamar : I Can Make You Feel Good	1
9	Indie:Florence + The Machine : Dog Days Are Over	1

Est quasiment la même que celle de du user1

Sa prédiction Item:

	name	sentiment_score
0	Starred:Duncan Dhu : En Algun Lugar	1
1	Starred:Dropkick Murphys : Fields of Athenry	1
2	Starred:Dzihan & Kamien : Homebase	1
3	Starred:Dyland & Lenny : Nadie Te Amará Como Yo	1
4	Starred:Duran Duran : Ordinary World	1
5	Starred:Dubstep : Dubstep 6 - Dubstep Mix	1
6	Starred:Dropkick Murphys : I'm Shipping Up To Boston	1
7	Starred:Earl Sweatshirt : Whoa	1
8	Starred:Duke Dumont : Need U (100%) - Radio Edit	1
9	joni mitchell :Prince : A Case of You	1

Est complètement différente de celle du user1 : Cet utilisateur n'a qu'une seule ligne dans le dataset.

on garde l'utilisateur le plus présent et l'on fait des prédictions item et user avec le second utilisateur le plus présent
Le deuxième utilisateur le plus présent dans le datset est 99deafd9b792af8e6a535483088faef2

Ses 10 préferences sont:

	name	sentiment_score
1168318	All Spotify Stuff:Useless ID : Erratic	1
534133	Latin / Flamenco / Spanish:Juanes : Para Tu Amor	1
159119	All Spotify Stuff:Dwarves : Bang Up	1
182852	All Spotify Stuff Vol 3:Mord Fustang : Super Meat Freeze - Original Mix	1
699381	All Spotify Stuff Vol 2:October : Born In A Rolling Barrel	1
17670	All Spotify Stuff Vol 3:Sam Roberts Band : Shapeshifters - Youth Banda Remix	1
345895	All Spotify Stuff:Gil Evans : Wait Till You See Her - master	1
1127076	All Spotify Stuff Vol 2:Rusko : Yeah	1
1226178	All Spotify Stuff Vol 3:Farruko : Miro El Reloj	1
1209993	July 4 2014 Daytime:Best Coast : I Want To	1

Sa prédiction User:

	name	sentiment_score
0	Regina Spektor:Regina Spektor : Lacrimosa	1
1	SI SEÑOR !!:Los Miticos Del Ritmo : Willy's Merengue	1
2	Shooter Jennings – Black Ribbons:Shooter Jennings : Fuck You (I'm Famous)	1
3	Starred:Le Castle Vania : Nobody Gets Out Alive Part II	1
4	Fancy Dancy:Tiësto feat. Nelly Furtado : Who Wants To Be Alone - Phillip D Remix	1
5	16/05/13_31/12/13:Red Fang : Blood Like Cream	1
6	2014 Jan Touring:Pure Bathing Culture : Ever Greener	1
7	Upbeat:Sara Bareilles : Uncharted	1

8	July2013:Shalamar : I Can Make You Feel Good	1
9	Indie:Florence + The Machine : Dog Days Are Over	1

Encore un fois les prédiction user sont très proches

Sa prédiction Item:

	name	sentiment_score
0	Starred:Eminem : Encore/Curtains Down	1
1	Starred:Emmy The Great : We Almost Had A Baby	1
2	Starred:Emily Maguire : Over the Waterfall	1
3	Starred:Eminem : Criminal - Album Version (Edited)	1
4	Starred:Elton John : I Don't Wanna Go On With You Like That	1
5	Starred:Emmy The Great : Paper Forest - In The Afterglow of Rapture	1
6	Starred:English Concert Choir/Trevor Pinnock : Behold, and see if there be any sorrow (tenor, arioso)	1
7	Starred:Eminem : Cinderella Man	1
8	joni mitchell :Prince : A Case of You	1
9	Starred:Eminem : Tonya - Skit (Explicit)	1

La prédiction Item est très différente

Remarque

Les prédictions items entre les deux utilisateurs sont proches.

Le fait que les préférences limitées à 1 de l'utilisateur le plus proche , signifie qu'il n'est présent qu une seule fois dans le data set (cela est du aux réductions drastiques de lignes effectuées en amont).
Encore une fois les datasets sont tellement déséquilibrés, les réductions de taille dues aux manques de performance de nos machines nous amène à des résultats très difficiles à interpréter

Test 3 user_and_track_sentiment.csv:

Est un merge de deux datasets de twitter: Il réunit les variables user_id , track_id , hashtag et sentiment_score . N'ayant pas les noms de chansons et d'artiste , nous pouvons utiliser le nom du hashtag pour palier à ce manque. Il arrive que les hashtag soient parlant . Dans ce dataset nous avons un score d'un ensemble d'utilisateurs que nous considérerons comme la note de l'utilisateur

target: 'sentiment'
user_id: 'user_id'
item_id: 'track_id'
track_name: 'hashtag'

Nous avons remarqué que le hashtag 'nowplaying' est écrasant en majorité: 5765350 sur 6053259: la solution est de le supprimer tout simplement .

De plus nom n'est pas très parlant par rapport à d'autres tels que deathrock, punk, romanticlove…

L'utilisateur choisi au hasard est 106628331

Ses 10 préferences sont:

	track_id	hashtag	sentiment_score
2142205	b2b84348b8ce3a5348092fff46dd1203	slowmotion	0.0
2699354	b58b2b2070a794a7fbcc949c9257b7cc	slowmotion	0.0
2560076	e01d28c988407f58db4da765e1e34bab	slowmotion	0.0
2698669	e87ef5136d31a8fd9c037007544234c4	slowmotion	0.0
2698793	10591cc205c699a574d181ea06c329fb	slowmotion	0.0
2698947	7767886f58ba855383ac6a629f11c5fd	slowmotion	0.0
2699024	0d6cabb2a84b70336f698df3851d9445	slowmotion	0.0

2699163	998719918bc7751404b274d64e60fb9e	slowmotion	0.0
2840082	d6b1124fdd64c4b1afa59f967397111d	slowmotion	0.0
2559868	e5cca8172905a7d0f4f7fd71b69e502a	slowmotion	0.0

L’utilisateur a beaucoup participé à un hashtag

Sa prédiction user:

	track_id	hashtag
0	0000e47c1207e2c637a44753a713456f	rock
1	000248c97c5991b9900360aca97d9879	disturbed
2	000359abe73f500143335921d078a7b0	kiss92
3	000359abe73f500143335921d078a7b0	boringsville
4	000359abe73f500143335921d078a7b0	unstunst
5	000359abe73f500143335921d078a7b0	twittamp
6	000359abe73f500143335921d078a7b0	justloved
7	000359abe73f500143335921d078a7b0	tophits
8	000359abe73f500143335921d078a7b0	tophits
9	000359abe73f500143335921d078a7b0	redhotchilipeppers

On constate que le même morceau revient souvent , c’est un problème lié au dataset : des morceaux font partis de commentaires issus de hasttag différents

Sa prédiction Item:

	track_id	hashtag
0	7d63a98b631fbb193b577b7081f766e2	starhustler
1	7d63a98b631fbb193b577b7081f766e2	great
2	459462ccf0a47e68cc5e115ca67e2658	chill
3	d47175c6b461bf16604af0901c4be256	fighting
4	39144013a2e9891fc3c7c3dc02a0ec02	twittamp
5	39144013a2e9891fc3c7c3dc02a0ec02	afterhourplaylist
6	39144013a2e9891fc3c7c3dc02a0ec02	playlistmalam
7	39144013a2e9891fc3c7c3dc02a0ec02	twittamp
8	39144013a2e9891fc3c7c3dc02a0ec02	twittamp
9	39144013a2e9891fc3c7c3dc02a0ec02	twittamp

Les résultats ne parlent beaucoup , les hashtag ne sont pas forcément un bon choix d’identification visuelle, mais ce sont les seuls à disposition

L’utilisateur le plus proche est 81496937

Ses 10 préferences sont:

	track_id	hashtag	sentiment_score
1	cd52b3e5b51da29e5893dba82a418a4b	deathrock	1.3
3240199	8c7312077c77616d89dd7db43102c251	deathrock	1.3
3234120	7629287dcd73bd997153355fbf607092	deathrock	1.3
3242793	2d41ff3b2153e948bedcba02a291ebca	deathrock	1.3
3242632	71eeb0ca05fa2399ef7e2ec051e94429	deathrock	1.3
3242505	9d01cb9daebba8f53275b49536c0c060	deathrock	1.3
3242210	ec1a6bcb18aebc4fbba1affc9e568a78	deathrock	1.3
3242054	5f2c00a4e0cd395d08f30842e79ba605	deathrock	1.3

3241518	aaf13d347e55b3e997b3572cfaa60fbb	deathrock	1.3
3241065	49938fce6609e5cbd7df1eaaec7e4e41	deathrock	1.3

De part le nom du hashtag on change de style

Sa prédiction user:

	track_id	hashtag
0	0000e47c1207e2c637a44753a713456f	rock
1	000248c97c5991b9900360aca97d9879	disturbed
2	000359abe73f500143335921d078a7b0	kiss92
3	000359abe73f500143335921d078a7b0	boringsville
4	000359abe73f500143335921d078a7b0	unstunst
5	000359abe73f500143335921d078a7b0	twittamp
6	000359abe73f500143335921d078a7b0	justloved
7	000359abe73f500143335921d078a7b0	tophits
8	000359abe73f500143335921d078a7b0	tophits
9	000359abe73f500143335921d078a7b0	redhotchilipeppers

Certains hashtag semblent se rapprocher de ‘deathrock’, mais on constate à nouveau le même morceau remonté

Sa prédiction Item :

	track_id	hashtag
0	9f019258072bd0ca3dc4864af5333216	wreckonthehighway
1	1afc6d55a1f9685cce9208b094cae61a	please
2	31892ef6457495cd6054fa25d545fa18	postpunk
3	31892ef6457495cd6054fa25d545fa18	darkwave
4	31892ef6457495cd6054fa25d545fa18	postpunk
5	31892ef6457495cd6054fa25d545fa18	darkwave
6	d1bd3af257ffd466f1b9038bff27fb9e	brilliant
7	835a634e6e4d6e51706868ca6aebc842	gunaydinsarkisi
8	835a634e6e4d6e51706868ca6aebc842	gunaydinsarkisi
9	f3208446b05ee02129691c88c194442a	deathrock

Un morceau revient encore plusieurs fois … Les noms de hashtag se rapprochent avec le hashtag de préferences ‘deathrock’

Les résultats de ce test sont peu concluants , la lisibilité des résultats ainsi que la qualité des datasets font de la vérification une tâche plutôt ardue.

Filtrage hybride

Présentation des résultats:

Test 1 dataset de simulation

***L’utilisateur choisi au hasard est* user3**

Prediction faite avec le modèle Normal Predictor

	user_id	track_id	r_ui	est	details
0	user3	track2	6.38	5	{'was_impossible': False}
1	user3	track3	6.38	5	{'was_impossible': False}
2	user3	track1	6.38	5	{'was_impossible': False}
3	user3	track6	6.38	5	{'was_impossible': False}
4	user3	track7	6.38	5	{'was_impossible': False}

5	user3	track4	6.38	5	{'was_impossible': False}

Prediction faite avec le modèle SVD (Singular ValueDecomposition)

	user_id	track_id	r_ui	est	details
0	user3	track2	6.38	5	{'was_impossible': False}
1	user3	track3	6.38	5	{'was_impossible': False}
2	user3	track1	6.38	5	{'was_impossible': False}
3	user3	track6	6.38	5	{'was_impossible': False}
4	user3	track7	6.38	5	{'was_impossible': False}
5	user3	track4	6.38	5	{'was_impossible': False}

Les deux prédictions sont identiques

Test 2 spotify_user_track_reduced10.01.csv:

L'utilisateur choisi est 4398de6902abde3351347b048fcdc287

Prédictions faites avec le modèle Normal Predictor

	name
0	Kavinsky : Nightcall - Breakbot Remix
1	No way out : Lo mismo
2	Fat Freddy's Drop : Silver and Gold
3	Bear McCreary : Dark Woods
4	Manic Street Preachers : We Her Majesty's Prisoners
5	Gary Allan : All I Had Going Is Gone
6	Harry Nilsson : Girlfriend
7	BIGBANG : Bingle Bingle (빙글빙글)
8	Toy : Make It Mine
9	Sam Smith : Stay With Me

Prédictions faites avec le modèle SVD (Singular ValueDecomposition)

Evaluating RMSE, MAE of algorithm SVD on 5 split(s).

	Fold 1	*Fold 2*	*Fold 3*	*Fold 4*	*Fold 5*	*Mean*	*Std*
RMSE (testset)	*0.0523*	*0.0520*	*0.0522*	*0.0522*	*0.0522*	*0.0522*	*0.0001*
MAE (testset)	*0.0314*	*0.0312*	*0.0313*	*0.0314*	*0.0314*	*0.0314*	*0.0001*
Fit time	*42.34*	*41.78*	*41.40*	*45.06*	*42.79*	*42.67*	*1.28*
Test time	*7.64*	*5.89*	*7.31*	*6.00*	*6.92*	*6.75*	*0.70*

	name
0	Kavinsky : Nightcall - Breakbot Remix
1	No way out : Lo mismo
2	Fat Freddy's Drop : Silver and Gold
3	Bear McCreary : Dark Woods
4	Manic Street Preachers : We Her Majesty's Prisoners
5	Gary Allan : All I Had Going Is Gone
6	Harry Nilsson : Girlfriend
7	BIGBANG : Bingle Bingle (빙글빙글)
8	Toy : Make It Mine

9 Sam Smith : Stay With Me

Les deux prédictions sont identiques

Test 3 user_and_track_sentiment.csv:
Prédictions faite avec le modèle Normal Predictor
***L'utilisateur est* 106628331**

	track_id	hashtag
0	4e8390d4fb61ee55b9903562a4cd4a6c	goodmemories
1	1cc13a8f4d742d7e7895efd3728aefd0	realclassicrock
2	6f6850d459c7d54b88e666a29780adc9	rock
3	f126acbebc58dbdf530300708dc69ca2	doommetal
4	f6d3f84ac1f57ef542df6978cfdd2dcf	doommetal
5	544831cb8e4ae3b5c8e614e762888305	greatrequest
6	75bc08b82262cbcd68a0f33e8766924a	goodmemories
7	6be564fc9c82bf26e3f681b8af5fa90c	kiss92
8	d6b1124fdd64c4b1afa59f967397111d	kiss92
9	d6b1124fdd64c4b1afa59f967397111d	kiss92

Prédictions faite avec le modèle SVD (Singular ValueDecomposition)

	track_id	hashtag
0	4e8390d4fb61ee55b9903562a4cd4a6c	goodmemories
1	1cc13a8f4d742d7e7895efd3728aefd0	realclassicrock
2	6f6850d459c7d54b88e666a29780adc9	rock
3	f126acbebc58dbdf530300708dc69ca2	doommetal
4	f6d3f84ac1f57ef542df6978cfdd2dcf	doommetal
5	544831cb8e4ae3b5c8e614e762888305	greatrequest
6	75bc08b82262cbcd68a0f33e8766924a	goodmemories
7	6be564fc9c82bf26e3f681b8af5fa90c	kiss92
8	d6b1124fdd64c4b1afa59f967397111d	kiss92
9	d6b1124fdd64c4b1afa59f967397111d	kiss92

Les deux prédictions sont identiques

Les modèles hybrides sont beaucoup plus rapide que notre implémentation mémoire user et item

6) Conclusion

Pour valider nos modèle il est impératif d'avoir des résultats interprétables, mais certains sont peu parlants .

Nous avons appris qu'il ne faut pas négliger certains aspects quant la qualité des données d'entrées:

-Eviter les doublons

-Eviter les déséquilibres de valeurs: trop de tracks identiques, d'utilisateurs identiques

-Posséder un retour visuel fiable pour une interprétation humaine possible

-Avoir les bonnes variables entre les mains: Il manque souvent aux datatsets utilisés des variables que l'on peut contourner avec le risque de fausser les résultats .

-Eviter les données trop conséquentes: on peut rapidement à attendre des heures avant qu'un calcul se finisse et lorsque les résultats ne sont pas concluants , on peut ressentir une forme de frustration ...

Pour l'instant nous avons encore du travail de debuggage et de fabrication de datatsets moins pollués

Le content filtering est cependant amusant

Notre objectif est de partir de datasets issus de sources variées , et de tester tous les modèles dans les même conditions afin d'obtenir une comparaison objective des résultats obtenus…

Pour cela nous allons passer par des datasets de simulations , en générant des utlisateurs aux gouts différents codés dans leurs ids , et des tracks de genre différents codés dans leurs ids, les notes attribuées devront correspondre aux goûts . Cela nous permettra de vérifier rapidement la qualité des prédictions , en connaissant à l'avance ce que doit prédire le modèle.