UNIVERSITÉ DE SHERBROOKE

Faculté des sciences
Département d'informatique

IFT799 – Science des données

Rapport de travail individuel

TP4 : Prédiction basée sur un principe de collaboration; évaluation hors-ligne.

JEU DE DONNÉES ABR

RÉALISÉ PAR:

KAOUTAR L'HASSNAOUI 22 148 702

ENCADRÉ PAR: SHENGRUI WANG
ETIENNE GAEL TAJEUNA

Abstract

Sur Amazon, les clients peuvent laisser leurs opinions sur des livres. Ce feedback s'avère être une source d'information avec beaucoup de potentiel. En effet, comparé au marketing de masse, proposer une expérience personnalisée peut se trouver très utile pour augmenter les ventes.

A partir des reviews effectués par un internaute sur des livres donnés, on veut savoir quel serait l'opinion de cet internaute sur un autre livre. Ainsi, on devrait réaliser une prédiction basée sur les méthodes de système de recommandation et des algos de clustering.

Ce TP4 a pour but de meilleure expérience aux lecteurs de notre base de données en se basant sur leurs ratings des livres déjà lus et sur les lecteurs qui leurs sont similaires .

Table de matières

1.Introduction

2.Problème de Dataset

3.Méthodologie proposée dans le TP

4.Implémentation et analyse

5.Conclusion

6.Ressources

1.Introduction

Maintenant qu'on a une idée générale de nos données et qu'on a obtenu des structures pertinentes, on veut pousser notre travail pour pouvoir faire la prédiction des préferences des lecteurs et donc leur recommander des livres intéressents.

Le TP1 nous a permis de comprendre nos données de façon globale grâce aux outils statistiques et à des visualisations après une projection en composantes principales.

Dans le TP2_3, on a utilisé différentes approches de clustering tout en visualisant et en évaluant les segmentations obtenues pour enfin sortir avec celle la plus représentative de nos données.

Pour ce TP4, le but est de faire une prédiction basée sur un principe de collaboration et puis accomplir une évaluation hors-ligne ceci sera utile pour notre analyse prédictive.

2.Dataset

Comme pour le TP1 je travaille avec les 50000 premières lignes de la base données 'ABR.json'. Notons que l'on garde cet aspect aléatoire puisque la base de donnée initiale ne présente aucun ordre sur ces lignes.

Elle est composée de 9 colonnes dont la description est la suivante :

	reviewerID	asin	reviewerName	helpful	reviewText	overall	summary	unixReviewTime	reviewTime
0	A10000012B7CGYKOMPQ4L	000100039X	Adam	[0, 0]	Spiritually and mentally inspiring! A book tha...	5	Wonderful!	1355616000	12 16, 2012
1	A2S166WSCFIFP5	000100039X	adead_poet@hotmail.com "adead_poet@hotmail.com"	[0, 2]	This is one my must have books. It is a master...	5	close to god	1071100800	12 11, 2003
2	A1BM81XB4QHOA3	000100039X	Ahoro Blethends "Seriously"	[0, 0]	This book provides a reflection that you can a...	5	Must Read for Life Afficianados	1390003200	01 18, 2014
3	A1MOSTXNIO5MPJ	000100039X	Alan Krug	[0, 0]	I first read THE PROPHET in college back in th...	5	Timeless for every good and bad time in your l...	1317081600	09 27, 2011
4	A2XQ5LZHTD4AFT	000100039X	Alaturka	[7, 9]	A timeless classic. It is a very demanding an...	5	A Modern Rumi	1033948800	10 7, 2002
...	...	...	...	...	...	...	...	...	...
49995	A2R2VN5X77D66O	0028633873	Donna H.	[0, 0]	i gave this as a gift so i don't know how good...	5	bought as a gift	1402704000	06 14, 2014
49996	A38UCPTYS6LBHE	0028633873	Lindsay Harrison "film and book aficionado"	[22, 23]	I have been attempting to learn Hebrew, but I ...	5	For the Jew or the Goy	1051488000	04 28, 2003
49997	A2UENE1PINAKCT	0028633873	Marian E. Wells	[0, 0]	Wanted to know more about Yiddish and this boo...	5	YIDDISH for a Gentile?	1388620800	01 2, 2014
49998	ALA77HERW2U0J	0028633873	Michael Peterson	[1, 3]	While searching for a book that explains the Y...	5	This Yiddish book is a Feast for the Senses	1198627200	12 26, 2007
49999	A1LH5914M6CLTH	0028633873	Piet F. Van Allen	[2, 14]	It's a series - not to worry.Say, there's a "B...	3	Complete Idiot?	1181606400	06 12, 2007

50000 rows × 9 columns

3.Méthodologie proposée dans le TP:

Dans ce TP on voudrait tirer profit de la connaissance des collaborations entre entités pour effectuer une tâche prédictive.

Le TP suit les étapes suivantes :

1.Stratification base de données et visualisation

2.Calcul de similarités et voisinage

3.Prédiction des ratings et évaluation

4.Evaluation des résultats

a)

4.Implementation et analyse

Question 1 :

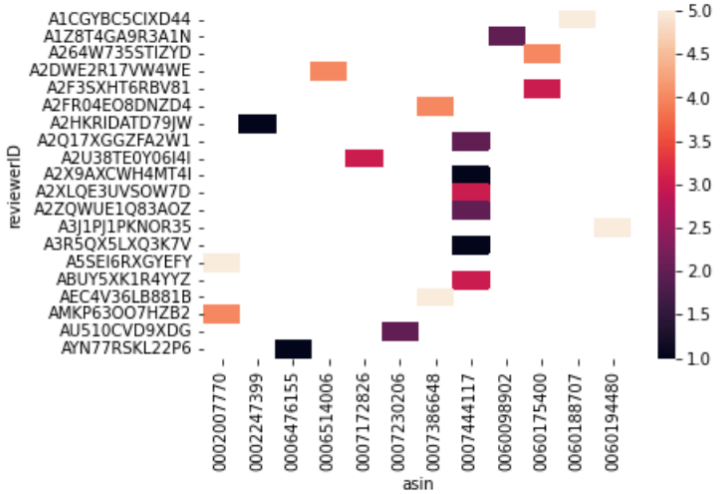
Je travaille uniquement avec un sous- échantillon prélevé de l'ensemble de données. De manière stratifié, on a p livres y compris tous les n reviewers qui ont participé aux ratings de ces livres. On obtient en fin de compte un dataframe similaire à celui ci-dessous:

asin	0002007770	0002247399	0006476155	0006514006	0007172826	0007230206	0007386648	0007444117	0060098902	0060175400	0060188707	0060194480
reviewerID												
A1CGYBC5CIXD44	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	5.0	NaN
A1Z8T4GA9R3A1N	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2.0	NaN	NaN	NaN
A264W735STIZYD	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	4.0	NaN	NaN
A2DWE2R17VW4WE	NaN	NaN	NaN	4.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
A2F3SXHT6RBV81	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	3.0	NaN	NaN
A2FR04EO8DNZD4	NaN	NaN	NaN	NaN	NaN	NaN	4.0	NaN	NaN	NaN	NaN	NaN
A2HKRIDATD79JW	NaN	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
A2Q17XGGZFA2W1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2.0	NaN	NaN	NaN	NaN
A2U38TE0Y06I4I	NaN	NaN	NaN	NaN	3.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
A2X9AXCWH4MT4I	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN
A2XLQE3UVSOW7D	NaN	NaN	NaN	NaN	NaN	NaN	NaN	3.0	NaN	NaN	NaN	NaN
A2ZQWUE1Q83AOZ	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2.0	NaN	NaN	NaN	NaN
A3J1PJ1PKNOR35	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	5.0
A3R5QX5LXQ3K7V	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN
A5SEI6RXGYEFY	5.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ABUY5XK1R4YYZ	NaN	NaN	NaN	NaN	NaN	NaN	NaN	3.0	NaN	NaN	NaN	NaN
AEC4V36LB881B	NaN	NaN	NaN	NaN	NaN	NaN	5.0	NaN	NaN	NaN	NaN	NaN
AMKP63OO7HZB2	4.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
AU510CVD9XDG	NaN	NaN	NaN	NaN	NaN	2.0	NaN	NaN	NaN	NaN	NaN	NaN
AYN77RSKL22P6	NaN	NaN	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

J'ai veillé à voir les lecteurs qui figurent le plus dans notre bas de données ce qui implique l'existence de leurs ratings sur plusieurs livres **Sauf que les livres les plus lus et ayant des ratings ne sont pas communs chez les lecteurs dominants .** Et du coup l'intersection (BONS LECTEURS) INTERSEC (LIVRES TROP LUS) contient peut de données et ceci peut être traduit par les valeurs NaN figurantes sur notre DataFrame ci-dessus**.**

On conclut donc à ce que les lecteurs lisent des livres mais pas trop de lecteurs lisent les mêmes livres on se trouve donc avec des lignes avec peu de données .

Le heatmap correspondant ressemble à ceci:



Remarque :

Lorsque je continue le travail avec le dataframe deja présenté je me retrouve avec des résultats non interprétables et non représentant le but attendu.
Tout d'abord une **similarité très très faible quasiment nulle** entre les differents livres .

asin	0002007770	0002247399	0006476155	0006514006	0007172826	0007230206	0007386648	0007444117	0060098902	0060175400	0060188707	0060194480
asin												
0002007770	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0002247399	0.0	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0006476155	0.0	0.0	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0006514006	0.0	0.0	0.0	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0007172826	0.0	0.0	0.0	0.0	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0007230206	0.0	0.0	0.0	0.0	0.0	1	0.0	0.0	0.0	0.0	0.0	0.0
0007386648	0.0	0.0	0.0	0.0	0.0	0.0	1	0.0	0.0	0.0	0.0	0.0
0007444117	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1	0.0	0.0	0.0	0.0
0060098902	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1	0.0	0.0	0.0
0060175400	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1	0.0	0.0
0060188707	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1	0.0
0060194480	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1

Ceci implique donc un **voisinnage aléatoire (non identifiant l'item)** de chaque livre avec les autres livres et chaque lecteurs avec les autres lecteurs .

reviewerID			
A1CGYBC5CIXD44	A1Z8T4GA9R3A1N	A264W735STIZYD	A2DWE2R17VW4WE
A1Z8T4GA9R3A1N	A1CGYBC5CIXD44	A264W735STIZYD	A2DWE2R17VW4WE
A264W735STIZYD	A1CGYBC5CIXD44	A1Z8T4GA9R3A1N	A2DWE2R17VW4WE
A2DWE2R17VW4WE	A1CGYBC5CIXD44	A1Z8T4GA9R3A1N	A264W735STIZYD
A2F3SXHT6RBV81	A1CGYBC5CIXD44	A1Z8T4GA9R3A1N	A264W735STIZYD
A2FR04EO8DNZD4	A1CGYBC5CIXD44	A1Z8T4GA9R3A1N	A264W735STIZYD
A2HKRIDATD79JW	A1CGYBC5CIXD44	A1Z8T4GA9R3A1N	A264W735STIZYD
A2Q17XGGZFA2W1	A1CGYBC5CIXD44	A1Z8T4GA9R3A1N	A264W735STIZYD
A2U38TE0Y06I4I	A1CGYBC5CIXD44	A1Z8T4GA9R3A1N	A264W735STIZYD
A2X9AXCWH4MT4I	A1CGYBC5CIXD44	A1Z8T4GA9R3A1N	A264W735STIZYD
A2XLQE3UVSOW7D	A1CGYBC5CIXD44	A1Z8T4GA9R3A1N	A264W735STIZYD
A2ZQWUE1Q83AOZ	A1CGYBC5CIXD44	A1Z8T4GA9R3A1N	A264W735STIZYD
A3J1PJ1PKNOR35	A1Z8T4GA9R3A1N	A264W735STIZYD	A2DWE2R17VW4WE
A3R5QX5LXQ3K7V	A1Z8T4GA9R3A1N	A264W735STIZYD	A2DWE2R17VW4WE
A5SEI6RXGYEFY	A1Z8T4GA9R3A1N	A264W735STIZYD	A2DWE2R17VW4WE
ABUY5XK1R4YYZ	A1Z8T4GA9R3A1N	A264W735STIZYD	A2DWE2R17VW4WE
AEC4V36LB881B	A1Z8T4GA9R3A1N	A264W735STIZYD	A2DWE2R17VW4WE
AMKP63OO7HZB2	A1Z8T4GA9R3A1N	A264W735STIZYD	A2DWE2R17VW4WE
AU510CVD9XDG	A1Z8T4GA9R3A1N	A264W735STIZYD	A2DWE2R17VW4WE
AYN77RSKL22P6	A1Z8T4GA9R3A1N	A264W735STIZYD	A2DWE2R17VW4WE

asin			
0002007770	0002247399	0006476155	0006514006
0002247399	0002007770	0006476155	0006514006
0006476155	0002007770	0002247399	0006514006
0006514006	0002007770	0002247399	0006476155
0007172826	0002007770	0002247399	0006476155
0007230206	0002007770	0002247399	0006476155
0007386648	0002007770	0002247399	0006476155
0007444117	0002007770	0002247399	0006476155
0060098902	0002007770	0002247399	0006476155
0060175400	0002007770	0002247399	0006476155
0060188707	0002007770	0002247399	0006476155
0060194480	0002007770	0002247399	0006476155

On visualise que l'on obtient les mêmes voisins chez tous les les livres ainsi que chez tous les lecteurs

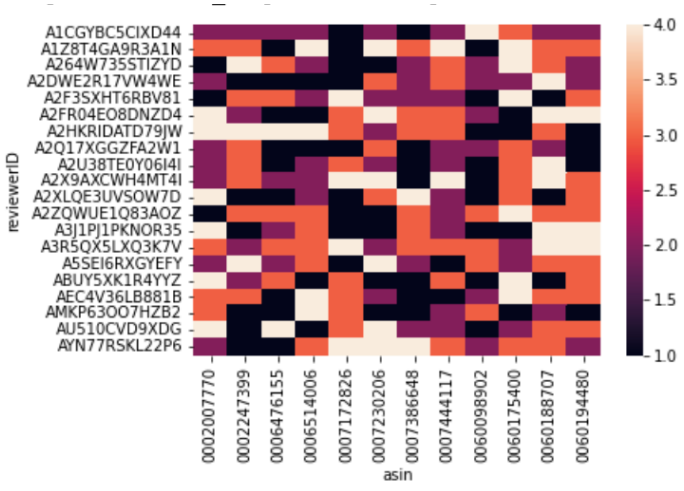
J'ai donc jugé intéressant de remplir aléatoirement un bon nombre de cases vides par des valeurs entre 1 et 5 .

b)

Dans tout ce qui suit je travaille avec le dataframe df_rand suivant qui présente plus de valeurs non nulles et de ratings de lecteurs sur la plupart des livres. Mon nouveau dataframe est donc :

asin	0002007770	0002247399	0006476155	0006514006	0007172826	0007230206	0007386648	0007444117	0060098902	0060175400	0060188707	0060194480
reviewerID												
A1CGYBC5CIXD44	2	2	1	4	1	4	4	1	3	1	3	1
A1Z8T4GA9R3A1N	4	4	4	1	4	4	3	4	4	2	4	4
A264W735STIZYD	2	1	3	3	4	3	2	3	1	2	4	4
A2DWE2R17VW4WE	4	1	3	3	1	2	3	1	3	4	4	2
A2F3SXHT6RBV81	1	3	4	4	3	4	3	3	3	3	2	2
A2FR04EO8DNZD4	4	3	1	3	3	3	4	4	2	1	4	4
A2HKRIDATD79JW	3	2	1	2	1	1	2	2	3	2	3	4
A2Q17XGGZFA2W1	4	3	3	2	2	3	2	3	4	4	4	1
A2U38TE0Y06I4I	2	2	3	1	4	4	4	4	1	3	3	4
A2X9AXCWH4MT4I	4	4	3	2	2	3	4	1	4	2	2	3
A2XLQE3UVSOW7D	1	1	2	1	2	2	4	3	2	3	3	3
A2ZQWUE1Q83AOZ	3	4	3	4	4	4	1	3	1	1	4	2
A3J1PJ1PKNOR35	3	4	1	3	2	2	1	2	3	2	4	1
A3R5QX5LXQ3K7V	3	2	1	1	2	4	1	1	2	2	3	4
A5SEI6RXGYEFY	1	3	3	4	1	3	2	4	3	1	1	1
ABUY5XK1R4YYZ	3	2	2	1	1	1	2	1	2	2	4	1
AEC4V36LB881B	2	1	1	3	1	3	1	3	4	1	4	4
AMKP63OO7HZB2	1	4	3	4	2	4	3	2	2	2	2	4
AU510CVD9XDG	2	2	4	2	1	1	4	3	1	3	4	3
AYN77RSKL22P6	2	3	1	3	4	2	2	2	1	1	1	4

D'un heatmap suivant:



b)

i) Maintenant que je sais à quoi ressemble mon nouveau jeu de données Fr1,on va extraire en extraire 3 sous-ensembles de données.
Je construis une fonction DataSelection() qui permet de mettre en valeurs nulles quelques cases de mon DataFrame on obtient donc Fr2:

```
Fr11=dataselection(D1)
Fr11
```

asin reviewerID	0002007770	0002247399	0007149824	0007172826	0007205236	0007386648	0007441290	0007442920	0007444117	0025853503
A188NTJ5LV8LA4	2.0	3	4	4	1	4	3	4.0	4	2
A1DX7THDO236Z8	2.0	1	2	4	2	1	1	1.0	4	2
A1Q55AHIMCT0YU	1.0	2	2	2	1	1	2	1.0	3	3
A1R2MFGY33970U	2.0	2	3	1	4	4	3	3.0	4	3
A201PBRJY4IZ3J	3.0	3	2	2	3	2	1	NaN	2	2
A25AKIPL88W14S	2.0	2	4	1	3	2	1	1.0	2	4
A27CH9AA9DCYI0	4.0	4	4	1	4	2	3	4.0	3	2
A2AOI18ADJZXBR	2.0	4	1	2	3	2	4	4.0	2	3
A2NJO6YE954DBH	4.0	2	2	3	2	2	4	1.0	4	2
A2PQXUXOFTSK5	2.0	2	3	3	2	2	1	2.0	3	4
A2WOGEE6LKMA4R	1.0	3	4	1	2	3	2	1.0	1	3
A2YB2H2BYQMIP1	4.0	3	4	1	1	2	1	2.0	1	4
A32D9LD2KU9C9C	1.0	2	2	1	4	2	2	4.0	4	4
A34J1LC1GEF3Q5	1.0	4	1	3	3	4	3	4.0	2	4
A3AZP441UVMTNO	NaN	2	3	3	2	4	4	4.0	2	2

ii) Repéte l'étape 1(b)i trois (03) fois de suite et retourner 3 sous-ensembles d'entrainement Train Sets = {Fr1 ,Fr2 ,Fr3 }.

```
Fr13=dataselection(D3)
Fr13
```

asin reviewerID	0002007770	0002247399	0007149824	0007172826	0007205236	0007386648	0007441290	0007442920	0007444117	0025853503
A188NTJ5LV8LA4	2.0	3	4	4.0	1	4	3.0	4.0	4.0	2
A1DX7THDO236Z8	2.0	1	2	4.0	2	1	1.0	1.0	4.0	2
A1Q55AHIMCT0YU	1.0	2	2	2.0	1	1	2.0	1.0	NaN	3
A1R2MFGY33970U	2.0	2	3	1.0	4	4	3.0	3.0	4.0	3
A201PBRJY4IZ3J	3.0	3	2	2.0	3	2	1.0	NaN	2.0	2
A25AKIPL88W14S	2.0	2	4	1.0	3	2	1.0	1.0	2.0	4
A27CH9AA9DCYI0	4.0	4	4	1.0	4	2	3.0	4.0	3.0	2
A2AOI18ADJZXBR	2.0	4	1	2.0	3	2	4.0	4.0	2.0	3
A2NJO6YE954DBH	4.0	2	2	NaN	2	2	4.0	1.0	NaN	2
A2PQXUXOFTSK5	2.0	2	3	3.0	2	2	1.0	2.0	3.0	4
A2WOGEE6LKMA4R	1.0	3	4	1.0	2	3	2.0	1.0	1.0	3
A2YB2H2BYQMIP1	4.0	3	4	1.0	1	2	1.0	2.0	1.0	4
A32D9LD2KU9C9C	1.0	2	2	1.0	4	2	2.0	4.0	4.0	4
A34J1LC1GEF3Q5	1.0	4	1	3.0	3	4	3.0	4.0	2.0	4
A3AZP441UVMTNO	NaN	2	3	3.0	2	4	4.0	4.0	2.0	2
A3INUWOQUE8RFF	NaN	1	2	2.0	3	4	2.0	2.0	3.0	1
A3Q3Y39W2F0E3J	3.0	1	2	3.0	2	1	1.0	4.0	2.0	1
A8E9VWP5EQH6B	4.0	4	3	2.0	3	1	2.0	3.0	4.0	3
AWNNXG7J5UD98	1.0	3	3	1.0	3	3	4.0	1.0	1.0	3
AY2527VC3C5GG	1.0	4	3	3.0	4	4	NaN	2.0	2.0	2

asin reviewerID	0002007770	0002247399	0006476155	0006514006	0007172826	0007230206	0007386648	0007444117	0060098902	0060175400	0060188707	0060194480
A1CGYBC5CIXD44	4	3.0	1.0	4	4	4	2	2.0	4	4.0	2	1
A1Z8T4GA9R3A1N	2	4.0	1.0	4	1	1	3	3.0	1	2.0	2	4
A264W735STIZYD	2	2.0	4.0	4	2	3	1	2.0	4	1.0	3	3
A2DWE2R17VW4WE	4	4.0	3.0	3	3	1	1	3.0	4	2.0	3	1
A2F3SXHT6RBV81	4	1.0	1.0	1	2	1	4	4.0	3	NaN	1	4
A2FR04EO8DNZD4	3	4.0	1.0	2	2	3	4	2.0	2	1.0	4	1
A2HKRIDATD79JW	3	3.0	1.0	4	2	3	2	3.0	1	3.0	3	2
A2Q17XGGZFA2W1	4	4.0	1.0	4	3	4	3	4.0	2	4.0	2	4
A2U38TE0Y06I4I	2	3.0	2.0	3	2	3	3	2.0	3	1.0	2	2
A2X9AXCWH4MT4I	3	4.0	3.0	3	3	2	2	4.0	3	2.0	4	4
A2XLQE3UVSOW7D	1	2.0	4.0	4	1	4	1	3.0	4	2.0	4	1
A2ZQWUE1Q83AOZ	3	1.0	1.0	4	3	4	1	2.0	4	4.0	2	2
A3J1PJ1PKNOR35	2	2.0	3.0	2	3	3	2	3.0	3	4.0	4	3
A3R5QX5LXQ3K7V	2	3.0	NaN	2	1	1	4	2.0	3	2.0	1	2
A5SEI6RXGYEFY	1	1.0	3.0	3	1	4	3	NaN	1	4.0	2	4
ABUY5XK1R4YYZ	4	4.0	1.0	2	3	4	4	4.0	3	1.0	1	2
AEC4V36LB881B	3	1.0	1.0	2	3	2	4	2.0	4	2.0	1	3
AMKP63OO7HZB2	2	NaN	2.0	4	4	3	4	4.0	3	4.0	2	3
AU510CVD9XDG	2	4.0	4.0	2	3	1	1	2.0	1	2.0	4	3
AYN77RSKL22P6	4	4.0	1.0	1	4	4	3	4.0	4	3.0	2	1

C)

i) les matrices de similitudes Bk ∈ Rp×p des livres et les matrices de similitudes Uk ∈ Rn×n des reviewers.

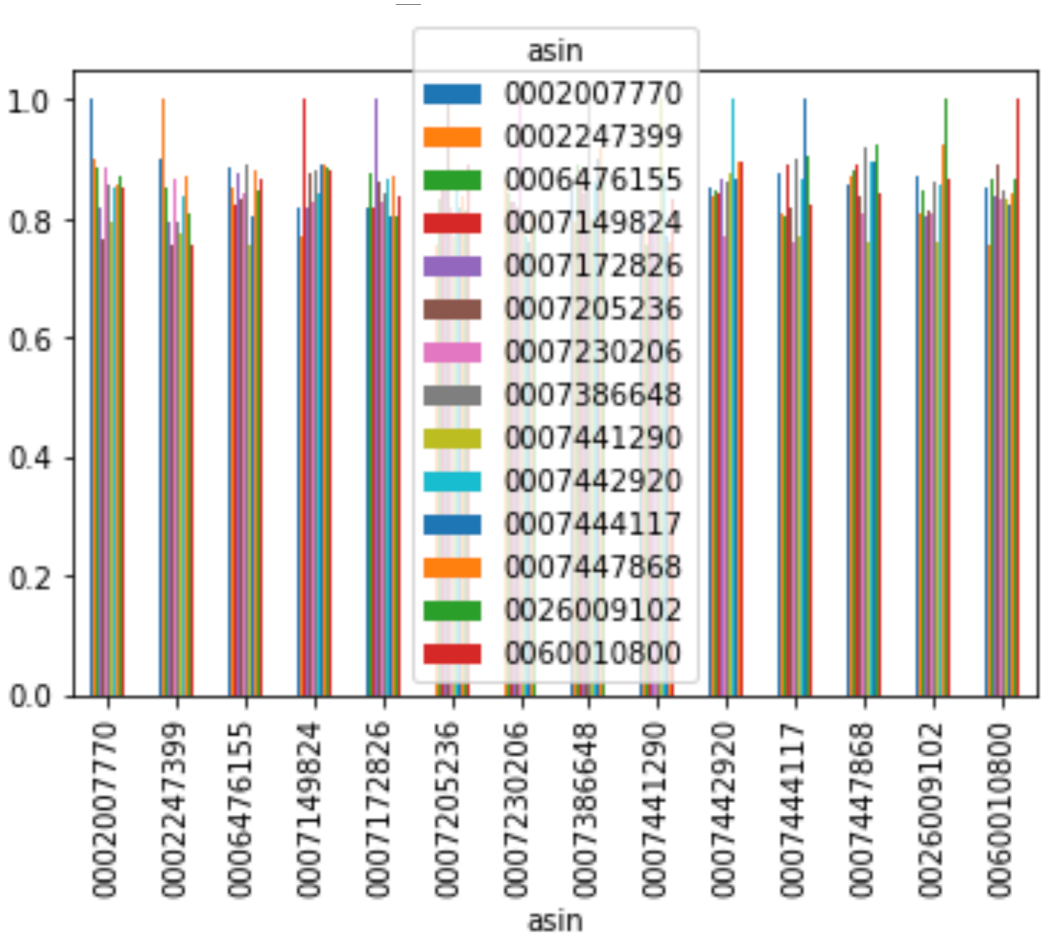
```
data_ibs
```

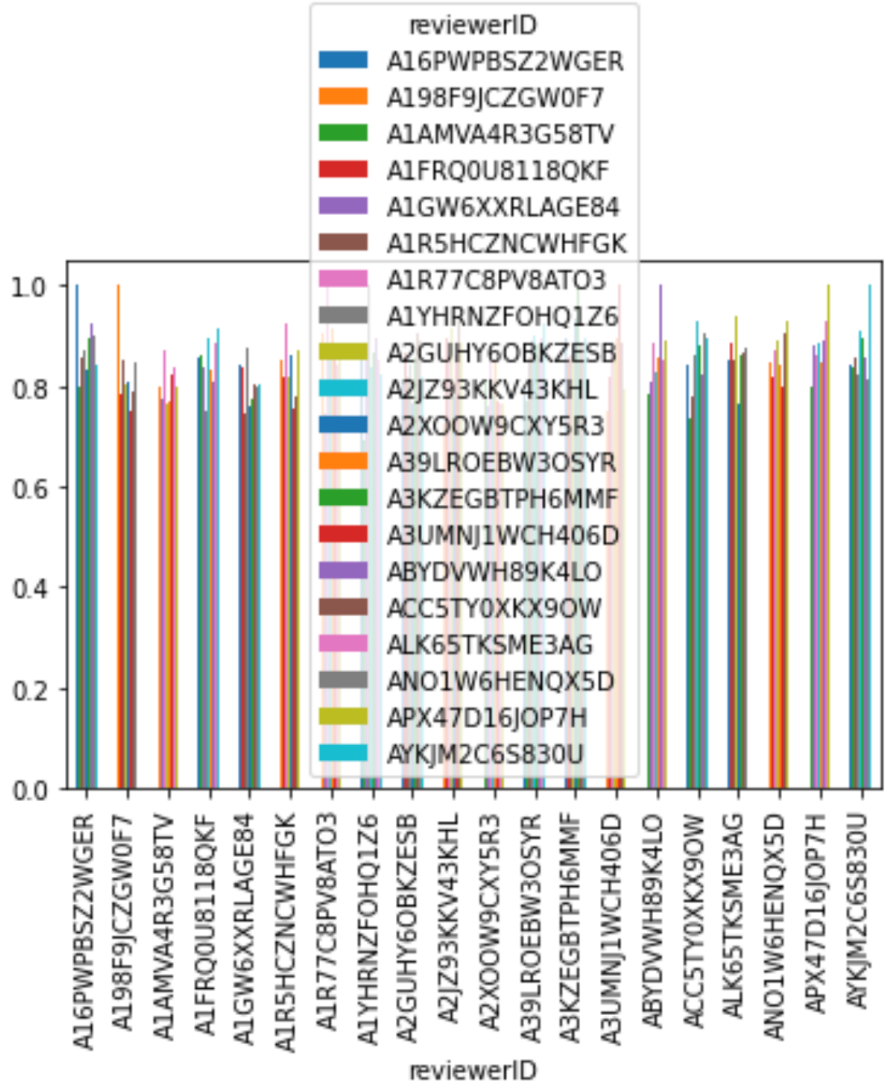
asin asin	0002007770	0002247399	0007149824	0007172826	0007205236	0007386648	0007441290	0007442920	0007444117	0025853503	0026009102	0060175400	0060256656	0060393491	0060510862
0002007770	1	0.832188	0.833772	0.667124	0.77666	0.648074	0.715525	0.728103	0.740013	0.79688	0.77746	0.799219	0.750446	0.809101	0.661186
0002247399	0.832188	1	0.894029	0.800795	0.909696	0.875985	0.848788	0.850721	0.80423	0.912785	0.884277	0.841361	0.825896	0.882245	0.850727
0007149824	0.833772	0.894029	1	0.796333	0.868354	0.879856	0.800001	0.804378	0.837547	0.902439	0.885592	0.827255	0.841142	0.8769	0.856701
0007172826	0.667124	0.800795	0.796333	1	0.774371	0.82466	0.707592	0.82717	0.859767	0.781019	0.857689	0.720479	0.751921	0.860291	0.804738
0007205236	0.77666	0.909696	0.868354	0.774371	1	0.888235	0.825567	0.849691	0.884891	0.887231	0.895974	0.888115	0.809174	0.865333	0.863195
0007386648	0.648074	0.875985	0.879856	0.82466	0.888235	1	0.858324	0.841078	0.820157	0.847977	0.901789	0.774314	0.804984	0.853872	0.916949
0007441290	0.715525	0.848788	0.800001	0.707592	0.825567	0.858324	1	0.845432	0.735609	0.820871	0.878275	0.837235	0.81325	0.850365	0.777291
0007442920	0.728103	0.850721	0.804378	0.82717	0.849691	0.841078	0.845432	1	0.86998	0.817045	0.805652	0.817711	0.792267	0.842543	0.728717
0007444117	0.740013	0.80423	0.837547	0.859767	0.884891	0.820157	0.735609	0.86998	1	0.817605	0.832911	0.800005	0.722691	0.86035	0.808911
0025853503	0.79688	0.912785	0.902439	0.781019	0.887231	0.847977	0.820871	0.817045	0.817605	1	0.897169	0.853941	0.826886	0.942667	0.79599
0026009102	0.77746	0.884277	0.885592	0.857689	0.895974	0.901789	0.878275	0.805652	0.832911	0.897169	1	0.892942	0.866131	0.931257	0.883672
0060175400	0.799219	0.841361	0.827255	0.720479	0.888115	0.774314	0.837235	0.817711	0.800005	0.853941	0.892942	1	0.881308	0.827505	0.693769
0060256656	0.750446	0.825896	0.841142	0.751921	0.809174	0.804984	0.81325	0.792267	0.722691	0.826886	0.866131	0.881308	1	0.832923	0.741284
0060393491	0.809101	0.882245	0.8769	0.860291	0.865333	0.853872	0.850365	0.842543	0.86035	0.942667	0.931257	0.827505	0.832923	1	0.848843
0060510862	0.661186	0.850727	0.856701	0.804738	0.863195	0.916949	0.777291	0.728717	0.808911	0.79599	0.883672	0.693769	0.741284	0.848843	1

data_ubs

reviewerID	A16PWPBSZ2WGER	A198F9JCZGW0F7	A1AMVA4R3G58TV	A1FRQ0U8118QKF	A1GW6XXRLAGE84	A1R5HCZNCWHFGK	A1R77C8PV8ATO3	A1YHRNZFOHQ1Z6	A2GUHY6OBKZESB	A2JZ93KKV43KHL	A2XOOW9CXY5R3	A39LROEBW3OSYR
reviewerID												
A16PWPBSZ2WGER	1	0.811381	0.796942	0.85728	0.839789	0.858441	0.834377	0.868421	0.81263	0.81785	0.833561	0.7964
A198F9JCZGW0F7	0.811381	1	0.800163	0.783092	0.731664	0.850487	0.902077	0.790306	0.804547	0.726163	0.810063	0.795169
A1AMVA4R3G58TV	0.796942	0.800163	1	0.859738	0.775849	0.888028	0.871183	0.68989	0.762866	0.783531	0.812801	0.771365
A1FRQ0U8118QKF	0.85728	0.783092	0.859738	1	0.836132	0.819853	0.874894	0.75012	0.863636	0.895937	0.819978	0.831809
A1GW6XXRLAGE84	0.839789	0.731664	0.775849	0.836132	1	0.744839	0.746131	0.876703	0.807948	0.860415	0.758704	0.805204
A1R5HCZNCWHFGK	0.858441	0.850487	0.888028	0.819853	0.744839	1	0.925973	0.796491	0.819853	0.816152	0.859869	0.855653
A1R77C8PV8ATO3	0.834377	0.902077	0.871183	0.874894	0.746131	0.925973	1	0.817689	0.840918	0.854033	0.873053	0.914936
A1YHRNZFOHQ1Z6	0.868421	0.790306	0.68989	0.75012	0.876703	0.796491	0.817689	1	0.83942	0.83563	0.786732	0.889486
A2GUHY6OBKZESB	0.81263	0.804547	0.762866	0.863636	0.807948	0.819853	0.840918	0.83942	1	0.914037	0.76277	0.884456
A2JZ93KKV43KHL	0.81785	0.726163	0.783531	0.895937	0.860415	0.816152	0.854033	0.83563	0.914037	1	0.77831	0.901426
A2XOOW9CXY5R3	0.833561	0.810063	0.812801	0.819978	0.758704	0.859869	0.873053	0.786732	0.76277	0.77831	1	0.850323
A39LROEBW3OSYR	0.7964	0.795169	0.771365	0.831809	0.805204	0.855653	0.914936	0.889486	0.884456	0.901426	0.850323	1
A3KZEGBTPH6MMF	0.89568	0.810127	0.800163	0.858183	0.776007	0.861118	0.861985	0.864068	0.858183	0.84363	0.933823	0.882141
A3UMNJ1WCH406D	0.80418	0.750097	0.821053	0.818671	0.796266	0.754059	0.863919	0.851485	0.857197	0.862915	0.767716	0.892421
ABYDVWH89K4LO	0.9229	0.860971	0.785481	0.809595	0.805666	0.881577	0.887192	0.893446	0.75962	0.82584	0.880559	0.856651
ACC5TY0XKX9OW	0.840433	0.789099	0.733548	0.875245	0.802869	0.779681	0.891297	0.859753	0.904748	0.930033	0.825137	0.899821
ALK65TKSME3AG	0.852936	0.791277	0.835939	0.8855	0.870673	0.851998	0.843435	0.852936	0.937082	0.958526	0.766419	0.896155
ANO1W6HENQX5D	0.898687	0.846715	0.71683	0.816217	0.797151	0.853342	0.871583	0.898687	0.887973	0.883964	0.837242	0.84147
APX47D16JOP7H	0.87381	0.884559	0.798794	0.83958	0.87797	0.871672	0.859175	0.913082	0.89955	0.885539	0.765248	0.845075
AYKJM2C6S830U	0.840433	0.707868	0.83834	0.914582	0.802869	0.857649	0.909674	0.821112	0.81624	0.910453	0.866394	0.922602

D'après les matrices ci dessus et les bareaux si dessous on a une similarité remarquable entre plusieurs livres et plusieurs lecteurs . Ceci va nous aider pour pouvoir recommander a l'aide du filtrage collaboratif :





i) On obtient donc les plus proches voisins de chacun de nos livres ainsi que le voisinage de chaque lecteur:

```
livres_voisins #la première colonne est a supprimer ,c'est le livre lui même
```

asin	1	2	3	4
0002007770	0002007770	0007149824	0002247399	0060393491
0002247399	0002247399	0025853503	0007205236	0007149824
0007149824	0007149824	0025853503	0002247399	0026009102
0007172826	0007172826	0060393491	0007444117	0026009102
0007205236	0007205236	0002247399	0026009102	0007386648
0007386648	0007386648	0060510862	0026009102	0007205236
0007441290	0007441290	0026009102	0007386648	0060393491
0007442920	0007442920	0007444117	0002247399	0007205236
0007444117	0007444117	0007205236	0007442920	0060393491
0025853503	0025853503	0060393491	0002247399	0007149824
0026009102	0026009102	0060393491	0007386648	0025853503
0060175400	0060175400	0026009102	0007205236	0060256656
0060256656	0060256656	0060175400	0026009102	0007149824
0060393491	0060393491	0025853503	0026009102	0002247399
0060510862	0060510862	0007386648	0026009102	0007205236

```
revs_voisins
```

reviewerID	1	2	3	4	5	6	7	8
A188NTJ5LV8LA4	A188NTJ5LV8LA4	A1Q55AHIMCT0YU	A1DX7THDO236Z8	A34J1LC1GEF3Q5	A2YB2H2BYQMIP1	A2PQXUXOFTSK5	A2WOGEE6LKMA4R	A201PBRJY4IZ3J
A1DX7THDO236Z8	A1DX7THDO236Z8	A2PQXUXOFTSK5	A201PBRJY4IZ3J	A1Q55AHIMCT0YU	A2WOGEE6LKMA4R	A34J1LC1GEF3Q5	A25AKIPL88W14S	A3AZP441UVMTNO
A1Q55AHIMCT0YU	A1Q55AHIMCT0YU	A2PQXUXOFTSK5	A1DX7THDO236Z8	A2WOGEE6LKMA4R	A25AKIPL88W14S	A34J1LC1GEF3Q5	A201PBRJY4IZ3J	A3AZP441UVMTNO
A1R2MFGY33970U	A1R2MFGY33970U	A34J1LC1GEF3Q5	A2NJO6YE954DBH	A2WOGEE6LKMA4R	A2AOI18ADJZXBR	A25AKIPL88W14S	A3AZP441UVMTNO	A1DX7THDO236Z8
A201PBRJY4IZ3J	A201PBRJY4IZ3J	A1DX7THDO236Z8	A2WOGEE6LKMA4R	A25AKIPL88W14S	A2YB2H2BYQMIP1	A2PQXUXOFTSK5	A2NJO6YE954DBH	A1Q55AHIMCT0YU
A25AKIPL88W14S	A25AKIPL88W14S	A3AZP441UVMTNO	A2WOGEE6LKMA4R	A201PBRJY4IZ3J	A1Q55AHIMCT0YU	A1DX7THDO236Z8	A27CH9AA9DCYI0	A34J1LC1GEF3Q5
A27CH9AA9DCYI0	A27CH9AA9DCYI0	A2WOGEE6LKMA4R	A25AKIPL88W14S	A34J1LC1GEF3Q5	A1DX7THDO236Z8	A2AOI18ADJZXBR	A2YB2H2BYQMIP1	A201PBRJY4IZ3J
A2AOI18ADJZXBR	A2AOI18ADJZXBR	A2NJO6YE954DBH	A1DX7THDO236Z8	A201PBRJY4IZ3J	A27CH9AA9DCYI0	A34J1LC1GEF3Q5	A25AKIPL88W14S	A1R2MFGY33970U
A2NJO6YE954DBH	A2NJO6YE954DBH	A201PBRJY4IZ3J	A2AOI18ADJZXBR	A34J1LC1GEF3Q5	A1R2MFGY33970U	A1Q55AHIMCT0YU	A2WOGEE6LKMA4R	A25AKIPL88W14S
A2PQXUXOFTSK5	A2PQXUXOFTSK5	A34J1LC1GEF3Q5	A1DX7THDO236Z8	A1Q55AHIMCT0YU	A2WOGEE6LKMA4R	A201PBRJY4IZ3J	A2YB2H2BYQMIP1	A25AKIPL88W14S
A2WOGEE6LKMA4R	A2WOGEE6LKMA4R	A34J1LC1GEF3Q5	A25AKIPL88W14S	A2PQXUXOFTSK5	A201PBRJY4IZ3J	A2YB2H2BYQMIP1	A1Q55AHIMCT0YU	A1DX7THDO236Z8
A2YB2H2BYQMIP1	A2YB2H2BYQMIP1	A2WOGEE6LKMA4R	A201PBRJY4IZ3J	A32D9LD2KU9C9C	A2PQXUXOFTSK5	A1DX7THDO236Z8	A27CH9AA9DCYI0	A34J1LC1GEF3Q5
A32D9LD2KU9C9C	A32D9LD2KU9C9C	A2YB2H2BYQMIP1	A2WOGEE6LKMA4R	A1Q55AHIMCT0YU	A34J1LC1GEF3Q5	A2PQXUXOFTSK5	A1DX7THDO236Z8	A27CH9AA9DCYI0
A34J1LC1GEF3Q5	A34J1LC1GEF3Q5	A2PQXUXOFTSK5	A2WOGEE6LKMA4R	A1DX7THDO236Z8	A1Q55AHIMCT0YU	A201PBRJY4IZ3J	A2NJO6YE954DBH	A1R2MFGY33970U
A3AZP441UVMTNO	A3AZP441UVMTNO	A25AKIPL88W14S	A2WOGEE6LKMA4R	A201PBRJY4IZ3J	A1Q55AHIMCT0YU	A1DX7THDO236Z8	A34J1LC1GEF3Q5	A2NJO6YE954DBH
A3INUWOQUE8RFF	A188NTJ5LV8LA4	A1DX7THDO236Z8	A1Q55AHIMCT0YU	A1R2MFGY33970U	A201PBRJY4IZ3J	A25AKIPL88W14S	A27CH9AA9DCYI0	A2AOI18ADJZXBR
A3Q3Y39W2F0E3J	A188NTJ5LV8LA4	A1DX7THDO236Z8	A1Q55AHIMCT0YU	A1R2MFGY33970U	A201PBRJY4IZ3J	A25AKIPL88W14S	A27CH9AA9DCYI0	A2AOI18ADJZXBR
A8E9VWP5EQH6B	A188NTJ5LV8LA4	A1DX7THDO236Z8	A1Q55AHIMCT0YU	A1R2MFGY33970U	A201PBRJY4IZ3J	A25AKIPL88W14S	A27CH9AA9DCYI0	A2AOI18ADJZXBR
AWNNXG7J5UD98	A188NTJ5LV8LA4	A1DX7THDO236Z8	A1Q55AHIMCT0YU	A1R2MFGY33970U	A201PBRJY4IZ3J	A25AKIPL88W14S	A27CH9AA9DCYI0	A2AOI18ADJZXBR
AY2527VC3C5GG	A188NTJ5LV8LA4	A1DX7THDO236Z8	A1Q55AHIMCT0YU	A1R2MFGY33970U	A201PBRJY4IZ3J	A25AKIPL88W14S	A27CH9AA9DCYI0	A2AOI18ADJZXBR

d)

Dans cette partie le but est de se baser sur les 5 voisins d'un lecteur i pour prédire son rating à un certain livre j.

L'idée est donc de calculer le rating que chaque lecteur pourrait donner à un livre et de le lui recommander si jamais il l'apprécie.

Pour ce calcul de rating on a la formule suivante:

$$\mathcal{R}(livre_j|reviewer_i,\Omega) = \left\| (\omega_1,\omega_2,\omega_3,\omega_4,\omega_5) \cdot \sum_{livre_{j^*} \in \boldsymbol{I}^k_{\sim j}} Z^k_* \right\|$$

avec $\Omega = \{\omega_1,\omega_2,\omega_3,\omega_4,\omega_5\}$, l'ensemble des paramètres.
$Z^k_* \in R^{5\times 8}$ est une matrice dont les éléments $z^k_{i^*,j^*}$ sont définis comme suit:

$$z^k_{i^*,j^*} = \begin{cases} r_{i^*,j^*} \; si \; le \; reviewer_{i^*} \; a \; une \; appréciation \; au \; livre_{j^*} \\ 0 \; sinon \end{cases}$$

A l'aide de celle ci j'arrive à retrouver une nouvelle matrice de ratings qui se base sur les 5 voisinages de chaque lecteur et sur leurs ratings pour donner le rating du lecteur en question:

```
R=RatingPredict(W,D)
R
```

```
[asin            0002007770  0002247399  0006514006  0006551807  0007205236  \
 reviewerID
 A12PSSIA6RUKC6    1.841476    0.920738    3.682952    0.920738    3.682952
 A1G9UEWALDT9JI    0.920738    3.682952    2.762214    2.762214    3.682952
 A1KJO5VP4K3CHU    1.841476    3.682952    0.920738    3.682952    0.920738
 A2EBLL2OYEQJN9    2.762214    2.762214    2.762214    3.682952    3.682952
 A2F5JRRB7URWMC    0.920738    3.682952    0.920738    3.682952    1.841476
 A2FAX01N7TD75Z    2.762214    3.682952    0.920738    0.920738    3.682952
 A2HAEO73497YIY    1.841476    2.762214    3.682952    0.920738    2.762214
 A34A1BNPEFZ6FC    3.682952    2.762214    1.841476    0.920738    0.920738
 A35O7F1LIXICYD    3.682952    2.762214    2.762214    3.682952    0.920738
 A3E9MQ6LX3H06B    0.920738    3.682952    1.841476    2.762214    2.762214
 A3FTQ64U1DJNDR    0.920738    2.762214    2.762214    1.841476    0.920738
 A3FUMXE6QB1RM1    2.762214    1.841476    0.920738    3.682952    3.682952
 A3O3XWYEY1MKZA    3.682952    3.682952    2.762214    2.762214    1.841476
 A3SI9SQF6US0IB    1.841476    2.762214    3.682952    1.841476    3.682952
 A3UCKXBO457YD6    3.682952    1.841476    0.920738    3.682952    2.762214
 A6V1EHSAN0GUI     3.682952    2.762214    0.920738    0.920738    0.920738
 AEUR5WPP9BKNJ     1.841476    0.920738    1.841476    2.762214    2.762214
 AH6SCCSDOTNWM     3.682952    1.841476    1.841476    2.762214    3.682952
 AKP40I31XYX1D     2.762214    2.762214    2.762214    3.682952    1.841476
```

6.Ressources

https://mrmint.fr/gradient-descent-algorithm

https://pandas.pydata.org/pandasdocs/stable/user_guide/visualization.html

http://www.python-simple.com/python-pandas/dataframes-indexation.php

Fin