UNIVERSITÉ DE SHERBROOKE

IFT 599/799 – Science des données

TP2 & 3: Segmentation des données et visualisation.

Professeur:

Shengrui Wang & Etienne G. Tajeuna

Réalisé par :

BENHAMMOU Nouhayla 22 149 177
DERFOUFI Asmae 22 148 150

Mail :

Nouhayla.Benhammou@USherbrooke.ca
Asmae.Derfoufi@USherbrooke.ca

Année Universitaire 2022/2023

Contents

List of Figures

Introduction générale

La revue des livres est un indicateur décisif aux clients pour les aider à chisir de lire un livre ou pas. Dans cette série de travaux pratiques nous allons exploiter la base de données ABR,afin de donner des recommendations aux utilisateurs.Précedemment dans le TP1,nous avons effectuer un pretraitement de données et une étude statistique pour visualiser les différents résultats et les analyser par la suite.Dans ce rapport nous allons effectuer une segmentation sur les données tout en exploitant les différentes techniques de clustering vu au cours notamment le KMeans et le Clmustering spectral.

Partie A

1 Description des données

Le jeu de données ABR est un dataset qui décrit les opinions sur les livres. Ce dataset a été collecté depuis Amazon et contient:

- reviewerID - ID of the reviewer
- asin - ID of the product
- reviewerName - name of the reviewer
- helpful - helpfulness rating of the review
- reviewText - text of the review
- overall - rating of the product
- summary - summary of the review
- unixReviewTime - time of the review (unix time)
- reviewTime - time of the review (raw)

Tout au long de ce Tp nous allons travaillé avec 200000 données et nous avons choisi qe les visualier sur un tableau avant d'entamer les questions. La matrice obtenue a la forme 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
...	...	...	...	...	...	...	...	...	...
199995	A15P0E7IYRBXAX	0060975776	Tom Badyna	[15, 22]	I tried to like this collection, and there wer...	2	Just Not Sure	1208390400	04 17, 2008
199996	A1WX7UUXN4WNL0	0060975776	T. Tucker	[3, 3]	I love this book. The ending lines about "I n...	5	brilliant and beautiful	1111276800	03 20, 2005
199997	A2R6LKLYSJIQH8	0060975776	T T Walker	[0, 0]	I don't think that Jesus' Son should be put wi...	4	Left me wanting more	1358380800	01 17, 2013
199998	A3QYDL5CDNYN66	0060975776	Verita "a devoted reader"	[0, 0]	Stories involving drug addiction. Johnson's gr...	5	Spectacularly accurate and beautifully written	1285372800	09 25, 2010
199999	AI0A3WN9QZVSL	0060975776	Voice of Chunk	[6, 7]	If you're a fan of Johnson's manic, drug-fuele...	5	A groundbreaking short story collection	956620800	04 25, 2000

200000 rows × 9 columns

Figure 1: Dataframe des données

2 Bibliothèques

Afin d'effectuer le travail demandé, il nous faut importer quelques bibliothèques de python qui vons nous permettre d'implementer le code plus facilement.Tout au long de ce TP nous avons utiliser les bibliothèques suivantes:

- **numpy** nous a permis de manipuler des matrices ou tableaux multidimensionnels ainsi que des fonctions mathématiques opérant sur ces tableaux.Elle nous a servi aussi par son module **linalg** dans la partie de la décomposition en valeurs singulières.
- **pandas** nous a permi la manipulation et l'analyse des données.
- **scipy** nous a permi de calculer la matrice de distance euclidienne grace à **distance_matrix**.
- **sklearn** destinée à l'implementation des algorithmes d'apprentissage automatique.Nous avons exploiter les modules **StratifiedKFold,silhouette_score,PCA, KMeans, StandardScaler, cosine_similarity.**
- **matplotlib** nous l'avons utilisée dans pour pouvoir visualiser nos figures.

3 Prétraitement des données

Afin de faciliter le travail, nous avons effectuer un petit traitement sur nos données.Grace à la fonction matrice_score que nous avons codée,nous avons pu former une nouvelle dataframe qui illustre pour chaque livre son score collecté par les utilisateurs suivant les ratings 1,2,3,4 et 5.La figure suivante montre le résultat obtenu:

	000100039X	0001055178	0001473123	0001473727	0001473905	0001712772	000171287X	0001714538	0002005395	0002006715	...	0028631919	002863196X	002863201X	0028632028
1	6	0.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	...	2.0	0.0	0.0	0.0
2	4	4.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	...	0.0	1.0	2.0	1.0
3	8	2.0	0.0	0.0	1.0	0.0	2.0	0.0	5.0	1.0	...	0.0	0.0	1.0	1.0
4	15	10.0	2.0	0.0	0.0	1.0	1.0	2.0	3.0	0.0	...	2.0	2.0	4.0	1.0
5	173	2.0	13.0	7.0	5.0	13.0	9.0	3.0	6.0	4.0	...	14.0	10.0	15.0	6.0

Figure 2: Prétraitement des données

4 Réponses aux questions:

4.1 Question 1

Construire la matrice de données X Rp×c avec c étant le nombre de caractérisitques statistiques extraites.
Nous avons exploiter la matrice obtenue dans le prétraitement des données pour trouver trois groupes de ratings , Aprécié qui correspondent aux ratings 4-5 , Neutre correspondent à 3, et finalement les Moins Appréciés qui correspondent aux ratings 1-2.Nous avons aussi ajouter une dimensions qui correspond au nombre de votes pour chaque livres. Dans une deuxième matrice nous avons trouver pour chaque livre sa moyenne, sonécart-type, et sa médiane des scores.

	Apprécié	Neutre	Moins_Apprécié	Nombre_Votes
000100039X	188.0	8.0	10.0	206.0
0001055178	12.0	2.0	4.0	18.0
0001473123	15.0	0.0	1.0	16.0
0001473727	7.0	0.0	0.0	7.0
0001473905	5.0	1.0	0.0	6.0
...	...	...	...	...
0060534095	5.0	0.0	0.0	5.0
0060534214	41.0	4.0	10.0	55.0
0060534249	29.0	5.0	4.0	38.0
0060534389	53.0	9.0	4.0	66.0
0060534397	26.0	3.0	0.0	29.0

	overall		
	mean	std	median
asin			
000100039X	4.674757	0.875712	5.0
0001055178	3.555556	0.983524	4.0
0001473123	4.625000	1.024695	5.0
0001473727	5.000000	0.000000	5.0
0001473905	4.666667	0.816497	5.0
...	...	...	...
0028632613	3.722222	1.363626	4.0
0028632753	2.400000	0.547723	2.0
0028633504	4.625000	0.875388	5.0
0028633784	4.428571	0.786796	5.0
0028633873	4.625000	0.744024	5.0

Figure 3: Caractéristiques ds livres

Ensuite nous avons fusionner les deux dataframes pour obtenir la matrice X demandée par la questions et qui illustre les différentes caractériqtiques extraites pour chaque livres.

	Apprécié	Neutre	Moins_Apprécié	Nombre_Votes	(overall, mean)	(overall, std)	(overall, median)
000100039X	188.0	8.0	10.0	206.0	4.674757	0.875712	5.0
0001055178	12.0	2.0	4.0	18.0	3.555556	0.983524	4.0
0001473123	15.0	0.0	1.0	16.0	4.625000	1.024695	5.0
0001473727	7.0	0.0	0.0	7.0	5.000000	0.000000	5.0
0001473905	5.0	1.0	0.0	6.0	4.666667	0.816497	5.0
...	...	...	...	...	...	...	...
0060534095	5.0	0.0	0.0	5.0	4.800000	0.447214	5.0
0060534214	41.0	4.0	10.0	55.0	3.872727	1.291777	4.0
0060534249	29.0	5.0	4.0	38.0	4.052632	1.137740	4.0
0060534389	53.0	9.0	4.0	66.0	4.075758	0.899754	4.0
0060534397	26.0	3.0	0.0	29.0	4.275862	0.648986	4.0

Figure 4: Matrice finale des caractéristiques de chaque livre

4.2 Question 2

Effectuer une segmentation basée sur les k-moyennes avec k = 3 en utilisant dans un premier temps la distance euclidienne et ensuite la similarité cosinus pour comparer les objets.

Tout d'abord nous avons effectuer la méthode elbow(ou bien la méthode su coude) sur nos données afin de s'assurer si le nombre de clusters déterminé par la question est le meilleur à choix(k=3).

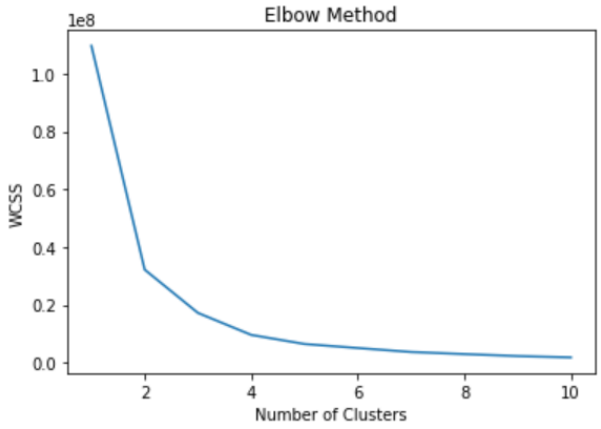

Figure 5: Méthode du coude

Par la suite, nous avons réaliser le clustering sur notre matrice X moyennant l'algorithme KMeans.Dans un premier temps nous avons tester l'algorithme sur la matrice X toute entière et par la suite nous avons teser avec des couples de dimensions aléatoires pour voir ceux qui donnent le meilleur résultat.Finalement nous avons trouver que l'association des dimensions Nombres de votes avec la moyenne donnent les meilleurs résultats.Par conséquent, nous avons choisi de les garder pour effectuer notre clustering. Il faut noter que nous avons effectuer une standarisation pour les données en entrée du KMeans avec la distance euclidienne.

Dans une seconde étape, nous avons effectuer le KMeans avec la métrique "cosine".Tout d'abord nous avons chercher les dimensions de X qui vont donner les meilleurs résultats et nous avons choisi les trois premières dimensions à savoir Apprécié,Neutre, et Moins Apprécié.Par la suite nous avons cherché la matrice de similarité cosinus que nous avons noméme similarity_df.Finalement nous avons effectuer le Kmeans clustening en exploitant les labels donnés avec la métrique cosine.Ci-dessous les deux figures obtenues avec KMeans en exploitant les deux métriques.

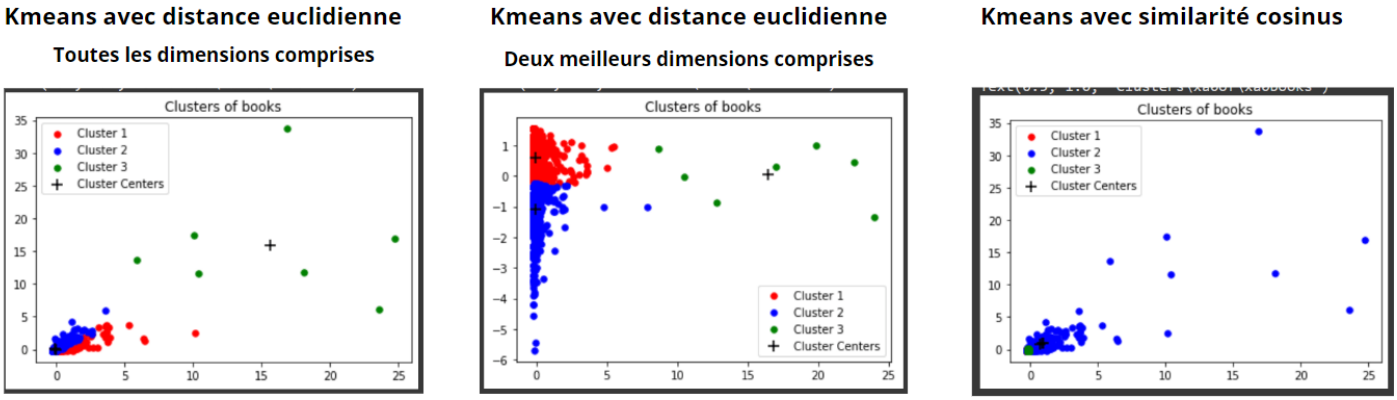

Figure 6: KMeans Clustering avec deux métriques différentes

Nous remarquons que le KMeans appliqué sur notre matrice moyennant la métrique **euclidienne** donne des clusters plus définis qu'avec la métrique **cosine**.Notre clustering avec la distance euclidienne est raffiné davantage lorsque nous selectionnons deux dimensions.

4.3 Question 3

En utilisant l'analyse en composante principale, projectez vos segments obtenus suivant les deux premiers composantes principales. De manière visuelle, les segments sont-ils différents avec la distance euclidienne par rapport aux segments obtenus avec la similarité cosinus?

Pour cette question nous avons effectuer la projection en composantes principales sur toute notre matrice de données X. Par la suite nous avons récupéré le résultat et effectué le clustering par KMeans directement vu que celui ci adopte par défaut la distance euclidienne comme métrique.En sortie nous avons obtenu la figure à gauche.Nous remarquons que le résultat après PCA est presque semblable au résultat du KMeans avant PCA pour les dimensions que nous avons choisi(Nombre de votes et moyenne).Ceci indique que ces deux dimensions sont réellement les meilleurs pour modeliser et visualiser les clusters. En ce qui concerne le résultat du clustering sur toute la matrice X, nous remarquons que l'analyse par composantes principales donne des résultats plus clairs.

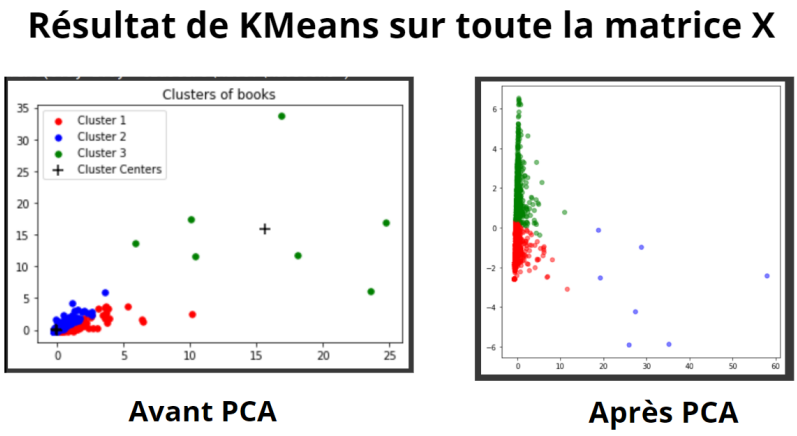

Figure 7: KMeans Clustering avec distance euclidienne avant et après PCA

Nous avons également effectuer une étude comparative entre le KMeans qui adopte la métrique cosine avant et après PCA.Nous avons obtenu les résultats suivants:

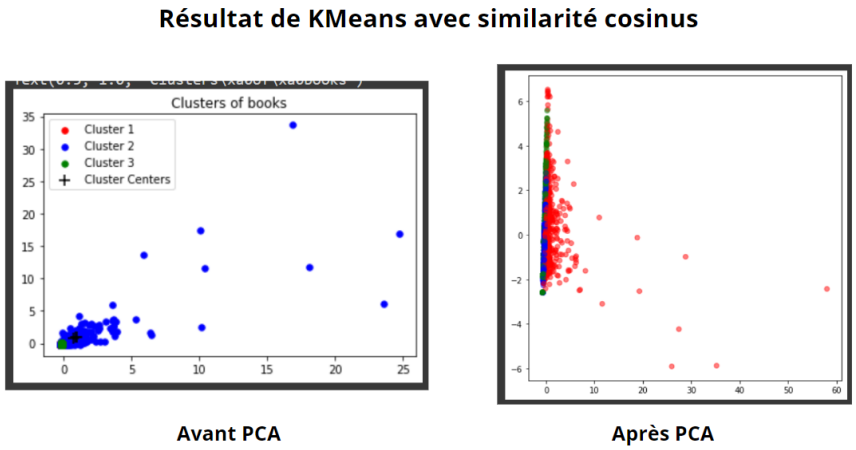

Figure 8: KMeans Clustering avec deux métriques différentes après PCA

Pour résumer voici les deux figures du clustering par KMeans apèrs PCA en utilisant deux métriques différentes.On remarque que meme après avoir effectuer la PCA, les résultats du clutering KMeans avec la distance euclidienne demeurent plus distingués que ceux obtenus avec la similarité cosinus.

Néomoins, nous allons revérifier nos résultats ultérieurement en effectuant une analyse par score silouaitte et info-mutuelle pour pouvoir choisir la meilleure.

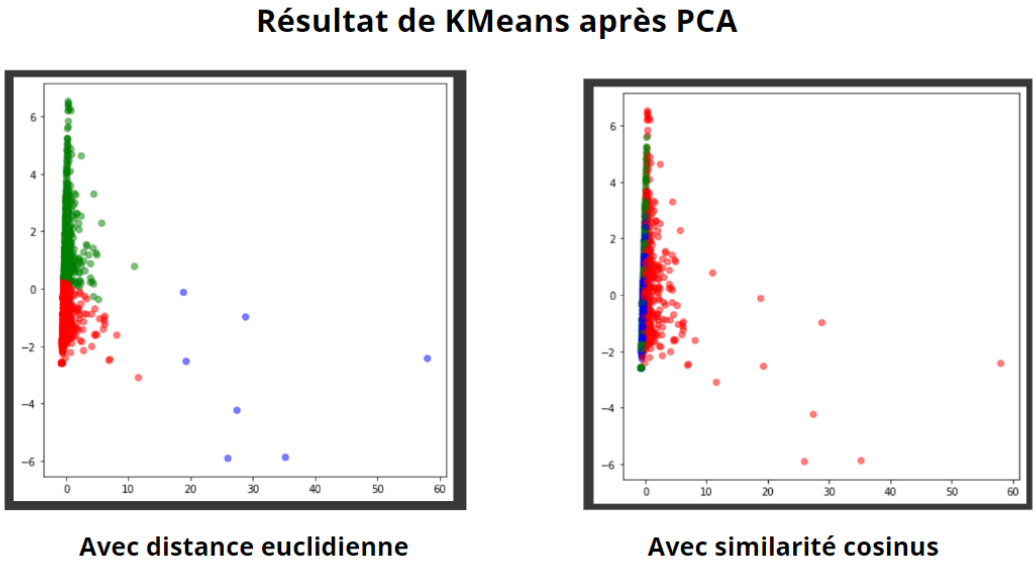

Figure 9: KMeans Clustering avec deux métriques différentes après PCA

4.4 Question 4

Effectuer une segmentation par le clustering spectral sur deux dimensions en prenant k = 3 avec pour distances euclidienne et similaruté cosinus

Comme suggéré par notre professeur de cours, nous allons procéder dans cette question par deux méthodes différentes.Nous allons répondre à la question directement puis dans une seconde étape nous allons effectuer la PCA et par la suite répondre à la question.Nous allons finalement comparer et analyser les résultats de cette question par les deux méthodes.

Pour la première méthode avec distance euclidienne, nous avons selectionner les deux premières dimensions de la matrice P, par la suite nous avons effectuer la décomposition en valeurs singulières.Le résultat de la décomposition donne la matrice P et D et par conséquent nous avons trouver M.
Pour la métrique cosine similarity nous avons procéder de la meme manière tout en exploitant la matrice de similarité au cours de la décomposition.

Pour la seconde méthode avec PCA, nous avons projeté notre matrice X moyennant la PCA puis exploiter le résultat en sortie pour effectuer la décomposition en valeurs singulières pour les deux types de métriques.Finalement nous avons effectuer le KMeans sur la matrice $P\sqrt{D}$.

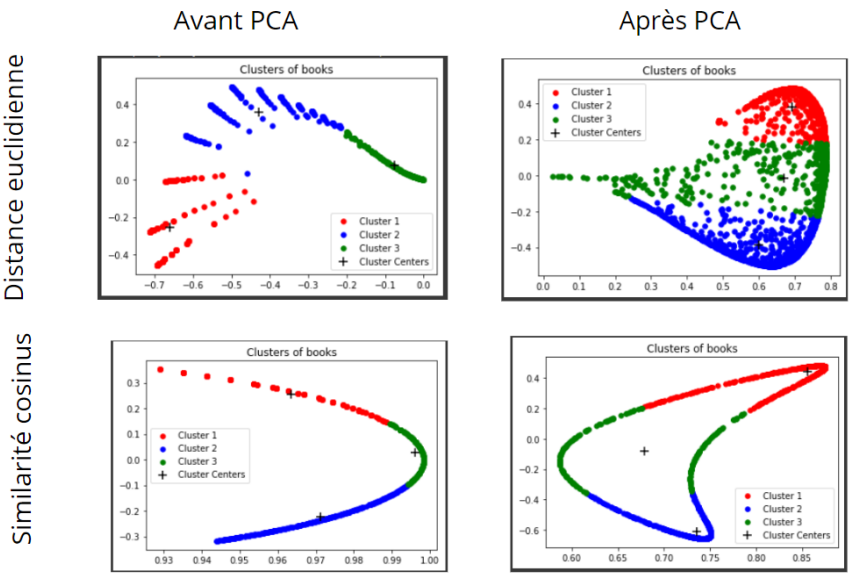

Figure 10: KMeans Clustering avec deux métriques différentes après PCA

Le clustering spectral est flexible et nous permet également de regrouper des données non graphiques. Il ne fait aucune hypothèse sur la forme des clusters. Les techniques de clustering, comme K-Means, supposent que les points affectés à un cluster sont sphériques autour du centre du cluster.On remarque que dans de tels cas, le clustering spectral aide à créer des clusters plus précis. Il peut regrouper correctement les observations qui appartiennent en fait au même cluster, mais qui sont plus éloignées que les observations d'autres clusters, en raison de la réduction de dimension.

4.5 Question 5

Utilisez les métriques silhouette et information mutuelle pour évaluer la qualité de votre segmentation (dans le cas des k-moyennes et le cas spectral) lorsque vous utilisez dans un premier temps une distance euclidienne et ensuite dans un deuxième temps la similarité cosinus.

Nous avons évaluer la qualité de nos algorithmes de clustering moyennant le score silhouette et l'information mutuelle.Nous avant tester la qualité des résultats avant et après PCA.Nous pouvons remarquer que la projection en composantes principales améliore les résultats du clustering, et ceci est causé par le fait que la projection choisi les meilleurs dimensions.De plus, nous remarquons que le KMeans donne de meilleurs résultats avec la distance euclidienne après PCA alors que le clustering spectrale donne de meilleiurs résultats avec la similarité cosinus après PCA.

Indices de qualité avant PCA

	Distance euclidienne_Siloutte	Distance euclidienne_IonfoMut	Similarité cosinus_Silouette	Similarité cosinus_InfoMut
Kmeans	0.911435	0.916617	0.468143	0.533456
Spectral	0.568217	0.016876	0.781868	0.008170

Indices de qualité après PCA

	Distance euclidienne_Siloutte	Distance euclidienne_IonfoMut	Similarité cosinus_Silouette	Similarité cosinus_InfoMut
Kmeans	0.931563	0.249842	0.467456	0.525214
Spectral	0.524094	0.132681	0.901090	0.160042

Figure 11: Indicateurs de qualité pour clustering

Partie B

4.6 Questin 1

Selon vous quel serait le risque de prendre aléatoirement un sous-ensemble de données pour effectuer les taches a-1) à a-5) ?
Le fait de sélectionner aléatoirement les quantités de données dans chacune des catégories ne donne pas des résultats significatifs.En effet, il se peut que les données d'une catégorie donnée prédominent sur l'échantillon et par conséquent le résultat du clustering sera biaisé. La subdivision des donn ees en des sous-ensembles pourraient generer des cas de figure favorables/defavorables.Pour s'assurer qu'on ne favorise pas un cas en particulier, on pourrait generer à partir de nos donnees, plusieurs sous-ensemble entrainements et tests.La meilleure façon pour éviter ce problème est d'exploiter les méthodes de stratification et de K-cross validation.

4.7 Questin 2

En procédant par une sélection stratifiée, on voudrait s'assurer que toutes les catégories soient représentées dans notre sous-ensemble. Pour chacune des trois catégories 1 2, 3 et 4 5, sélectionnez aléatoirement des quantités p1, p2, p3, p4 et p5 de données tels que p1 = p2 = p3 = p4 = p5 et p = p1 + p2 + p3 + p4 + p5 et reconstruire votre matrice X.

Nous avons construit un DataFrame qui comptabilise le nombre de fois qu'un livre est apprécié dans une catégorie. Ensuite nous avons pris le maximum sur le nombre de vote dans une appréciation et associé par la suite la catégorie correspondante (tel que dans un vote de majorité).Vu que nous avons initialement commencer à un nombre de p livres qui correspondrait à une proportion de x% de la totalité des livres,nous avons choisit dans chacune des catégories générales x% de livres de tel sorte que ce x% soit le minimum des proportions. Nous avons obtenu notre nouvelle matrice X laquelle nous allons exploiter par la suite et qui respecte la règle p1=p2=p3=p.

	Apprécié	Neutre	Moins_Apprécié	Nombre_Votes	(overall, mean)	(overall, std)	(overall, median)	Catégorie
000100039X	188.0	8.0	10.0	206.0	4.674757	0.875712	5.0	Apprécié
0001055178	12.0	2.0	4.0	18.0	3.555556	0.983524	4.0	Apprécié
0001473123	15.0	0.0	1.0	16.0	4.625	1.024695	5.0	Apprécié
0001473727	7.0	0.0	0.0	7.0	5.0	0.0	5.0	Apprécié
0001473905	5.0	1.0	0.0	6.0	4.666667	0.816497	5.0	Apprécié
...	...	...	...	...	...	...	...	...
0060975504	26.0	5.0	2.0	33.0	4.151515	1.093195	4.0	Apprécié
0060975547	13.0	2.0	1.0	16.0	4.25	0.930949	4.5	Apprécié
0060975598	15.0	3.0	0.0	18.0	4.388889	0.777544	5.0	Apprécié
0060975768	208.0	6.0	6.0	220.0	4.668182	0.742627	5.0	Apprécié
0060975776	62.0	1.0	8.0	71.0	4.338028	1.170393	5.0	Apprécié

5894 rows × 8 columns

Figure 12: Matrice après stratification

	Apprécié	Neutre	Moins_Apprécié	Nombre_Votes	(overall, mean)	(overall, std)	(overall, median)
006052149X	9.0	2.0	2.0	13.0	4.076923	1.38212	5.0
0060745312	7.0	0.0	0.0	7.0	4.714286	0.48795	5.0
0060834390	7.0	2.0	4.0	13.0	3.384615	1.660244	4.0
0060611391	13.0	3.0	0.0	16.0	4.625	0.806226	5.0
0028639510	3.0	1.0	2.0	6.0	3.333333	1.632993	3.5
...	...	...	...	...	...	...	...
0007328230	3.0	4.0	0.0	7.0	3.571429	0.786796	3.0
0060882190	3.0	4.0	0.0	7.0	3.714286	0.95119	3.0
0060937173	1.0	3.0	2.0	6.0	2.666667	1.032796	3.0
0060781610	2.0	3.0	1.0	6.0	3.5	1.224745	3.0
0060599316	3.0	4.0	3.0	10.0	3.1	1.286684	3.0

165 rows × 7 columns

Figure 13: Ma trtice de regroupement par catégorie

4.8 Questin 3

Refaire les étapes a-2) 'a a-5). Faite une comparaison des résultats reportés dans votre tableau à ceux reportés dans le tableau obtenu en a-5).

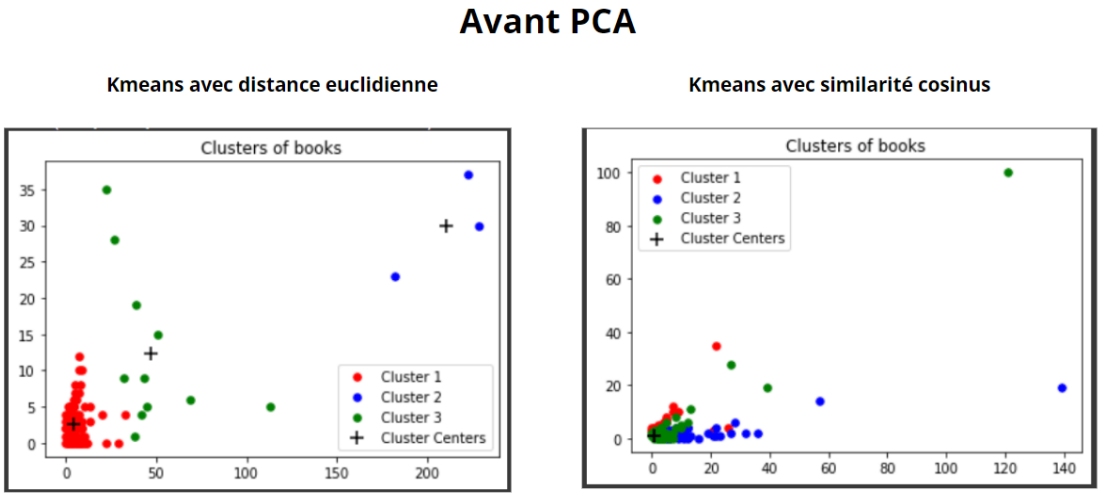

Figure 14: KMeans par deux métriques après stratification

Lorsque nous avons refait le KMeans sur la nouvelle matrice X stratifiée.Nous remarquons que la métrique cosinus donne de meilleurs résultats avec cette matrice qu'avec celle utilisée à la question 2.On peut maintenant visualisés les trois clusters qui sont plus au moins distincts.Pour la métrique euclidienne nous remarquons peu de différence par rapport à la question 2.Meme si nous avons fait une stratification nous nous pouvons point visualiser des clusters équilibrés.

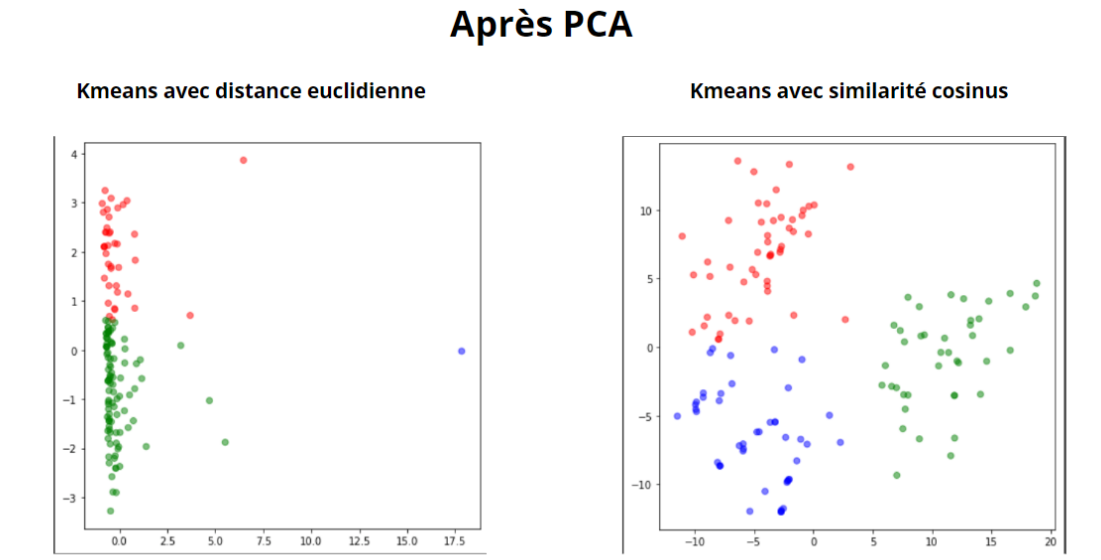

Figure 15: KMeans avec deux métriques différentes après PCA et après stratification

Comme à la question 3,encore une fois la projection en composantes principales a permis de mieux visualiser nos clusters.Cette fois ci la métrique cosinus permet de donner une meilleur séparation que l'euclidienne.
Nous soulignons le désiquilibre des clusters visualisés avec la distance euclidienne malgré l'exploitation de la méthode de stratification,contrairement à la métrique cosinus qui a permis de garder l'équilibre.

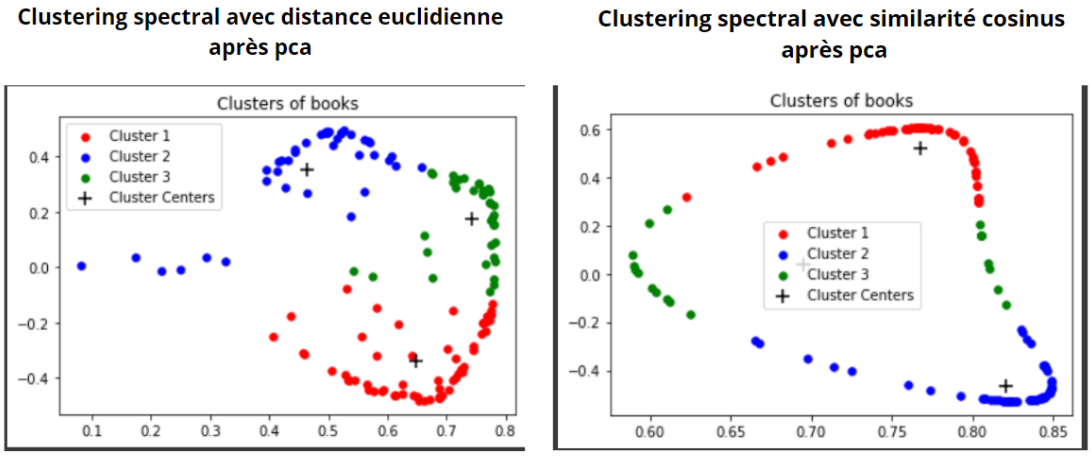

Figure 16: Clustering spectral avec deux métriques différentes après PCA et après stratification

Pour le clustering spectral,les résultats avant et après stratification sont presque semblables.

Indices de qualité du clustering après stratification

	Distance euclidienne_Siloutte_strat	Distance euclidienne_IonfoMut_strat	Similarité cosinus_Silouette_strat	Similarité cosinus_InfoMut_strat
Kmeans	0.075062	0.064631	0.555640	0.632697
Spectral	0.566485	0.395853	0.858186	0.355429

Figure 17: Indices de qualité après stratification

Finalement,nous avons obtenu les mesures de qualité ci-dessus.Nous remarquons que les valeurs ont diminués, mais nousnous pouvons point juger par rapport à l'efficacité de la stratification vu que le nombre de données utilisées est négligeable par rapport à celui utilisé avant stratification.

4.9 Questin 4

A supposer qu'on associe à chacune des catégories 1 2, 3 et 4 5 les étiquettes respectives l1, l2 et l3. On voudrait retrouver les étiquettes des données restantes (celles qui n'ont pas eté prises en considération lors de la segmentation). Quelle stratégie pensez-vous utiliser, expliquer ?

Nous proposons de construire une équation générée par un modèle qui prend en entrée les features et qui trouve une combinaison adéquate pour retrouver les clusters auquels appartiennent nos livrers.Nous allons par la suite exploiter cette équation sur un ensemble de test pour trouver les étiquettes de données restantes.

References

[1] :http://jmcauley.ucsd.edu/data/amazon/links.html.

[2] :https://docs.python.org/3/.

[3] : https://pandas.pydata.org/docs/.

[4] :https://www.geeksforgeeks.org/box-plot-in-python-using-matplotlib/.

[5] :https://sites.google.com/view/aide-python/statistiques/machine-learning-en-python/analyses-en-composantes-principales.

[6] :https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette$_s core.html$.

[7] :https://mrmint.fr/algorithme-k-means.

[8] :https://stackoverflow.com/questions/5529625/is-it-possible-to-specify-your-own-distance-function-using-scikit-learn-k-means.

[9] :https://scikit-learn.org/stable/modules/clustering.htmlclustering-performance-evaluation.

[10] :https://scikit-learn.org/stable/modules/generated/sklearn.model$_s election.StratifiedKFold.html$.

[11] : https://www.youtube.com/watch?v=YHz0PHcuJnkt=1048s.