IFT3700 Travail 2

December 23, 2021

Steve Levesque

Weiyue Cai

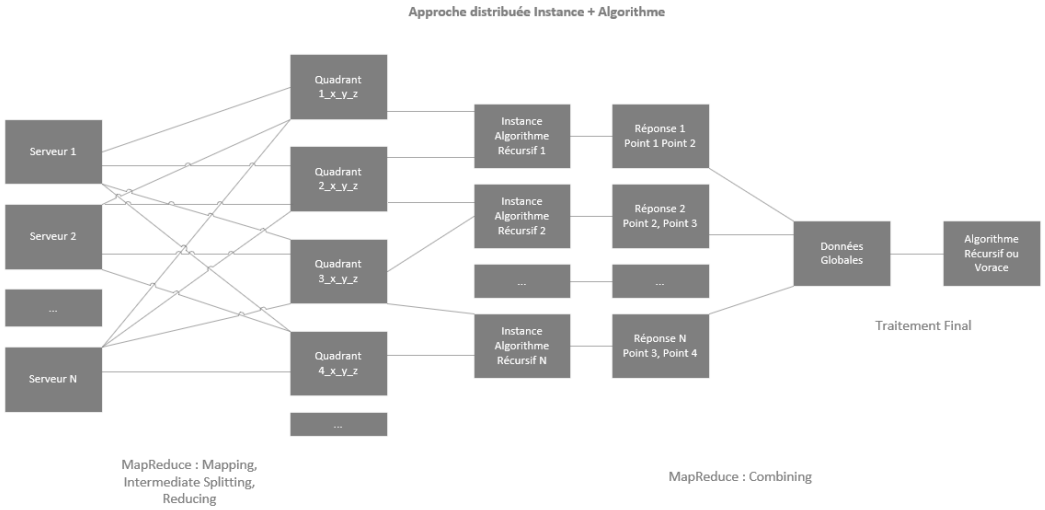
1 Problème 1

1.1 A

Il est demandé d'utiliser une approche distribuée pour trouver les points les plus proches. Il est possible de respecter cette demande pour toutes les étapes du processus. C'est-à-dire que la collecte des données pour l'analyse peut être faite grâce à MapReduce et la recherche des deux points les plus rapprochés est possible grâce à une version de l'algorithme diviser pour régner.

NB : Il est très pertinent d'utiliser l'approche distribuée de l'algorithme au lieu de naïvement l'appliquer à la fin puisque la division des données est un état de départ de notre tâche. Donc, nous pouvons l'utiliser à notre avantage au lieu de le voir comme un problème. (L'algorithme diviser pour régner donné en lien dans l'énoncé du devoir.

Voici une vue graphique, les explications détaillées suivent plus bas :



1. Au départ, les serveurs avec les données distribuées doivent être rassemblées à un certain point pour pouvoir appliquer l'algorithme et trouver les 1000 points les plus proches. Il est possible d'utiliser MapReduce pour cela en fonction d'une partie d'un quadrant pour avoir les données regroupées par rapport à leur position.

2. Nous allons utiliser l'algorithme version diviser pour régner sur chaque partie de quadrant, plus précisément les 2 partitions proches l'une de l'autre et chaîner du même ordre d'idée sur les autres instances (i.e. partie 1 et 2 proche l'une de l'autre avec l'algorithme, et ensuite utiliser 2 et 3, etc.). On veut utiliser une instance (ici la 2ème dans l'exemple) 2 fois ou plus puisqu'on est en 3 dimensions pour ne pas manquer la possibilité d'avoir une paire très proche, mais qui serait dans 2 partitions différentes.
3. On collecte les réponses et on les combine pour avoir un jeu de données global. Les paires comme 1 et 2 et 2 et 3 utilisent une donnée commune (2). MapReduce n'a pas la tâche de fusionner/supprimer cet ensemble parce que c'est de l'ordre de l'algorithme de faire une évaluation finale sur ce cas spécial qui est produit par le partage des partitions pour éviter de manquer des paires et des réponses potentielles.
4. L'algorithme est appliqué une dernière fois sur le jeu de données pour évaluer si c'est bien 1 et 2 ou 2 et 3 la paire la plus proche. Ici, il est arbitraire (à notre avis) d'utiliser l'algorithme récursif ou fouille exhaustive rendu à cette étape finale.

1.2 B

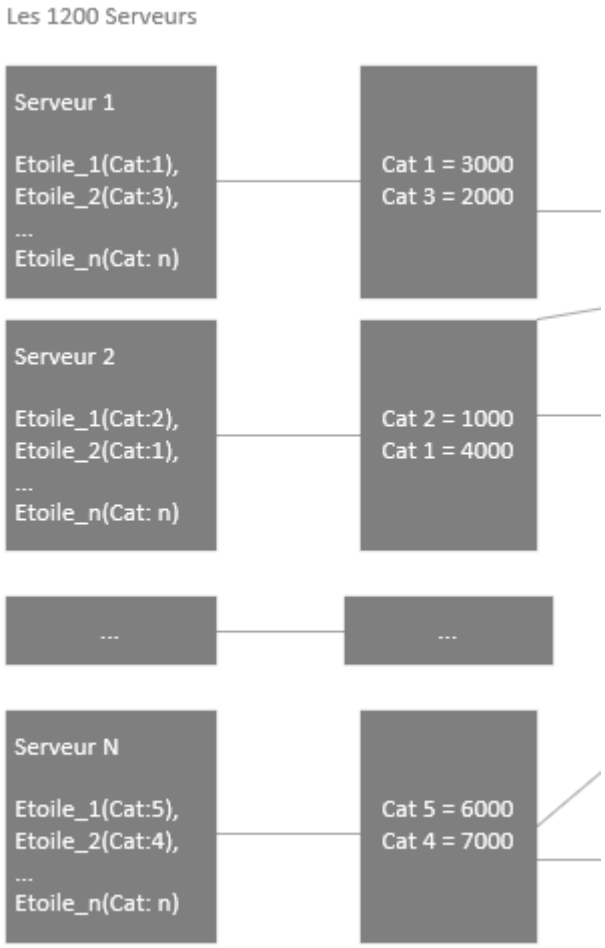
Il est possible d'utiliser la même approche que précédemment (A.) et de changer l'algorithme utilisé pour tout simplement répondre à notre tâche.

Nous allons en premier lieu avec MapReduce commencer à l'étape de MAP (mapping), puisque nos données sont sur des serveurs répartis à des endroits différents.

Ensuite, nous pouvons finir avec REDUCE. Cette étape permet de rassembler nos données bien rassemblées pour les réduire à notre résultat voulu ou un ensemble relativement propre et facile à traiter dans des temps raisonnables.

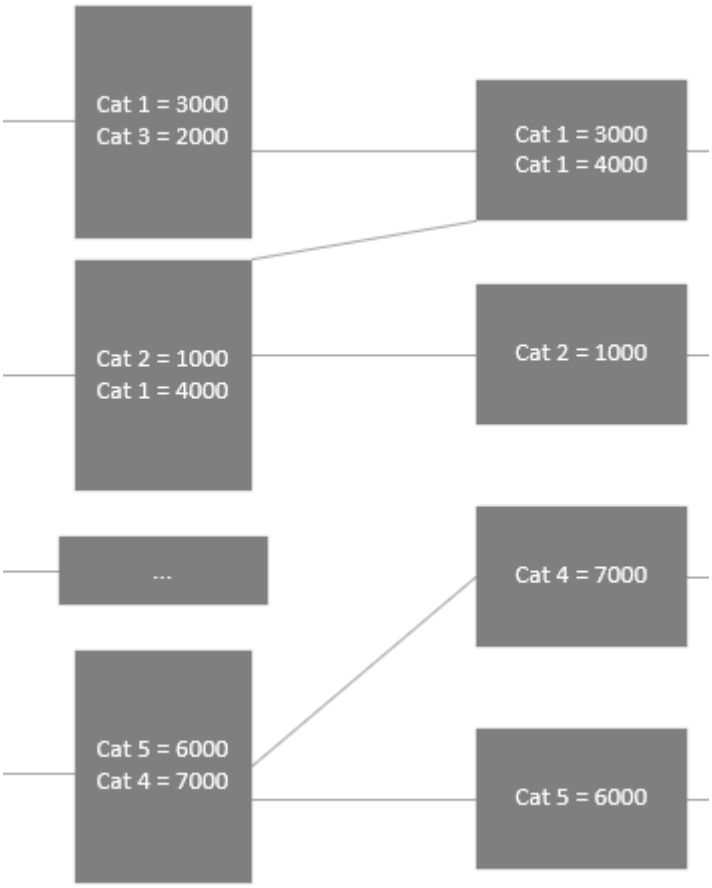
Map et Reduce comportent elles aussi des sous-étapes : Splitting, Mapping, Intermediate Splitting, Reducing et Combining. Elles sont décrites ci-dessous avec la vue globale pour compléter. (Splitting est esquivé dans ce problème puisque de base nous commençons avec des serveurs distincts.)

Mapping : Les données sont selon l'énoncé équitablement répartis entre les serveurs, donc on peut partir de ce fait et commencer à catégoriser les compteurs par rapport à chaque catégorie d'étoiles disponibles sur le serveur en question (la métrique demandé dans la question, cela pourrait être n'importe quoi d'autre) et chaque compteur sera incrémenté s'il y a une donnée d'étoile faisant partie de cette catégorie.



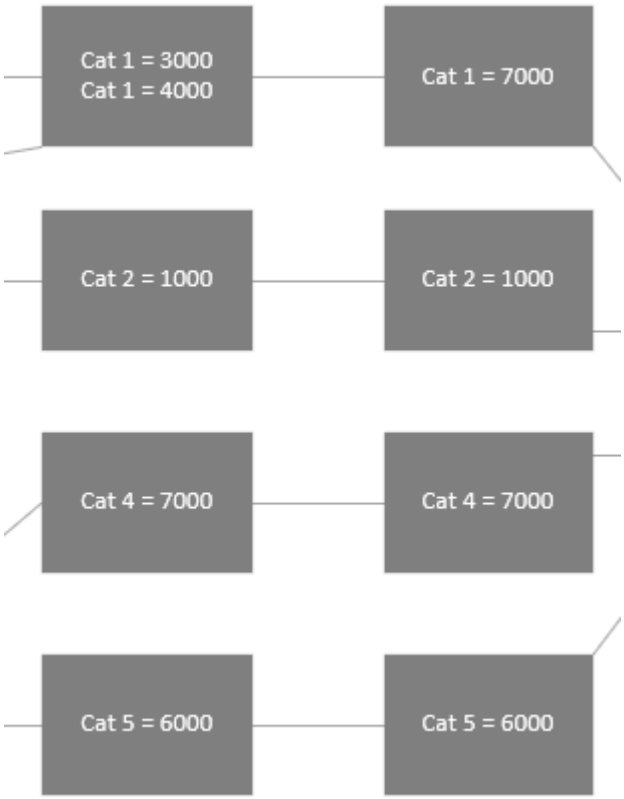
Intermediate Splitting : Cette étape peut être facultative, mais elle est toujours utile à faire pour avoir une réduction propre, concise et efficace. Celle-ci consiste à rassembler nos compteurs en catégories respectives.

Puisque l'étape précédente prend en compte les données des serveurs dans une portée locale et que celles-ci sont balancées de manières probablement non biaisées à une catégorie en particulié, il est fortement possible qu'un compteur ne comporte pas le nombre total pour une catégorie (Ou un "feature" plus généralement). Il faut donc rassembler le tout.

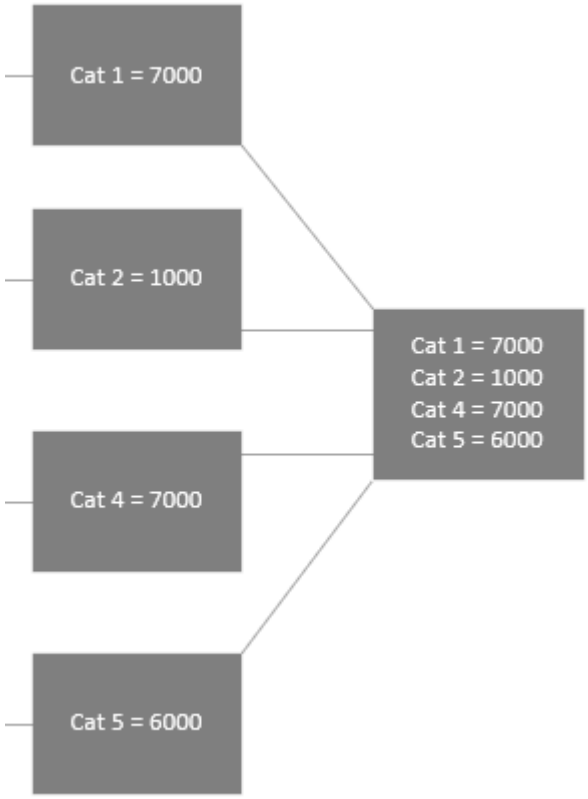


Reducing : Avec le rassemblement de certaines (pour ne pas dire toutes les instances) instances, il sera possible d'obtenir tous les compteurs des serveurs pour une catégorie d'étoile au même endroit.

Il est possible de faire d'autres choses à cette étape comme du prétraitement, mais pour le but de la question nous pouvons passer à la prochaine étape.

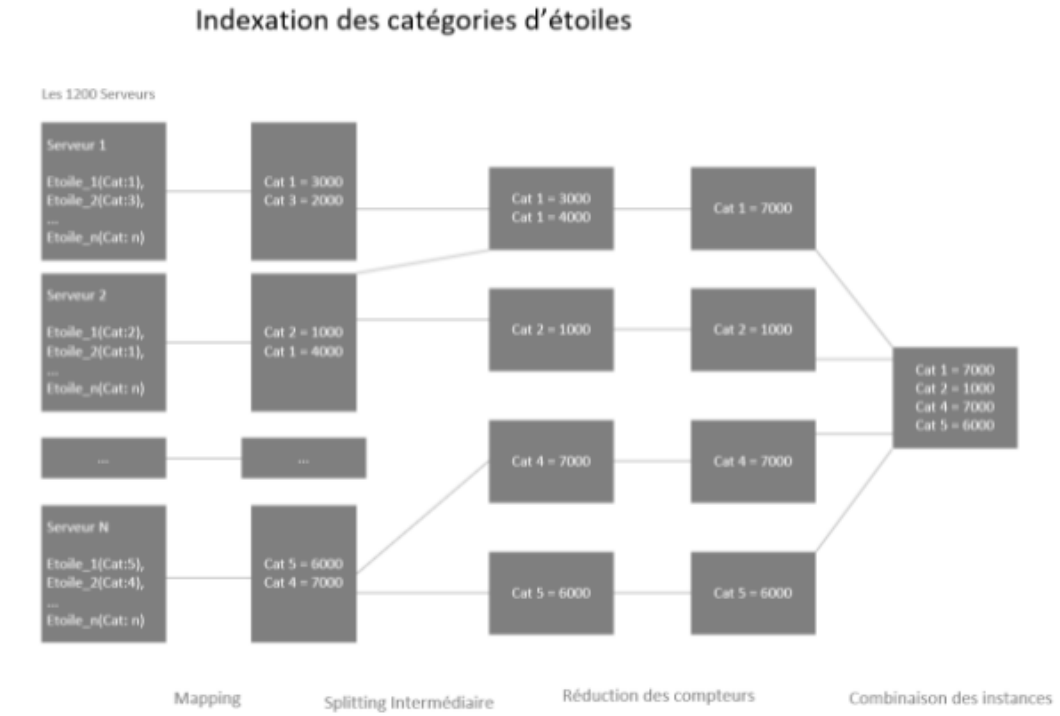


Combining : Cette étape (finale) permet de rassembler et réduire nos données bien traitées/pertinentes, mais en plusieurs instances dans la même.



Résultat Final :

Résultat Final :



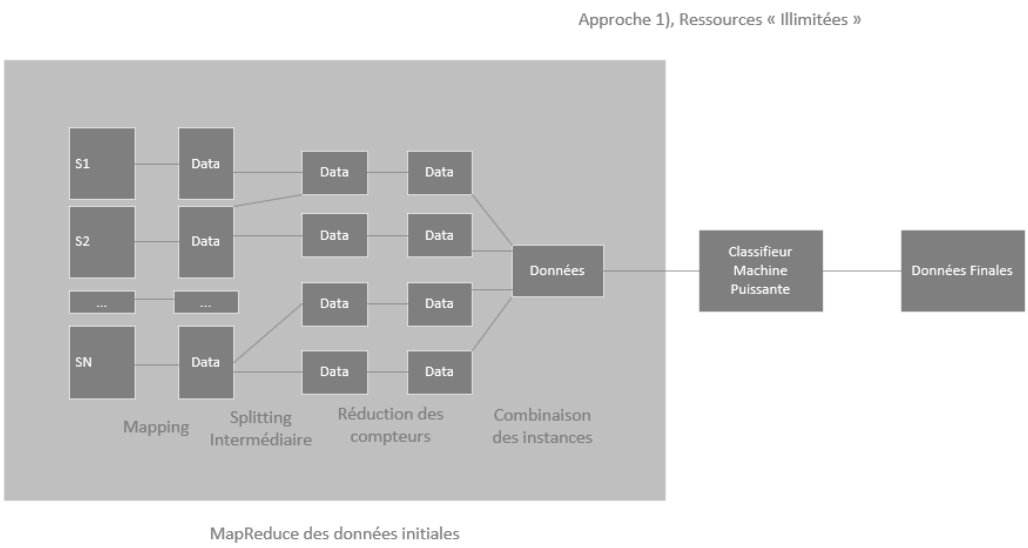
1.3 C

Le classifieur sera basé sur une réduction de dimension avec PCA/T-SNE et son application/entraînement se fera sur les connaissances suivantes :

1. Ressources illimitées : sur le rassemblement final des données, puisque la notion de distance pourra être globalement effective sur l'entraînement et l'utilisation du classifieur. MapReduce pourra être utilisé à priori, mais la machine qui reçoit le résultat (output) aura la responsabilité d'entraîner et d'utiliser le classifieur. (potentiellement une machine avec 1200 * 128 GB RAM + 1200 * 4 TB)
2. Ressources très restreintes : l'utilisation de MapReduce sera mise de l'avant après avoir appliquer le classifieur sur chaque serveur/instance par rapport à l'entraînement et l'utilisation. Cela permet d'utiliser une machine seulement pour rassembler les résultats et d'y appliquer ensuite une métrique rapide et peu coûteuse.
3. Intermédiaire : Il est possible d'utiliser les techniques en A. et en B., les jumeler et créer un modèle qui est dans la moyenne en matière de performance de classification ainsi qu'en utilisation de ressources.

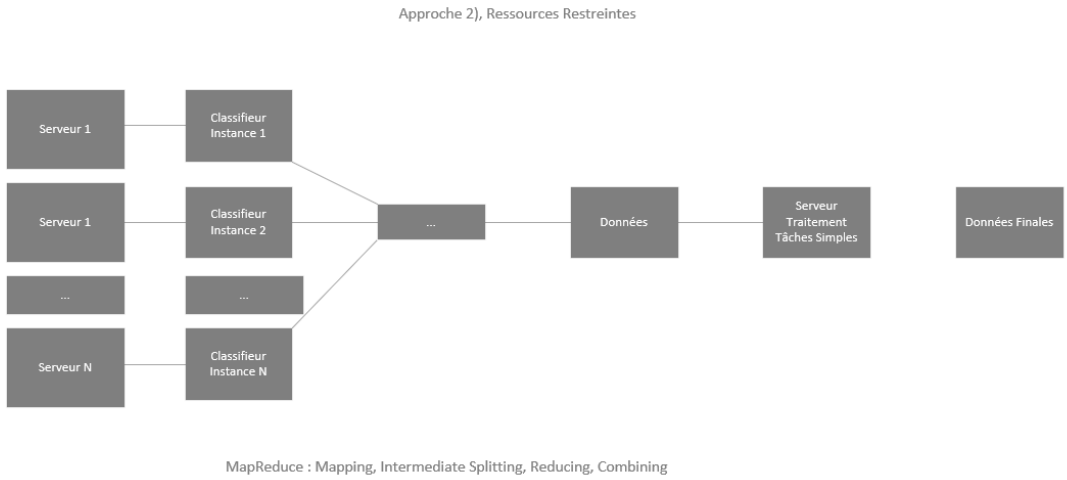
Voici plus de précision sur les 2 approches :

1.



MapReduce des données initiales

2.



MapReduce : Mapping, Intermediate Splitting, Reducing, Combining

2 Problème 2

Cette partie du rapport contient toutes les informations nécessaires lorsque la donnée à remettre pour une question en lien ne nécessite pas un fichier csv ou json ou qu'une précision est nécessaire.

Le tout est séparé par numéro de question.

2.1 1. Collecte des données

Les 40 fichiers pour les 40 colonnes se trouvent dans data/raws_from_wiki.

Voici le site utilisé pour compiler les .csv https://wikitable2csv.ggor.de/.

Il y a eu du prétraitement manuel lorsque les cas de modifications étaient très particuliers, où créer une automatisation pour cela aurait été trop complexe pour un gain peu pertinent.

Il est possible de créer/compiler dataA.csv à partir du code à chaque fois, il n'y a pas de données changées à la main après avoir lancé le code (juste avant pour la motivation ci-haut, i.e. dans les 40 .csv

Code utilisé : main.py, test_main.py et utils.py

2.2 2. Corrélations

Rien à complémenter en particulier.

Code utilisé : q2.py et utils.py

2.3 3 Prédictions avec classifieur bayésien et régression linéaire

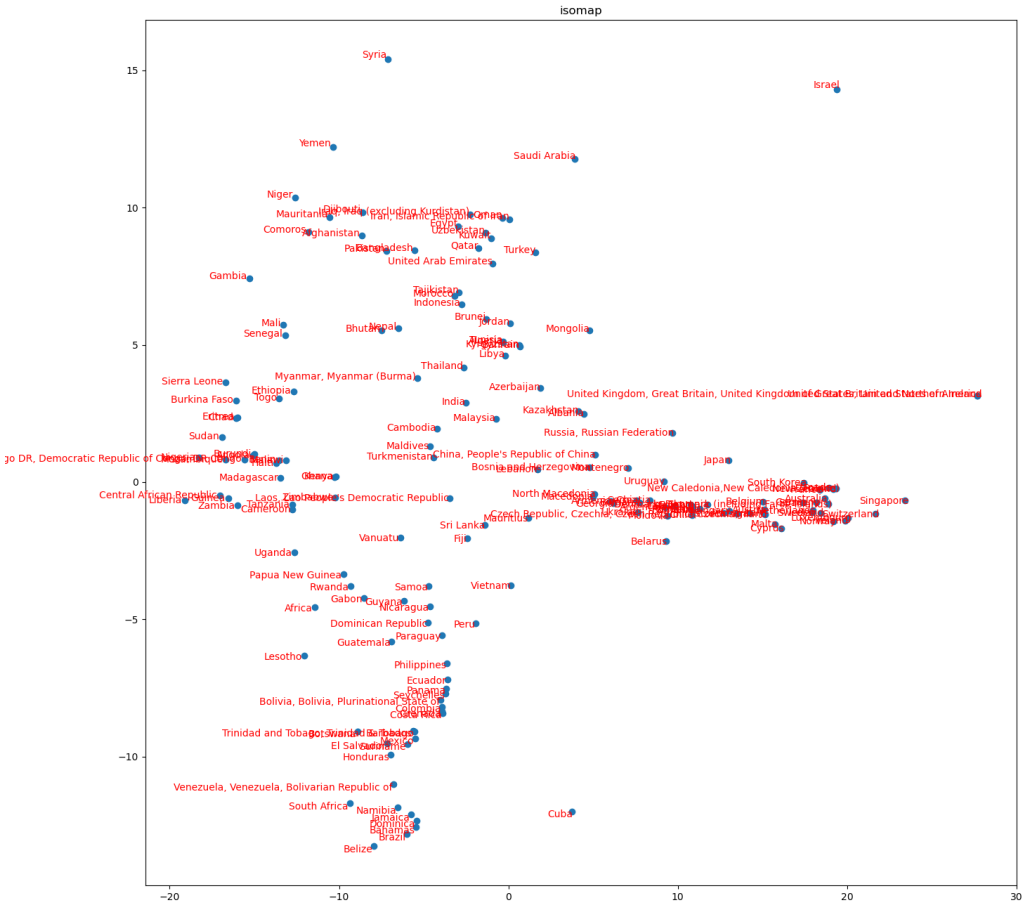
2.3.1 3.a Pour chaque colonne, analyser la qualité des prédicteurs en fonction des autres colonnes.

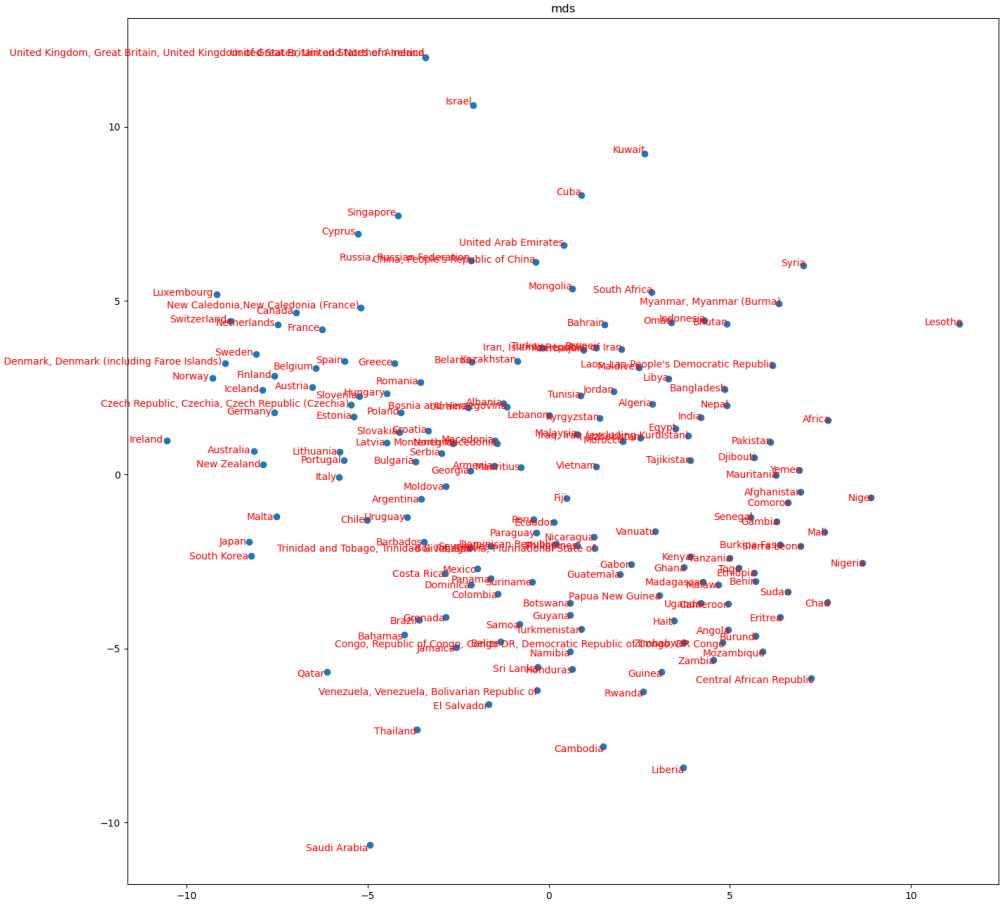
Le score R^2 de la régression linéaire qui prédit chaque colonne en fonction des autres colonnes. ('Human Development Index', 0.9755929638330451), ('GDP', 0.9598666283028567), ('Median Age', 0.9581167382824838), ('Life Expectancy', 0.9539162298649594), ('Minimal Wage', 0.9394451456623399), ('Population Growth Rate', 0.9376037238233875), ('Under-Five Mortality', 0.9269335196769433), ('Num. of Scientific and Technical Journals Articles', 0.9230858730331675), (('Fertility Rate', 0.9225249131605183), (('Oil Production', 0.9196795197117154), ('Importance of Religion', 0.9138721983904726), ('Military Expenditures', 0.9102375022547553), (('Age at First Marriage', 0.9046340518056656), ('Democracy Index', 0.9028873857642519), ('Literacy Rate', 0.9014167882330102), ('Internet Users', 0.8890859237227736), ('Obesity Rate', 0.8720399410621537), ('Meat Consumption per kg', 0.8329321069523623), ('Kilocalories', 0.8312289818439271), ('Economic Freedom Score', 0.8212929222032962), ('Tertiary Education', 0.8190844335680582), ('Health Expenditure', 0.8163214838651691), ('Consumption of Pure Alcohol', 0.7946467516423652), ('Internet Speed', 0.7798604361905022), ('Milk Consumption', 0.7751426748727382), ('Gini ('Incarceration Rate', 0.7705026380081861), ('Homeless Population', 0.7686003611148013), ('Avg. Yearly Temperature in Celsius', 0.7605173730862499), ('Age of Criminal Responsibility', 0.75836738773577), (('Intentional Homicide Victims', 0.7377686895345912), ('Suicide Rate', 0.7375920855909104), ('Books Published', 0.7045569504191674), ('Spending on Education', 0.6824180891577037), ('External Debt ((('Tabacco Consumption', 0.5826049091730161)]Le score "mean accuracy" du classificateur bayésien qui prédit chaque colonne en fonction des autres colonnes. ('Human Development Index', 0.9526627218934911), ('Median Age', 0.9230769230769231), ('Under-Five Mortality', 0.9171597633136095), ('Health Expenditure', 0.8994082840236687), ('GDP', 0.8816568047337278), ('Internet Users', 0.8816568047337278), ('Num. of Scientific and Technical Journals Articles', 0.8757396449704142), ('Fertility Rate', 0.8698224852071006), ('Life Expectancy', 0.8698224852071006), ('Meat Consumption per kg', 0.8579881656804734), ('Minimal Wage', 0.834319526627219), ('Age at First Marriage', 0.8224852071005917), ('Kilocalories', 0.8165680473372781), ('Economic Freedom Score', 0.8047337278106509), ('Milk Consumption', 0.8047337278106509), ('Population Growth Rate', 0.7988165680473372), ('External Debt ((('Importance of Religion', 0.7692307692307693), (('Literacy Rate', 0.757396449704142), ('Obesity Rate', 0.7396449704142012), ('Consumption of Pure Alcohol', 0.7337278106508875), ('Democracy Index', 0.7337278106508875), ('Avg. Yearly Temperature in Celsius', 0.727810650887574), ('Military Expenditures', 0.7218934911242604), ('Tertiary Education', 0.7218934911242604), ('Internet Speed', 0.7159763313609467), (('Intentional Homicide Victims', 0.7041420118343196), ('Gini (('Books Published', 0.6923076923076923), ('Age

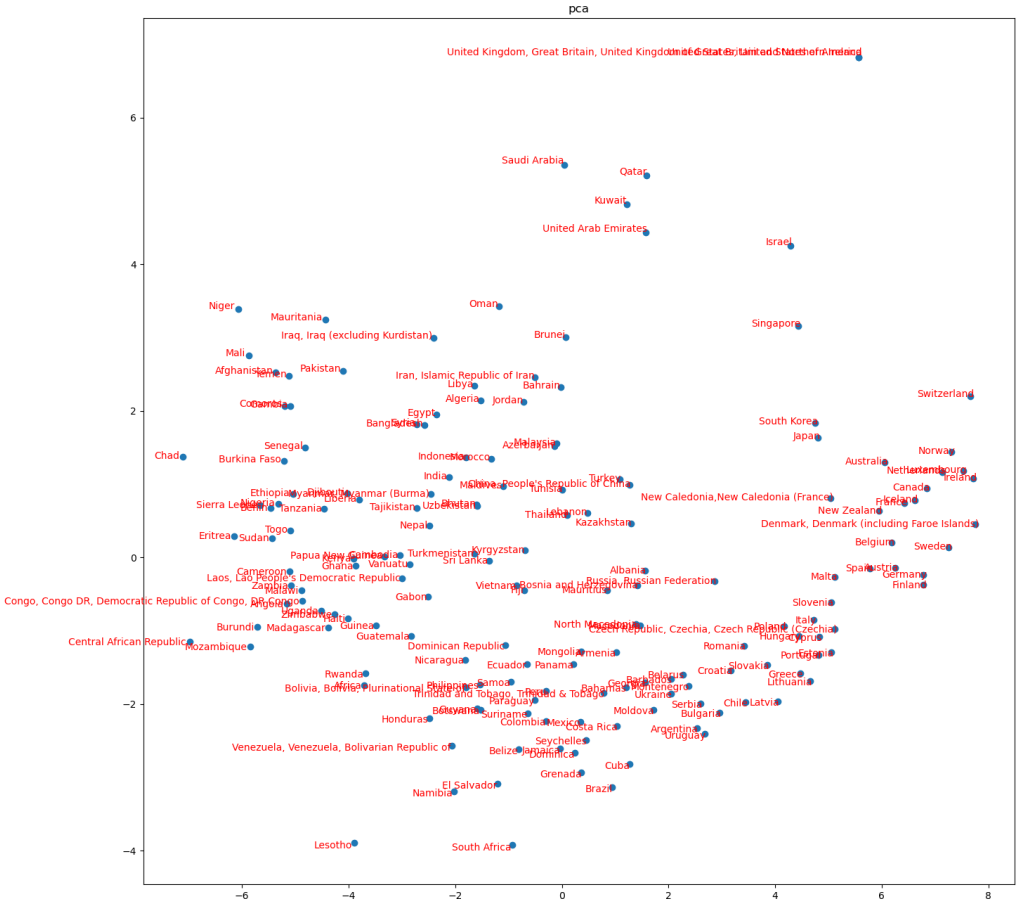
of Criminal Responsibility', 0.6863905325443787), ('Homeless Population', 0.6804733727810651), ('Tabacco Consumption', 0.6745562130177515), ('Suicide Rate', 0.6568047337278107), ('Oil Production', 0.6449704142011834), (''Incarceration Rate', 0.621301775147929), ('Spending on Education', 0.621301775147929) Code utilisé : q3.py et utils.py

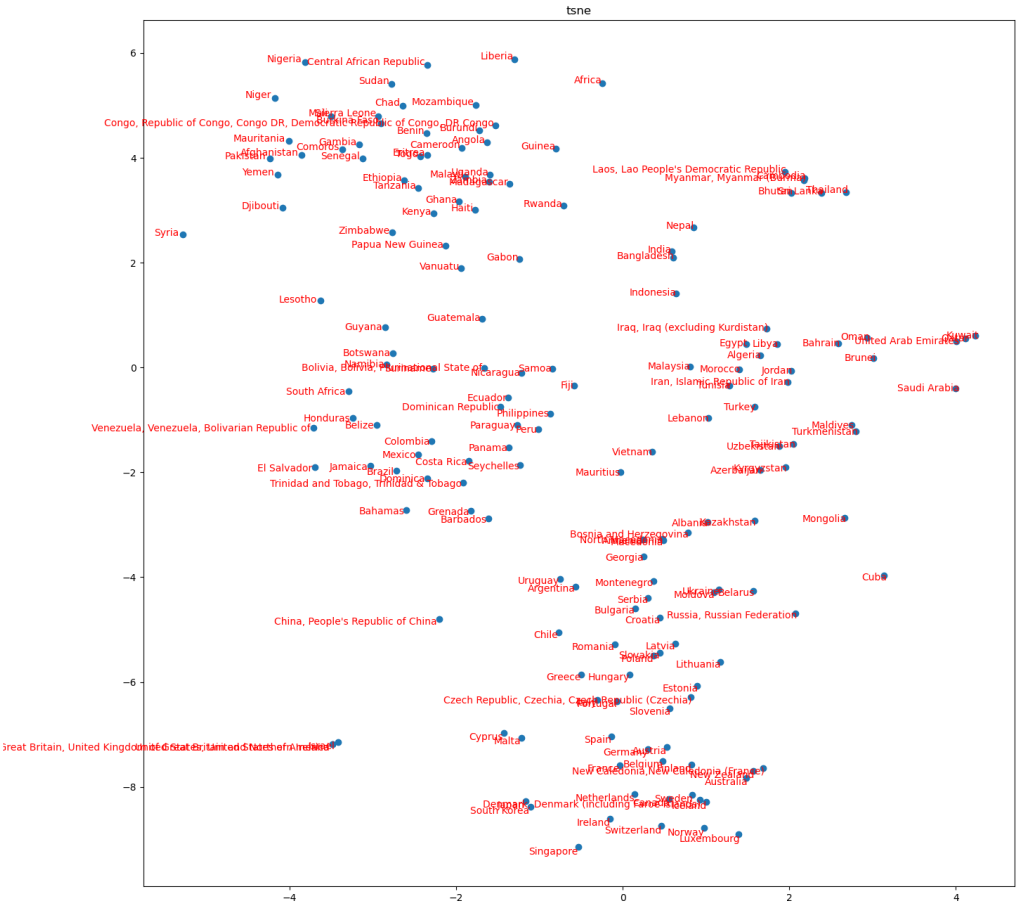
2.4 4. Visualisation et représentations

2.4.1 4.a Appliquer un algorithme de réduction de dimensionnalité pour pouvoir afficher en 2 dimensions l'ensemble des pays




mds
United Kingdom, Great Britain, United Kingdom of Great Britain and Northern Ireland
Israel
Kuwait
Cuba
Singapore
Cyprus
United Arab Emirates
Syria
Mongolia
South Africa
Myanmar, Myanmar (Burma)
Luxembourg
New Caledonia,New Caledonia (France)
Canada
Switzerland
Netherlands
France
Bahrain
Oman
Indonesia
Bhutan
Lesotho
Sweden
Denmark, Denmark (including Faroe Islands)
Belgium
Spain
Greece
Belarus
Kazakhstan
Laos
Maldives
People's Democratic Republic
Norway
Finland
Iceland
Austria
Romania
Libya
Tunisia
Jordan
Bangladesh
Czech Republic, Czechia, Czech Republic (Czechia)
Germany
Hungary
Slovenia
Bosnia and Herzegovina
Albania
Ukraine
Estonia
Poland
Lebanon
Algeria
Nepal
India
Kyrgyzstan
Africa
Ireland
Slovakia
Croatia
Latvia
Serbia
Egypt
Pakistan
Australia
New Zealand
Lithuania
Portugal
Bulgaria
Georgia
Tajikistan
Djibouti
Mauritania
Yemen
Italy
Vietnam
Moldova
Afghanistan
Niger
Argentina
Fiji
Comoros
Malta
Chile
Uruguay
Peru
Ecuador
Senegal
Gambia
Paraguay
Nicaragua
Vanuatu
Mali
Japan
Barbados
Sierra Leone
Trinidad and Tobago, Trinidad
South Korea
Kenya
Tanzania
Nigeria
Gabon
Ghana
Togo
Ethiopia
Costa Rica
Mexico
Guatemala
Benin
Panama
Suriname
Madagascar
Dominica
Malawi
Sudan
Colombia
Papua New Guinea
Botswana
Uganda
Cameroon
Chad
Guyana
Haiti
Eritrea
Brazil
Grenada
Samoa
Turkmenistan
Angola
Bahamas
Burundi
Jamaica
Congo, Republic of Congo
Belize
DR, Democratic Republic of the Congo
Congo
Namibia
Mozambique
Zambia
Qatar
Sri Lanka
Honduras
Guinea
Central African Republic
Venezuela, Venezuela, Bolivarian Republic of
Rwanda
El Salvador
Thailand
Cambodia
Liberia
Saudi Arabia
10
5
0
−5
−10
−10
−5
0
5
10


pca
United Kingdom, Great Britain, United Kingdom of Great Britain and Northern Ireland
Saudi Arabia
Qatar
Kuwait
United Arab Emirates
Israel
Niger
Mauritania
Oman
Singapore
Iraq, Iraq (excluding Kurdistan)
Brunei
Mali
Afghanistan
Yemen
Pakistan
Iran, Islamic Republic of Iran
Libya
Bahrain
Switzerland
Algeria
Jordan
Egypt
South Korea
Japan
Senegal
Malaysia
Norway
Chad
Burkina Faso
Australia
India
Turkey
Tunisia
Canada
New Caledonia,New Caledonia (France)
Iceland
France
Ireland
Ethiopia
Liberia
Tajikistan
New Zealand
Lebanon
Thailand
Kazakhstan
Denmark, Denmark (including Faroe Islands)
Nepal
Togo
Eritrea
Sudan
Belgium
Sweden
Turkmenistan
Kyrgyzstan
Sri Lanka
Spain
Austria
Germany
Albania
Malta
Cameroon
Laos, Lao People's Democratic Republic
Vietnam
Russia, Russian Federation
Finland
Mauritius
Gabon
Slovenia
Burundi
Madagascar
Guinea
Poland
Italy
Central African Republic
Guatemala
Hungary
Mozambique
Dominican Republic
Romania
Portugal
Mongolia
Armenia
Nicaragua
Ecuador
Panama
Slovakia
Rwanda
Croatia
Greece
Lithuania
Samoa
Belarus
Paraguay
Chile
Latvia
Honduras
Moldova
Serbia
Bulgaria
Suriname
Colombia
Mexico
Costa Rica
Ukraine
Argentina
Uruguay
Seychelles
Venezuela, Venezuela, Bolivarian Republic of
Belize
Jamaica
Dominica
Cuba
Grenada
El Salvador
Brazil
Namibia
Lesotho
South Africa
6
4
2
0
−2
−4
−6
−4
−2
0
2
4
6
8



2.4.2 4.b Analyser le résultat d'une réduction sur 2 dimensions et d'une réduction sur 5 dimensions

en essayant de trouver une interprétation aux valeurs de chacune des dimensions réduites.

pca 2 dimensions explained variance ratio: [0.37158704 0.09671977] pca 5 dimensions explained variance ratio: [0.37158704 0.09671977 0.06892782 0.06360297 0.04354653]

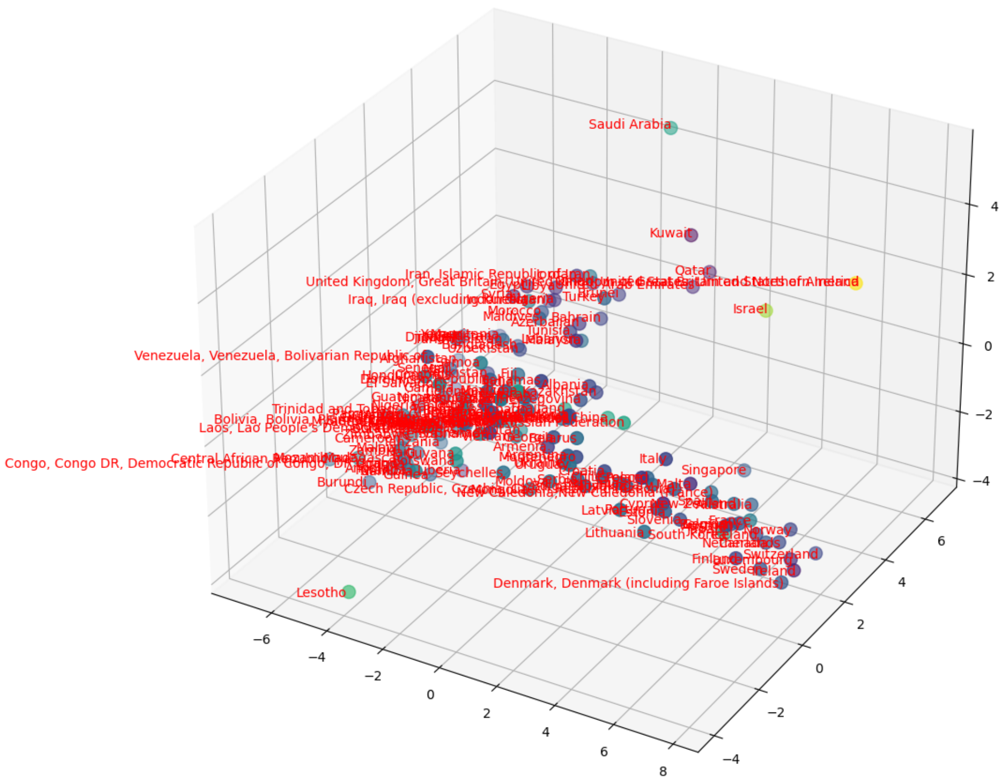
Il serait possible d'ajouter 3 autres dimensions perceptibles à l'oeil dans notre graphique.

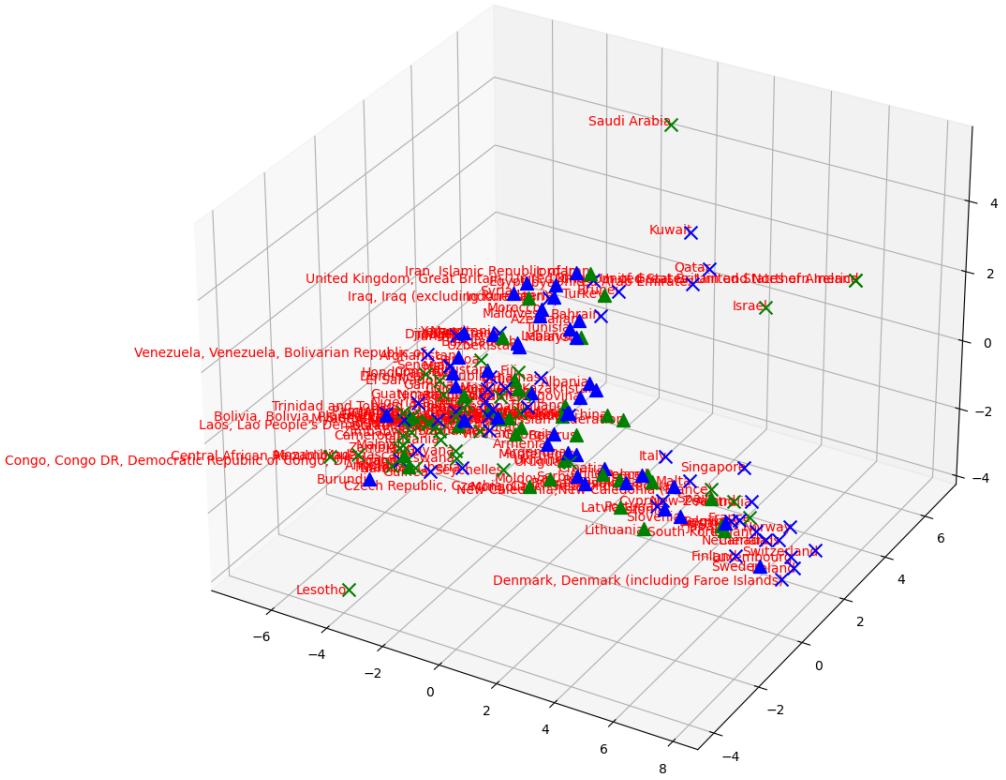
Par exemple, le gradient de couleur pour le text ou le point (quoique ce serait moins perceptible pour le point), la taille du point, une 3ème dimension, un préfix dans le nom de pays ou un marqueur (et bien d'autres).

En résumé, on pourrait avoir un markeur, un point qui change de gradient de couleur et un plan 3D ainsi que les 2 dimensions déjà présente pour interpréter visuellement les 5 dimensions demandées.

NB : J'ai utilisé la borne entre plus petit ou plus grand que zéro pour sélectionner une couleur et

un marqueur pour des fins de simplicité de codage. La version 4D avec gradient de couleur est disponible pour montrer ce que cela aurait donné.





On peut voir une perte d'information pour Lesotho, Israel et le pays en Jaune.

Par contre, ici, nous avons une vue plus directe puisque le reste avait un gradient peu distinctif vers le centre.

2.5 5. Modèle graphique probabiliste

Cette question n'a pas été faite.