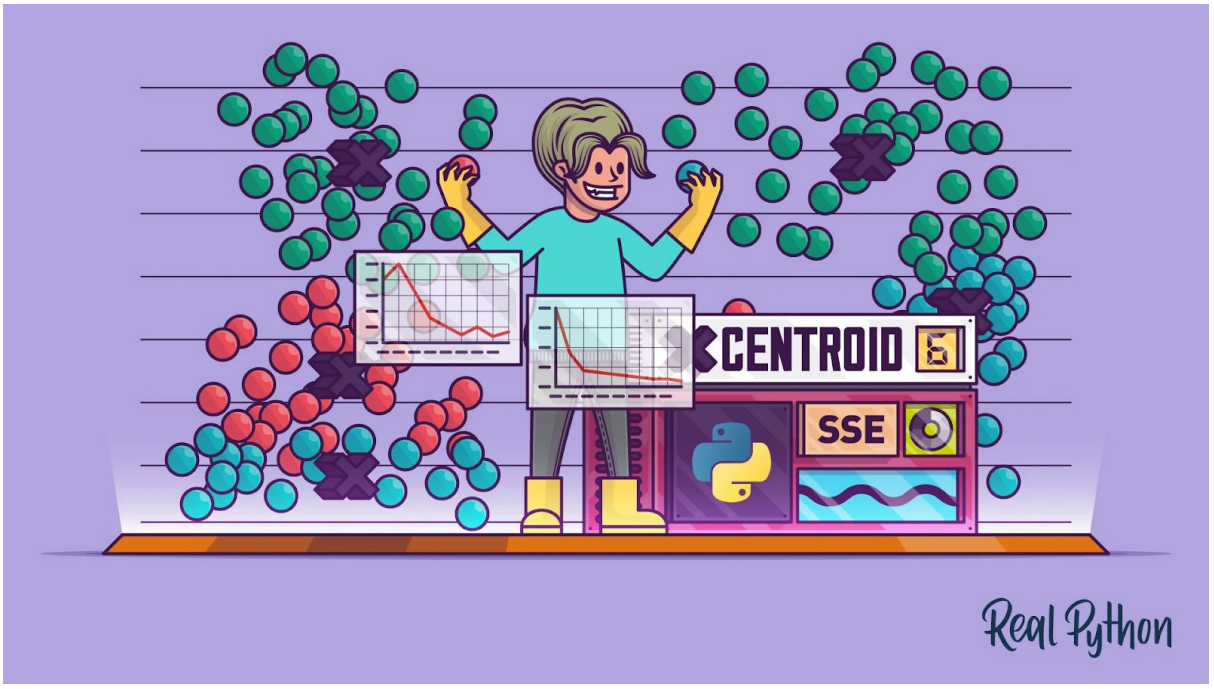
MARDAMA NAYAGOM Anaïs
MARONE Cheikh Ahmet Tidiane Chérif

ILC - TP2

ESIREM
ÉCOLE SUPÉRIEURE D'INGÉNIEURS
NUMÉRIQUE ET MATÉRIAUX

TP-Big Data : Projet-Clustering



Source :
https://files.realpython.com/media/K-Means-Clustering-in-Python_Watermarked.14dc56523461.jpg

TABLE DES MATIÈRES

Introduction **2**

Présentation du jeu de donnée **4**
- **Le Dataset** **4**
- **Champs intéressant** **5**

Fonctions utilisées **6**
- **Apache Spark** **6**
- **Assemblage des features** **6**
- **k-means** **7**
- **Pandas** **8**
- **PCA** **8**
- **Matplotlib** **8**

Interprétation du résultat **9**

Conclusion **12**

Annexes **13**
- **Webographie** **13**
- **Code Source** **13**
- **Lien vers la présentation** **13**

Introduction

L’objectif de ce projet est de se familiariser avec les techniques du clustering en utilisant Apache Spark (ou un autre logiciel), et les techniques de visualisation avec Matplotlib (ou un autre logiciel).

Une dataset doit être choisie et à partir de celle-ci, une analyse devra être faite.
Dans la majorité des dataset, il existe un attribut “prévision” permettant de connaître le résultat attendu et ainsi corriger et améliorer notre notre solution.

Mais cet attribut n’est pas toujours présent. Certaines dataset peuvent être plus imprévisible que d’autres et la dataset choisie en fait partie.

Présentation du jeu de donnée

Le Dataset

Pour ce projet, on a choisi parmi les dataset mise à disposition, celle des drapeaux.
Elle provient du livre Flags de collins gem de 1986.
Constituée de 193 instances, c'est-à-dire 193 pays avec les 30 caractéristiques de leurs drapeaux respectifs. Le but de ce dataset est de faire une classification.
Dans ce dataset, les champs sont séparés par des virgules, on remarque qu'on a des attributs numériques, des booléennes et d'autres textuelles.

Chaque ligne correspond à un pays et les colonnes sont les suivant :

1. name: nom du pays concerné
2. landmass: continent ou région du monde
    1 = Amérique du Nord, 2 = Amérique du Sud, 3 = Europe, 4 = Afrique, 4 = Asie, 6 = Océanie

3. zone: Quadrant géographique, basé sur Greenwich et l'équateur;
    1 = NE, 2 = SE, 3 = SW, 4 = NW
4. area : superficie en milliers de km2
5. population: en millions d'habitant
6. language : langue officielle
    1 = anglais, 2 = espagnol, 3 = français, 4 = allemand, 5 = slave, 6 = autre indo-européen, 7 = chinois, 8 = arabe, 9 = japonais / turc / finnois / magyar, 10 = Autres

7. religion : 0 = catholique, 1 = autre chrétien, 2 = musulman, 3 = bouddhiste, 4 = hindou, 5 = ethnique, 6 = marxiste, 7 = autres

8. bars: nombre de barres verticales dans le drapeau
9. stripes: nombre de rayures horizontales dans le drapeau
10. colors: nombre de couleurs différentes dans le drapeau
11. red: 0 si rouge absent, 1 si rouge présent dans le drapeau
12. green: idem pour le vert
13. blue: idem pour le bleu
14. or: idem pour l'or (également jaune)
15. white: idem pour le blanc
16. balck: idem pour le noir
17. orange: idem pour l'orange (également marron)
18. mainhue: couleur prédominante dans le drapeau
19. circle: nombre de cercles dans le drapeau
20. crosses: nombre de croix (verticales)
21. saltires: nombre de croix diagonales
22. quarters: nombre de sections en quartiers
23. sunstars: nombre de symboles soleil ou étoile
24. crescent: 1 si un symbole de croissant de lune est présent, sinon 0

25. triangle: 1 si des triangles sont présents, 0 sinon
26. icon: 1 si une image inanimée est présente (par exemple, un bateau), sinon 0
27. animate: 1 si une image animée (par exemple, un aigle, un arbre, une main humaine) est présente, 0 sinon
28. text: 1 s'il y a des lettres ou des écritures sur le drapeau (par exemple, une devise ou un slogan), 0 sinon
29. topleft: couleur dans le coin supérieur gauche (se déplacer vers la droite pour décider des tie-breaks)
30. botright: couleur dans le coin inférieur gauche (déplacer vers la gauche pour décider des tie-breaks)

Lien du dataset http://archive.ics.uci.edu/ml/datasets/Flags

Champs intéressant

Dû à cette absence de “prévisions”, il est indispensable de choisir un des autres attribut qui tiendra ce rôle.
Après concertation, on a décidé de classifier ce dataset selon les continents. C'est-à-dire, on associe chaque pays à son continent selon le drapeau. Puisque quelques paternes peuvent être observées, notamment au niveau des couleurs des drapeaux, par exemple, nombreux sont ceux du continent Africain ayant la couleur verte présente.

Pour cela, on a choisi le champ 2 correspondant au continent comme champ de prédiction et le champ 8 jusqu’au champ 30 comme donnée a traité.
Pour effectuer ce traitement, on convertit nos couleurs qui sont initialement en chaîne de caractère en entier, en gardant la valeur des booléens comme des entiers et tout cela grâce à un script shell.
Une fois ce traitement, on a en sortie le fichier flag.csv avec lequel on va travailler pour la suite.

Fonctions utilisées

Pour réaliser ce projet, on a décidé d'utiliser le moteur de traitement de données Apache Spark. Il semble être le meilleur choix puisqu'il est open-source et offre vitesse, simplicité d'utilisation.

Apache Spark

Apache spark est un framework qui permet d'effectuer un traitement de larges volumes de données de manière distribuée.
Pour cela, il faut d'abord le démarrer avec :

```
12  #start Spark Session
13  spark = SparkSession \
14          .builder \
15          .appName("Flag") \
16          .getOrCreate()
```

Le appName est le nom de l'application.

Une fois cette dernière créée, il est temps d'y intégrer la dataset spark.read.csv.
Pour ce faire, la construction d'un schéma avec **StructType** selon un format imposé par *Spark SQL* est nécessaire.

Spark SQL est un module Spark conçu pour le traitement de données structurées. Il apporte une couche d'abstraction en programmation appelée DataFrames et peut également faire office de moteur de requêtes SQL distribué.

Grâce à cela, il est possible d'exploiter des données venant même de différentes sources comme par exemple des fichiers JSON, des tables Hive ou même des bases de données avec JDBC avec une syntaxe SQL. Ce qui laisse ainsi de nombreuses possibilités à ses utilisateurs.

Assemblage des features

Par définition, une *feature* est une propriété ou une caractéristique individuelle mesurable d'un phénomène observé.
Ici les *features* sont les variables d'entrées. Elles sont équivalentes aux variables indépendantes en statistiques. Dans notre cas, ce sont les colonnes du dataset qui seront utilisées pour le calcul des clusters.

Malheureusement pour utiliser l'algorithme **K-means** de Spark, une seule colonne de "*features*" de type Vector est acceptée.
Actuellement, les *features* sont dans des colonnes différentes. Il faut donc les fusionner en une seule colonne de type Vector. Cette tâche sera confiée au **VectorAssembler** de Spark.

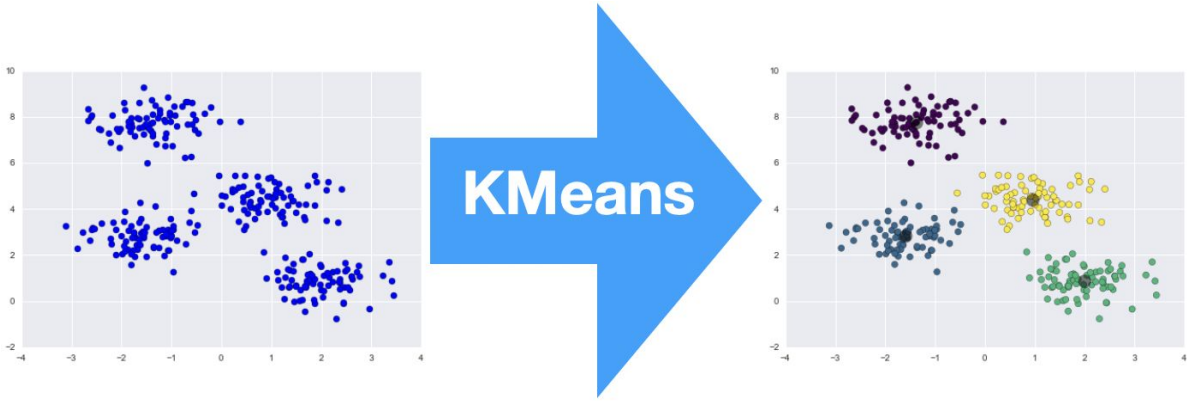
k-means

K-means est l'un des algorithmes de clustering les plus répandus.
Le Clustering est une tâche d'apprentissage non supervisé consistant à regrouper des objets similaires dans des catégories (les clusters).

Cet algorithme permet d'analyser un jeu de données caractérisées par un ensemble de descripteurs (*features*), afin de regrouper les données "similaires" en groupes (ou *clusters*).

La similarité entre deux données peut être inférée grâce à la "distance" séparant leurs *features*. Ainsi deux données très similaires sont deux données dont les descripteurs sont très proches.

Afin de créer des clusters, des centroïdes seront générés, ils représenteront les centres des différents *clusters*. L'algorithme associera chaque donnée à son centroïde le plus proche pour établir à quel groupe cette donnée appartient.



Après avoir initialisé ses centroïdes en prenant des données au hasard dans le jeu de données, K-means alterne plusieurs fois ces deux étapes dans le but d'optimiser la répartition des données :

1. Regrouper chaque objet autour du centroïde le plus proche.

2. Réévaluer chaque centroïde selon la moyenne des descripteurs de son groupe.

Après quelques itérations, l'algorithme trouve un découpage stable du jeu de données : on dit que l'algorithme a convergé.

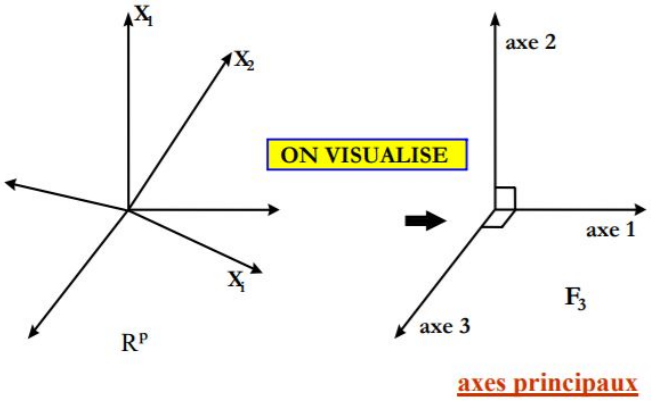
Dans ce projet notre K = 6 et correspond au nombre de "continent" du dataset.
Il est à noter que chaque observation (pays) appartient à un et un seul cluster.

Pandas

Pandas vient de “Python Data Analysis Library”. C’est une librairie python qui permet de manipuler facilement des données à analyser comme des tableaux de données. Ces tableaux sont appelés DataFrames.
Ce qui permet de facilement visualiser le partitionnement effectué par l’algorithme précédent en traçant des graphes à partir de ces DataFrames grâce à **matplotlib**.

PCA

L’analyse de composant principal ou PCA en anglais permet de détecter automatiquement les axes les plus importants pour les meilleures projections possibles.



En effet le PCA consiste à transformer des variables liées entre elles en nouvelles variables décorrélées les unes des autres. Ces nouvelles variables sont nommées « composantes principales », ou axes principaux. Elle permet au praticien de réduire le nombre de variables et de rendre l'information moins redondante.

Matplotlib

Inspiré de Matlab au départ, matplotlib est une librairie qui permet de tracer des graphes qui peuvent être complètement adaptés si besoin. Sur une figure, on peut tracer plusieurs graphes.

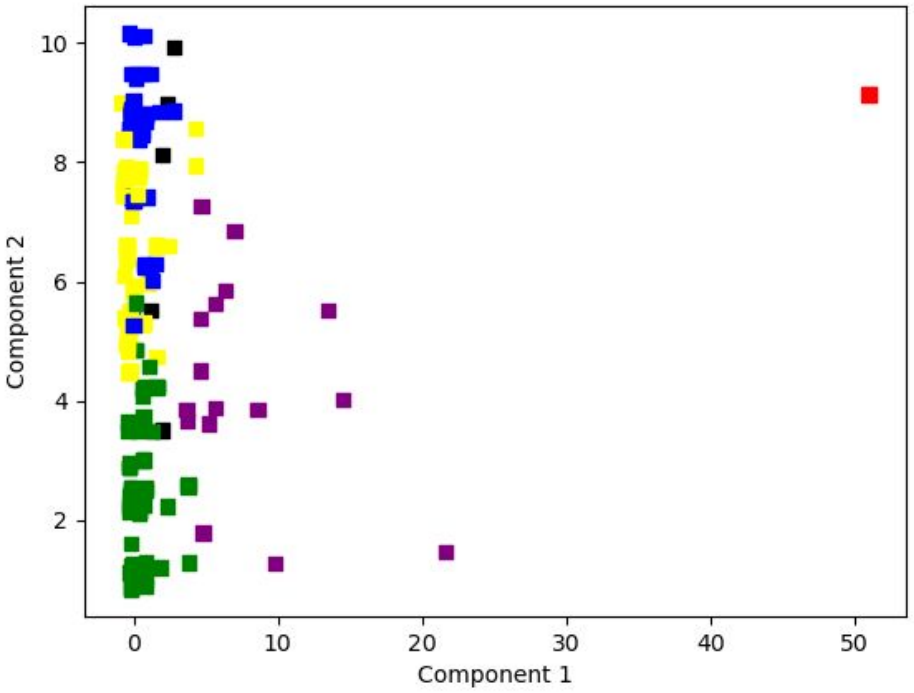
Interprétation du résultat

Grâce au PCA, on a eu les 3 projections suivantes:

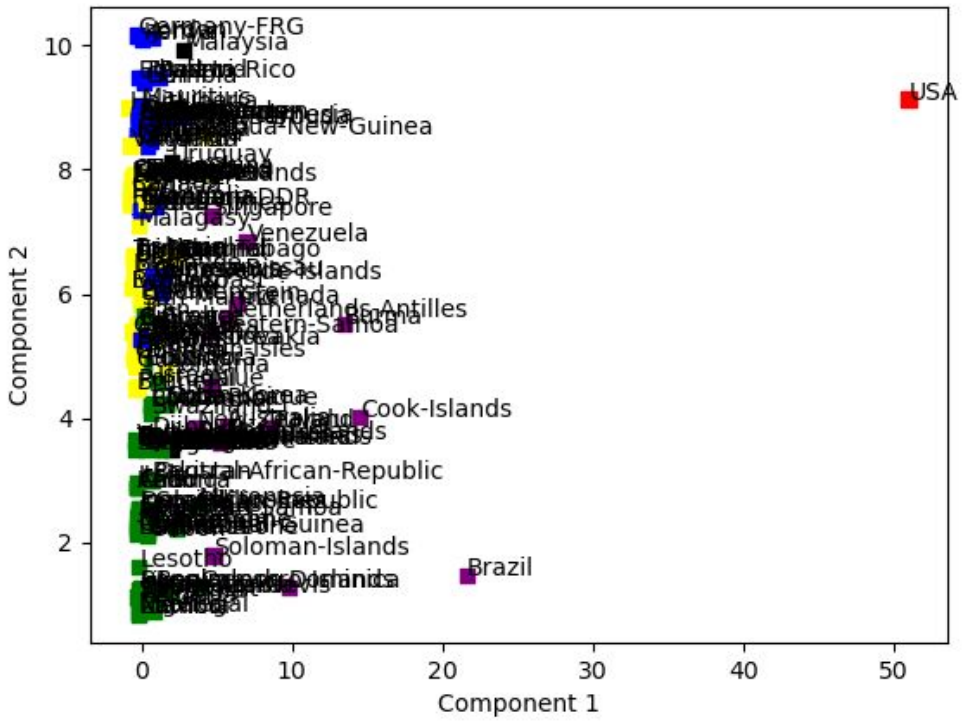
- component 1 correspond sunstart et stripes
- component 2 constitué de bar, botright, mainhue, sunstar
- component 3 : stripes, mainhue, crescent , botright

En 2D

- Sans abel

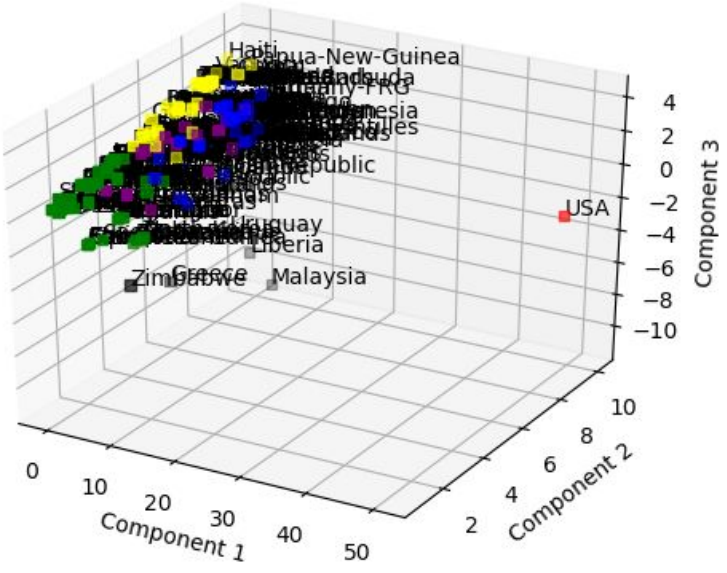


- Avec Label

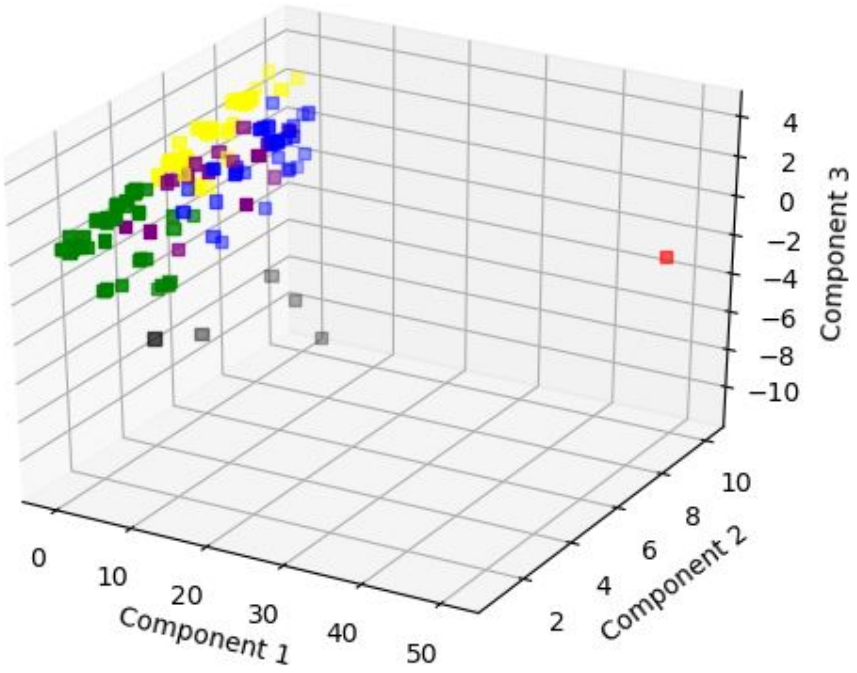


En 3D

- Avec Label



- Sans Label



En somme on a 6 clusters dont 1 est uniquement constitué d'un pays : les USA

On constate que la majorité des pays africains se ressemblent (drapeaux). En effet, le sao tomé, le lesotho, le djibouti appartiennent à ce cluster. Néanmoins on y trouve aussi la corée du nord , l'iran, l'israël…

Les Etats-unis ont un drapeau très singulier (51 étoiles dans son drapeau) .

Le liberia, la Malaisie dont le drapeau n'est pas très éloigné des USA constituent avec uruguay, la grèce et le zimbabwe un cluster. Donc on peut dire que ce cluster (Gris/Noir) réunit les pays dont le drapeau est constitué à 80% ou plus de bar horizontal.

Il est à noter que la correspondance pays et continent ne fonctionne pas. En effet, la majorité des pays africains sont en Amérique du Nord et les Etats-unis constituent un seul continent.

Remarque: Le dataset n'est pas à jour également, présence notamment de la Yougoslavie.

Conclusion

Cette mise en pratique a permis d'approfondir nos connaissances en Spark et d'observer le processus d'analyse et de traitement de données.

Le résultat final que nous avons obtenu ne reflète pas les véritables capacités de clustering qu'offre le Big Data.

Mais à travers cette dataset nous avons eu la possibilité de constater les limites auxquelles les professionnels d'analyse peuvent faire face. Il arrive quelquefois que les données dont nous disposons ne peuvent être regroupées selon des motifs définis, ce qui rappelle la théorie du Chaos.

Annexes

Webographie

http://www.python-simple.com/python-matplotlib/matplotlib-intro.php
https://mrmint.fr/algorithme-k-means
https://fr.wikipedia.org/wiki/Analyse_en_composantes_principales
https://www.lebigdata.fr/apache-spark-tout-savoir
http://www.python-simple.com/python-pandas/panda-intro.php

Code Source

https://github.com/bruaba/BigData-TP-Projet-Clustering

Lien vers la présentation

https://www.canva.com/design/DAEPvXX4xsE/Q8gdEXOMWvocYeLCyrwYRQ/view?utm_content=DAEPvXX4xsE&utm_campaign=designshare&utm_medium=link&utm_source=publishsharelink