Exercice Evenium : Aymeric le Proux de La Riviere

**Question 1 :**

You will find in the same folder the python code to extract, clean the Data from the excel file and count the number of occurrence by word.

**Question2 :**

To see the word cloud just open the html file in your Browser.

To keep the spiral shape i used the equations of the Archimede Spiral but the angle between 2 words and the radius augmentation coefficient decreases as the number of words added in the word cloud increases.

**Question 3 :**

Explain K-means to a 10 year old

The K-Means Algorithm is a well-known method to find similarity-group in a dataset. The idea is to gather individuals objects in a population together by similarity. The groups thus created are called « clusters ».

This may not be very clear, so this is an imaged example:

You have a farm with 3 kinds of animals (cow, duck, goat). You know the characteristics for each animal (for instance their speed and their size).

What you want to do is to gather all the animal of each species in an enclosure and thus have 3 enclosures: one for each species.

Moreover, everyday you receive a new animal in a black box, you don’t know what is its kind, you can’t see it, you just know its characteristic (speed and size) and you want to put it in the right enclosure. Just by knowing the characteristics you must be able to know the species and that’s what the K-means algorithm will help you do!

The algorithm :

You want to create K groups (K enclosure) for you 3 species. Here K=3

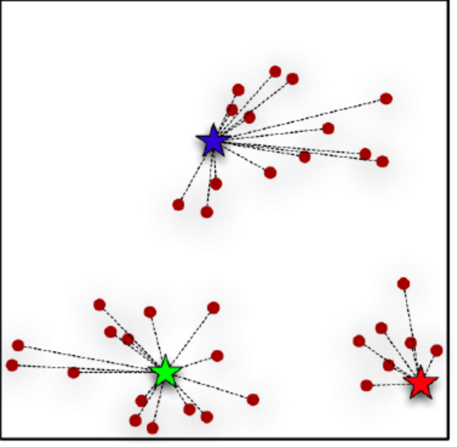
You put all your animal in a field according to their speed and size :the bigger one are on the top of the field and the heavier one on the right of the field.

You randomly put 3 flags in this field and create 3 groups :

First step:

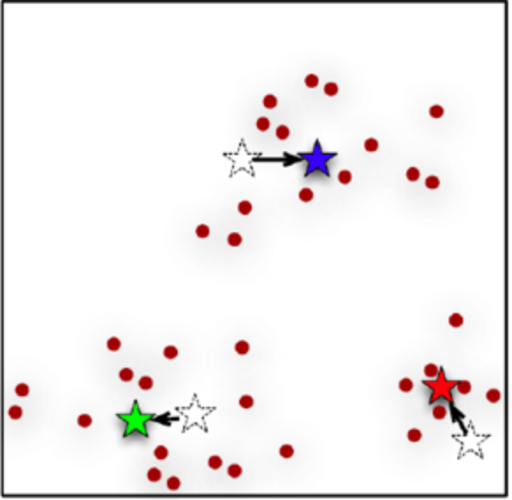
Add animals to their group : an animal belong to a group if among all the 3 flags ; this group’s flag is the closest to the animal.

SEE the picture below (the flags are represented by the stars)



Then you change the flag position: For each group the new position of the flag is the average position of all the animals of the group.

You now have 3 flags and you do the first step again and recreate 3 groups.



You keep doing it until the flag doesn’t move anymore and when it s done you have your 3 groups. The center of each group is the flag.

STEP 2 : add unknown animal to the right group

Use your groups to detect the kind of the mystery animal in the black box: you place your black box on the field (you know its speed and weight so you know where to put it) this animal will go to the enclosure whose position is the closest to the group’s flag.

For instance, if the animal has the size of a duck and run as fast as a duck then it must be a duck and so you can add the mysterious animal to the duck enclosure!

What have we done : we have created enclosure for each species of our farm and we have find a way to add unknown animals to the right enclosure

The main difficulty is to chose the good value of K :

For instance if you chose K=2 you will create 2 enclosure and have 2 kinds of animals in the same enclosure. You will have the same problem if K is too big. Most of the time you will try different value of K to know which one is the best.