COURS PYTHON TSI2

SOMMAIRE

Chapitre 1

Recursion and sorting recursion

■ Objectifs

Les incontournables :

savoir

Résumé de cours

■ Récurrence et fonctions récursives

Nous savons que l'on peut définir des suites en ne donnant pas une formule explicite pour u_n mais une formule de récurrence (et le premier terme de la suite). Par exemple, même si l'on ne sait pas en trouver une formule explicite, il existe une unique suite vérifiant

$$u_0 = 1 \text{ et } \forall n \in \mathbb{N},\ u_{n+1} = u_n^2 - n.$$

Si l'on veut écrire une fonction qui calcule le terme de rang n de cette suite, on peut écrire une boucle, dans laquelle on commence par calculer le terme de rang 1, puis le terme de rang 2...
On dit que la méthode est **impérative**.

```
In [1]: def u(n):
   ...:     x = 1
   ...:     for k in range(1,n+1):
   ...:         # calcul de u(k)=u^2(k-1) -(k-1)
   ...:         x = x**2 -k+1
   ...:     return x
In [2]: u(7),u(12)
Out[2]: (10, 310822058031477121387888456118 65)
```

On voit vu la tête de `u(12)` que l'on monte vite. Ici, nous avons détaillé la succession de calculs à faire pour arriver à u_n. Mais on peut faire autrement.

```
In [4]: def uRec(n):
   ...:     if n == 0 :
   ...:         return 1
   ...:     else :
   ...:         return uRec(n-1)**2 -(n-1)
   ...:
In [5]: uRec(7),uRec(12)
Out[5]: (10, 310822058031477121387888456118 65)
```

Une telle fonction est dite **récursive**, i.e. que dans la définition de la fonction, la fonction elle-même apparaît. Cette fois-ci, nous ne détaillons aucun calcul mais nous allons développer le cheminement pour calculer $uRec(4)$. Pour calculer u_4, la fonction s'appelle elle-même pour calculer u_3. Elle se réappelle encore pour calculer u_2, puis idem avec u_1 et enfin avec u_0. Elle connaît la réponse pour u_0 c'est 1. Elle renvoie cette valeur ce qui permet de calculer u_1. Après ce calcul, elle renvoie u_1 et on peut enfin calculer u_4.

On peut comparer le temps d'execution de la fonction `u` et de la fonction `uRec`
Pour cela, on use de la fonction prédefinie `time` que vous connaissez déjà de `perf_counter`

```
In [7]: from time import perf_counter as pc
In [8]: for n in range(2,10) :
   ...:     t = pc()
   ...:     u(n)
   ...:     print(pc() - t)

2.6999999818144715e-06 4.500000045482011e-06
2.7000000955013093e-06 2.499999936844688e-06
2.900000026784255e-06 3.199999923708674e-06
3.6000000136482413e-06 4.3999999661537e-06

In [9]: for n in range(2,10) :
   ...:     t = pc()
   ...:     uRec(n)
   ...:     print(pc() - t)

2.6000000161729986e-06 3.6000000136482413e-06
4.900000021734741e-06 5.600000008598727e-06
5.399999963628943e-06 6.199999916134402e-06
6.20000002982124e-06 7.3999999585794285e-06
```

Rien de bien significatif.

```
In [11]: for n in range(20,30) :
   ...:     t = pc()
   ...:     u(n)
   ...:     print(pc() - t)

0.00013369999999213178 0.0003566000000319036
0.00101859999995213 0.0034633999999869047
0.009808200000065881 0.03159629999993285
0.09322210000004816 0.26550889999998617
0.7767928999999185 2.3187297999999146

In [12]: for n in range(20,30) :
   ...:     t = pc()
   ...:     uRec(n)
   ...:     print(pc() - t)
0.00013879999994514947 0.0003527000000076441
0.0010068000000273969 0.0030861999999842737
0.010145299999976487 0.028480800000011186
0.08652330000006714 0.26422969999998713
0.7755059999999503 2.32723390000001
```

Encore rien de significatif. En fait les méthodes impératives ou récursives peuvent être comparables dans certains cas et dans d'autres il vaut mieux l'une que l'autre. Cela dépendra de l'algorithme.

Mise en œuvre : de l'exercice 01 à l'exercice 06

■ fonctions définies sur les listes

Les fonctions récursives que nous avons vues jusqu'à présent étaient définies sur $\mathbb{N}$, ce qui permettait de justifier leur terminaison :
• on commençe par traiter le (ou les) cas de base, pour lesquels la fonction renvoie directement une valeur (en général $n = 0$ et $n = 1$, voire d'autres situations plus compliquées) ;
• dans les autres cas, la fonction se réappelle elle-même, avec un argument strictement inférieur à celui de l'appel précédent.

Puisque l'argument diminue strictement, on finira, lors des appels récursifs, par tomber sur un des cas de base, pour lesquels la fonction renverra directement une valeur, ce qui permettra, par remontées successives, de terminer le calcul.

Puisque l'argument diminue strictement, on finira, lors des appels récursifs, par tomber sur un des cas de base, pour lesquels la fonction renverra une valeur, ce qui permettra, par remontées successives, de terminer le calcul.

Une autre situation dans laquelle il est commode de définir des fonctions récursives est celle des listes :
• on commence par traiter le(s) cas de base (en général, la liste vide, éventuellement le cas d'une liste de longueur 1, voire d'autres cas) ;
• dans le cas général, la fonction se réappelle elle-même, avec pour argument une nouvelle liste, dont la longueur est strictement inférieure à celle de l'appel précédent.

Dans ce cas aussi, on est assuré de la terminaison de la fonction.

Exemple : Écrivons une fonction récursive qui calcule la somme des termes d'une liste (de nombres).
• si la liste est vide, la somme est nulle ; on renvoie alors la valeur 0;
• sinon, il suffit de calculer (récursivement) la somme des termes de la sous-liste obtenue en retirant le dernier élément, et d'ajouter ce dernier élément.
On utilise $l[-1] = l[len(l) - 1]$ pour définir le dernier élément d'une liste et $l[: -1]$ pour définir la liste emputée de son dernier élément. On tape alors `sommesTermesRec(l)`

```
In [13]: def sommeTermesRec(l):
    ...:     if len(l) == 0:
    ...:         return 0
    ...:     else:
    ...:         return l[-1] + sommeTermesRec(l[:-1])
In [14]: sommeTermesRec([1,4,-4,7])
Out[14]: 8
```

Remarque : Cette façon d'écrire la fonction est un peu maladroite. En effet, en Python, l'appel à `L[:-1]` crée une copie de la liste `l` (privée de son dernier terme). Il y a donc ici autant de copies créés que d'appels récursifs. Il est préférable de ne créer aucune copie, en écrivant une nouvelle fonction (récursive) qui calcule la somme entre deux indices `deb` et `fin` À chaque appel récursif, on remplace l'indice `fin` par son prédécesseur. Le cas terminal est celui d'une plage vide, c'est-à-dire celui où `fin < debut` On tape :

```
In [15]: def sommeEntreRec(l,deb,fin):
    ...:     if fin < deb:
    ...:         return 0
    ...:     else :
    ...:         return l[fin] + sommeEntreRec(l,deb,fin-1)
```

La terminaison est now assurée par le fait que, lors des appels récursifs, la quantité `fin-deb` décroît strictement. Elle finira donc par devenir strictement négative, ce qui est le cas terminal.
La somme totale des éléments de la liste s'obstient en tapant :

```
In [16]: def NewsommeTermesRec(l):
    ...:     sommeEntreRec(l,0,len(l)-1)
In [17]: NewsommeTermesRec([1,4,-4,7])
Out[17]: 8
```

Mise en œuvre : de l'exercice 07 à l'exercice 09

■ fonctions définies sur d'autres types

L'argument d'une fonction récursive peut être plus compliqué qu'un entier ou une liste. Nous avons vu déjà quelques cas : `f(a,b)` de l'exercice 06 ou `sommeEntreRec(l,deb,fin)`
Dans ces deux exemples, cependant seul un des paramètres variait lors des appels récursifs (`b` remplacé par `b-1` ou `fin` par `fin-1`). La terminaison était alors facilement établie.

Exemples : 1. Un premier exemple
Que dire de la fonction suivante :

```
In [26]: def s(a,b) :
    ...:     if b == 0:
    ...:         return a
    ...:     else :
    ...:         return s(a+1,b-1)
```

Cette fois ci, les deux arguments évoluent et `a` augmente ! Ce n'est pas grave car le cas terminal est `b=0` et le deuxième argument diminue strictement à chaque appel récursif. Si `b` est un entier positif, la fonction se terminera.

Si l'on suppose que `a` et `b` sont deux entiers positifs, que renvoie `s(a,b)` ?

```
In [28]: s(2,4), s(2024,2025)
Out[28]: (6, 4049)
```

2. Nombre de chemins de déplacement d'un pion

Un pion se déplace sur un échiquier dont les cases sont numérotées.

Le départ est situé au point de coordonnées $(0,0)$. À chaque déplacement, le pion se déplace d'une et d'une seule case, soit vers la droite, soit vers le haut. On cherche à savoir combien de chemins distincts permettent de rejoindre le point de coordonnées (n,p) avec n et p deux entiers positifs connus.

• dans le cas général, les chemins qui aboutissent en (n,p) se divisent en deux catégories : ceux qui proviennent du point $(n,p-1)$ et ceux qui proviennent du point $(n-1,p)$. Ce raisonnement s'applique si $n,p \geqslant 1$.
On en déduit que $f(n,p) = f(n,p-1) + f(n-1,p)$.

• dans le cas où $n = 0$, il y a un seul type de chemin qui se termine en $(0,p)$: c'est de provenir de $(0,p-1)$ (du moins si $p \geqslant 1$). On en déduit $f(0,p) = f(0,p-1)$;

• de même, si $p = 0$, on a : $f(n,0) = f(n-1,0)$;

• enfin, il y a un unique chemin qui rermine en $(0,0)$: celui qui consiste à effectuer 0 déplacements.

```
In [29]: def nbChemins(n,p):
    ...:     if n == 0 and p == 0:
    ...:         return 1
    ...:     elif n == 0:
    ...:         return nbChemins(0,p-1)
    ...:     elif p == 0 :
    ...:         return nbChemins(n-1,0)
    ...:     else :
    ...:         return nbChemins(n-1,p)+nbChemins(n,p-1)
```

À la main calculer `nbChemins(1,1)` et `nbChemins(2,2)`

```
In [34]: nbChemins(4,4), nbChemins(4,7), nbChemins(8,10)
Out[34]: (70, 330, 43758)

In [35]: nbChemins(12,14)
Out[35]: 9657700
```

Pour `nbChemins(12,14)` ce ne fut pas instantané.

Comment justifier la terminaison de la fonction ? Cette fois-ci, au cours des appels récursifs, les deux paramètres évoluent, et aucun ne décroît strictement à chaque étape (parfois, le paramètre p est remplacé par lui-même ; parfois c'est n). Il faut cette fois ci s'intéresser à la quantité $s = n+p$. Celle-ci décroît strictement (d'une unité) à chaque appel récursif. Comme le cas $n = p = 0$ est l'unique cas de base, on finira par tomber sur ce cas lors des appels récursifs, et la fonction se terminera.

Remarque : on peut simplifier un peu la définition (et le nombre d'appels récursifs) en remarquant que si $n = 0$ alors il existe un unique chemin menant de $(0,0)$ à $(0,p)$, qui consiste à se déplacer vers le haut à chaque étape. On a donc $f(0,p) = 1$ (et de même $f(n,0) = 1$).

On obtient :

```
In [36]: def New_nbChemins(n,p):
    ...:     if n == 0 or p == 0 :
    ...:         return 1
    ...:     else :
    ...:         return New_nbChemins(n-1,p)+New_nbChemins(n,p-1)
    ...:

In [37]: New_nbChemins(4,4), New_nbChemins(4,7), New_nbChemins
   (8,10)
Out[37]: (70, 330, 43758)
```

Exercice d'application : Montrer que dans l'exemple précédent, le nombre de chemins est

$$f(n,p) = \binom{n+p}{n} = \binom{n+p}{p}.$$

3. Exponentiation rapide

Le problème est de calculer x^n avec $n \in \mathbb{N}$ le plus efficacement possible.

La solution qui vient à l'esprit est de calculer $x \times x \times ... \times x$ ce qui se fait au prix de $n-1$ multiplications.

Mais dans le cas où $n = 8$, on peut écrire : $x^8 = (x^2)^4 = (x^2)^{2\times 2} = ((x^2)^2)^2$.

Ce calcul ne nécessite que trois élévations au carré au lieu de sept multiplications dans la méthode classique.

De façon générale, si n est un nombre pair, on a utilisé la formule : $x^{2p} = (x^2)^p$.

Le problème est donc ramené de l'exposant $n = 2p$ à celui de l'exposant p qui est deux fois plus petit.

Dans le cas où n est impair, alors posons $n = 2p+1$ et : $x^{2p+1} = x(x^2)^p$. On tape :

```
In [38]: def puissRapide(x,n):
    ...:     if n == 0 :
    ...:         return 1
    ...:     elif n % 2 == 0:
    ...:         return puissRapide(x*x,n//2)
    ...:     else:
    ...:         return x * puissRapide(x*x, n//2)

In [39]: puissRapide(3,2) , puissRapide(5,15)
Out[39]: 9   30517578125
```

On peut donner une forme avec deux cas terminaux : $n = 0$ et $n = 1$.

On tape :

```
In [41]: def New_puissRapide(x,n):
    ...:     if n == 0 :
    ...:         return 1
    ...:     elif n == 1 :
    ...:         return x
    ...:     elif n % 2 == 0:
    ...:         return New_puissRapide(x*x,n//2)
    ...:     else:
    ...:         return x * New_puissRapide(x*x, n//2)

In [42]:  New_puissRapide(3,2),  New_puissRapide(5,15)
Out[42]: (9, 30517578125)
```

La terminaison est assurée par le fait que, lors de chaque appel récursif, l'exposant décrôit strictement, donc finit par tomber sur 0, pour lequel la fonction renvoie une valeur.

Cet algorithme est très intéressant car il permet de passer d'une complexité linéaire (en l'exposant) pour le calcul des puissances à une complexité logarithmique. En effet, à chaque appel récursif, l'exposant est au moins divisé par deux, et deux multiplications au plus sont effectuées.
À l'issue de p appels récursifs, l'argument est donc au plus égal à $n/2^p$.
On tombe sur un cas terminal dès que ce nombre est strictement inférieur à 2.
On cherche donc le plus petit entier p tel que :

$$\frac{n}{2^p} < 2 \text{ et } \frac{n}{2^{p-1}} \geqslant 2.$$

Cela donne : $p \leqslant \log_2(n)$.

4. Calcul de PGCD : algorithme d'Euclide

Rappelons que si `a` et `b` sont deux entiers positifs ou nuls (l'un au moins étant non nul), le PGCD de `a` et de `b` est le plus grand diviseur commun à `a` et `b`. La remarque intéressante pour le calculer est que les diviseurs communs à `a` et `b` sont les mêmes que les diviseurs communs à `a-b` et `b`

Donc : `PGCD(a,b) = PGCD(a-b,b)`

Par une récurrence immédiate, on a :

$$\forall q \in \mathbb{N} \text{ avec } \texttt{a-qb} \geqslant 0, \texttt{PGCD(a,b) = PGCD(a-qb,b)}.$$

Le plus grand possible est le quotient de la division euclidienne de `a` par `b` Alors `a-bq` est le reste `r` de cette division euclidienne et : `PGCD(a,b) = PGCD(r,b).`

Dans le cas où `b=0`, les diviseurs communs à `a` et `b=0` sont les diviseurs de `a` Le plus grand est donc `a`

Enfin, par convention, $PGCD(0,0) = 0$ et donc la formule `PGCD(a,0)=a` est encore vraie pour `a=0`

```
In [43]: def PGCD(a,b):
    ...:     if b == 0 :
    ...:         return a
    ...:     else :
    ...:         return PGCD(a%b,b)
```

Vérifions la terminaison de la fonction. Le cas terminal est `b=0` Or, dans notre codage, `b` n'est jamais modifié, donc s'il est initialement non nul, on ne tombera jamais sur le cas terminal.
Donc la fonction boucle sans fin...
Pour corriger cela, on remarque que : `PGCD (a,b) = PGCD(b,a)` et on tape alors :

```
In [44]: def GutPGCD(a,b):
    ...:     if b == 0 :
    ...:         return a
    ...:     else :
    ...:         return GutPGCD(b,a%b)
```

Cette fois la fonction se termine car `b` est remplacé par le reste de la division de `a` par `b` qui est strictement inférieur à `b` On finit par tomber sur le cas terminal.

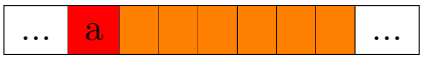
Exercice Faire tourner `PGCD(28,35)` à la main. Avec l'ordi, on trouverait :

```
In [45]: GutPGCD(28,35)
Out[45]: 7
```

■ Un premier tri récursif : Algorithme de tri rapide

Pour trier la sous-liste $l[i..j]$, l'algorithme de tri rapide consiste à :

1. choisir un élément pivot a (cellule rouge) dans la sous-liste $l[i..j]$ (plage rouge + orange). On choisit par exemple le premier `l[i]`



2. permuter les éléments de la sous-liste (symbolysée plus bas par les couleurs yellow, red et blue) de sorte que, après permutation, tous les éléments inférieurs à a (plage yellow) soient situés avant a, tous les éléments supérieurs à a soient situés après a (plage blue).

3. relancer le point 1 sur les sous-listes $l[i..k-1]$ et $l[k+1..j]$ si elles sont de longueur strictement supérieure à 1.

Commençons par écrire une fonction *partition* qui partitionne une liste entre les indices *deb* et fin, suivant l'élément a d'indice *deb*. Nous aurons besoin que cette fonction renvoie la nouvelle position de a, pour que l'on sache quelles sont les deux sous-listes à trier à l'étape suivante.
Pour écrire cette fonction, examinons les éléments l_i, pour $i \in [\![deb,\ fin]\!]$, en maintenant l'invariant de boucle suivant :

- $deb \leqslant m \leqslant i \leqslant fin$
- $\forall k \in [\![deb,\ m]\!],\ l_k \leqslant a$
- $\forall k \in [\![m+1,\ i]\!],\ l_k > a$.

Et passons maintenant à `partition(l,deb,fin)` qui est la fonction Python correspondante.

```
In [46]: def partition(l,deb,fin):
    ...:     a=l[deb] ; m =deb
    ...:     for i in range(deb+1,fin+1):
    ...:         if l[i] <= a:
    ...:             m += 1
    ...:             l[i],l[m] = l[m],l[i]
    ...:     l[deb],l[m] = l[m], l[deb]
    ...:     return m
```

Mise en œuvre : exercice 10

Remarque : On peut taper `return (m,l)` pour voir comment `l` est transformé. C'est ce que l'on fera à l'exercice 10.

Écrivons maintenant la fonction récursive qui trie le tableau entre les indices `deb` et `fin`

```
In [56]: def triEntre(l,deb,fin):
    ...:     if deb < fin:
    ...:         m = partition(l,deb,fin)
    ...:         triEntre(l,deb,m-1)
    ...:         triEntre(l,m+1,fin)
```

Il reste enfin la fonction principale, qui appelle la fonction récursive. On tape :

```
In [57]: def triRapide(l):
    ...:     triEntre(l,0,len(l)-1)
    ...:     return l
    ...:
In [58]: triRapide([3,-1,2,3,5,4,8,0,4,2,5,-5])
Out[58]: [-5, -1, 0, 2, 2, 3, 3, 4, 4, 5, 5, 8]
```

Complexité du tri rapide :

On peut montrer que le pire des cas est celui où, à chaque étape de partition, une des deux sous-listes est vide (et l'autre de taille $n-1$). La complexité maximale vérifie donc la relation de récurrence :

$$C_{max}(n) = C_{max}(n-1) + n - 1.$$

D'où l'on tire (comme $C_{max}(1) = 0$) que :

$$C_{max}(n) = \sum_{k=2}^{n}(k-1) = \frac{n(n-1)}{2} \sim \frac{n^2}{2}.$$

A l'inverse, le cas le plus favorable est celui, où, à chaque étape, les deux sous-listes résultant de la partition ont la même taille (à une unité près). Calculons la complexité dans ce cas.

Pour ce faire, nous nous limiterons au cas où la liste est de longueur $n = 2^p - 1$.

Notons alors

$$x_p = C_{min}(2^p - 1).$$

Si $p \geqslant 2$, la partition de la liste se fait au prix de $2^p - 2$ comparaisons et produit de deux sous-listes de tailles égales à $2^{p-1} - 1$.

On a donc la formule :

$$x_p = 2x_{p-1} + 2^p - 2 \Rightarrow x_p - 2 = 2(x_{p-1} - 2) + 2^p.$$

On divise tout par 2^p et l'on a, en posant $y_p = \dfrac{x_p - 2}{2^p}$,

$$y_p = y_{p-1} + 1.$$

C'est une suite arithmétique classique et comme $x_1 = 0$, $y_1 = -1$.

On a : $y_p = p - 1 + y_1 = p - 2$ puis $x_p = (p - 2)2^p + 2$.

Comme ici $n = 2^p - 1$, $2^p \sim n$ et $p \sim \log_2(n)$. On obtient donc :

$$C_{min}(n) \sim n \log_2(n).$$

Nous admettrons que ce résultat reste vrai même lorsque n est quelconque. On résume :

Proposition. Les complexités dans le pire des cas et dans le meilleur des cas du tri rapide sont respectivement :

$$C_{min}(n) \in O(n \log n) \text{ et } C_{max}(n) \in O(n^2)$$

Ce tri n'est pas très intéressant.

Notons que ce tri n'est pas stable : lors de l'étape de partition, un élément de même clé que le pivot dans la plage $l[deb..fin]$ sera permuté avec le pivot.

■ Un second tri récursif : Algorithme de tri fusion

Le principe de ce tri est de découper la liste en deux, de trier (récursivement) chaque partie, et de fusionner ces deux sous-listes pour obtenir une liste triée.

Commençons par écrire une fonction de fusion de deux listes l_1 et l_2, déjà triées.

Plus précisément, puisque notre fonction doit trier la liste l (et non renvoyer une autre liste, dont les éléments sont ceux de l mais triès.

Pour cela, nous créons une liste auxiliaire aux, dans laquelle nous réalisons le tri, avant de recopier le résultat dans la liste l.

On tape alors en Python la fonction `fusion(l,deb,mil,fin)` suivante.

```
In [59]: def fusion(l,deb,mil,fin):
    ...:     aux = []
    ...:     i,j = deb, mil+1
    ...:     while i <= mil and j <= fin :
    ...:         if l[i] < l[j]:
    ...:             aux.append(l[i]) ; i += 1
    ...:         else :
    ...:             aux.append(l[j]) ; j += 1
    ...:     while i <= mil :
    ...:         aux.append(l[i]) ; i += 1
    ...:     while j <= fin :
    ...:         aux.append(l[j]) ; j += 1
    ...:     for k in range(deb,fin + 1) :
    ...:         l[k] = aux[k-deb]

# On ecrit la fonction qui trie l entre deb et fin

In [60]: def triFusionEntre(l,deb,fin):
    ...:     if deb < fin :
    ...:         m = (deb+fin)//2
    ...:         triFusionEntre(l,deb,m)
    ...:         triFusionEntre(l,m+1,fin)
    ...:         fusion(l,deb,m,fin)

# Puis on tape la fonction principale

In [61]: def triFusion(l):
    ...:     triFusionEntre(l,0,len(l)-1)
    ...:     return l

In [62]: triFusion([-2,2,3,7,9,10,1,8,1,4])
Out[62]: [-2, 1, 1, 2, 3, 4, 7, 8, 9, 10]
```

Mise en oeuvre : Exercice 11

Proposition. La complexité (dans tous les cas) du tri fusion vérifie : $C(n) \in O(nlogn)$

Notons enfin que ce tri est stable.

Chapitre 2

Dictionary

■ Objectifs

Les incontournables :

savoir

Résumé de cours

■ Introduction

Les dictionnaires permettent de mémoriser des données. Ces données peuvent être des résultats intermédiaires qui sont produits par des calculs et qui doivent être réutilisés ensuite.
Les dictionnaires sont des types construits. Rappelons les types principaux de Python.
Le type `tuple` : C'est une liste ordonnée d'objets délimitée par des parenthèses. Ces objets sont indéxés par des entiers.

```
In [1]: a = (1,2,3) ; a = a + (5,6)
In [2]: a
Out[2]: (1, 2, 3, 5, 6)
```

On peut concaténer des `tuples`.

Le type `list` : C'est une liste ordonnée d'objets délimitée par des crochets. Ces objets sont indéxés par des entiers.

```
In [1]: l=[3,1,4]; l.append(5)
In [2]: l
Out[2]: [3, 1, 4, 5]
```

Le type `dict` : C'est une liste non ordonnée d'objets délimitée par des accolades, appelée dictionnaire. Ces objets sont accessibles par des clés (au lieu d'un indice). Une clé peut être n'importe quel objet, un `tuple` mais pas une liste ou un autre dictionnaire. On construit un dictionnaire en associant des successions de couples, chaque couple est constitué d'un premier élément, appelé clé, `keys` en Python et un second appelé valeur, `values` en Python.

```
In [1]: d={0:1,'login':'lppr','pass':'lppr','
   proc':64}
In [2]: d['login']
Out[2]: 'lppr'
In [3]: for cle in d.keys(): print(cle)
Out[3]:  0
         login
         pass
         proc
```

```
In [1]: for val in d.values(): print(val)
Out[1]:  1
         lppr
         lppr
         64
In [2]: for cle,val in d.items(): print(cle,val)
Out[2]: 0 1
        login lppr
        pass lppr
       proc 64
In [3]: valeur = d.pop('proc') ; valeur
Out[3]: 64
```

■ Implémentation

Une liste en Python est implémentée de manière à ce que certaines opérations soient très efficaces, c'est-à-dire avec une complexité en temps constante. Ces opérations sont par exemple : obtenir la longueur d'une liste, accéder à un élément et modifier un élément. La suppression et l'insertion d'un élément en fin de liste ont aussi un coût que l'on peut considérer comme constant.

Les dictionnaires ne sont pas des séquences ce qui siginifie que les éléments n'ont pas une place, repérée par un indice entier, qui permet d'ordonner l'ensemble pour le parcourir et par la même occasion accéder à un élément particulier en utilisant son indice. Dans un dictionnaire, un élément n'a ni prédécesseur ni successeur. L'objectif est de disposer d'opérations similaires à celles énumérées ci-dessus pour les listes, avec la même efficacité.

Il est important de noter que les éléments d'un dictionnaire n'ont pas à être ordonnés. Donc pour le parcours et l'affichage des clés ou des couples (clé, valeur), il n'y a pas d'ordre prévisible.

Il existe différentes manières d'implémenter un dictionnaire.

Exemple 1

Par exemple, si le dictionnaire **a pour clés des entiers naturels**, on peut utiliser un tableau. Si un indice est égal à une clé, on écrit la valeur correspondante, sinon on écrit une valeur par défaut. Dans ce cas, la complexité en espace est en $O(N)$, où N est la plus grande clé.

Rappelons le principe d'un tableau qui peut être utilisé par exemple pour stocker les adresses en mémoire des éléments d'une liste en Python.
On accède à un élément par son indice. Les adresses des éléments sont enregistrées sur des mots de taille fixe (8 ou 4 octets) de manière séquentielle. Ainsi, si l'on connait l'adresse du début, on peut obtenir l'adresse d'un élément quelconque en calculant par une opération simple sur l'indice la place où se trouve cette adresse. On prévoit un tableau assez grand pour pouvoir ajouter quelques éléments.

Illustration par la figure 1

Exemple 2

Plaçons nous dans le cas général où **les clés ne sont pas toutes des entiers naturels**. Une implémentation classique d'un dictionnaire utilise une **table de hachage**. On dispose d'une **fonction de hachage**, c'est-à-dire une fonction qui à une clé associe un entier appelé **valeur de hachage**

de la clé. Cet entier est utilisé pour calculer l'indice d'un tableau où est placée la clé.
On peut ainsi stocker suivant l'indice calculé pour chaque clé, l'adresse de la clé puis celle de la valeur et stocker la clé et la valeur aux adresses indiquées. Cela revient à utiliser une liste ordonnée suivant les indices calculés pour les clés avec successivement des clés et des valeurs et des places vides. Pour trouver un élément, on calcule l'indice, on trouve l'adresse de la clé et on vérifie. Ceci est exécuté en temps constant. On peut aussi stocker la valeur de hachage.

On doit donc disposer d'une fonction f avec $f(e) = p$ qui calcule la place p d'un élément e. Deux éléments distincts devant se trouver à deux places distinctes, la fonction f doit être bijective. Ce point a une première conséquence : si une clé `x` est mutable, sa place va changer à chaque mutation. Donc on ne peut utiliser comme clé que des objets non mutables, plus précisément non récursivement mutables.

Illustration par la figure 2

Exemple 3

Une autre possibilité est de stocker suivant l'ordre de création, la valeur de hachage calculée pour la clé, l'adresse de la clé puis celle de la valeur et stocker la clé et la valeur aux adresses indiquées. Cela revient à utiliser une liste ordonnée suivant l'ordre de création avec successivement une valeur de hachage, une adresse de clé et une adresse de valeur et ainsi de suite, sans place inoccupée. On complète alors avec un tableau d'octets où chaque octet, à la place dont l'indice est calculé pour une clé, donne la position dans la liste des trois éléments valeur de hachage, clé, valeur. Les octets ne correspondant à aucune valeur de hachage valent -1 pour indiquer une place vide et -2 pour un élément supprimé.
Donnons un exemple où seules les adresses des clés sont représentées.

Considérons le dictionnaire

```
In [1]: {97: "a", 101:"e", 114:"r", 111: " o"}
```

et un tableau `T` de huit octets supposé être la capacité du dictionnaire. ce tableau `T` va donner les positions des clés dans le dictionnaire.
Soit h la fonction de hachage et une clé `c` qui est la $k^{\text{ème}}$ insérée alors si `p` est la position calculée avec `p = h(c) modulo 8` alors `T[p] = k`.
Si l'on ne considère que les clés, le tableau des données contient dans l'ordre $97, 101, 114, 111, 0, 0, 0, 0$.
L'élément 97 a pour indice 0, l'élément 101 a pour indice 1, l'élément 114 a pour indice 2 et l'élément 111 a pour indice 3.
On choisit pour h la fonction identité (ce qui est logique, on verra plus loin).
Alors `h(97) modulo 8` est `97 modulo 8` soit 1. Alors $T[1] = 0$.
Puis `h(101) modulo 8` est `101 modulo 8` soit 5. Alors $T[5] = 1$.
Puis `h(114) modulo 8` est `114 modulo 8` soit 2. Alors $T[2] = 2$.
Enfin, `h(111) modulo 8` est `111 modulo 8` soit 7. Alors $T[7] = 3$.
Le tableau T contient dans l'ordre :

$$-1, 0, 2, -1, -1, 1, -1, 3$$

En effet, $T[0]$ par exemple n'est pas affecté et est une place vide et cet octet ne correspond à aucune donnée. On peut mettre 255 à la place de -1 (car $2^8 = 256$).
Pour accéder à la clé 101, comme sa position `p = h(101) modulo 8` vaut 5, on regarde dans T à la position 5 et on y lit la valeur 1 qui est la position de 101 dans le dictionnaire.

Si l'on supprime par exemple la clé 97, on remplace dans T à l'indice 1 la valeur 0 par -2 (ou 254).

Fonctions de hachage

Comment obtenir une bonne fonction de hachage ? Cette fonction, notée h doit permettre un calcul rapide de l'image d'une clé. Les indices calculés en général par h(clé) modulo n, où n est la taille du tableau T doivent être tous distincts.
En pratique, il est très souvent impossible d'avoir une correspondance bijective entre les clés et les indices. Lorsque le même indice est obtenu pour deux clés, on parle de **collision**.
Pour résumer, on procède en deux étapes :

Étape 1. Une fonction h de hachage associe un entier à chaque clé. Elle code donc les clés. En Python, la fonction est `hash`

```
In [1]: hash(55)
Out[1]: 55
In [2]: hash('pass') ; hash("login")
Out[2]: -8560737331547147328
    -4537507453712144586
In [3]: hash(2**61-1) ; hash(-1)
Out[3]: 0   -2
```

• Si i est un entier différent de -1 et si $-2^{61}+1 < i < 2^{61}-1$ alors `hash(i)` vaut i.
• Si x est un flottant égal à p/q alors $hash(x)$ vaut $(int(p*M/q))\,\%\,M$ avec $M = 2^{61}-1$.

```
In [1]: hash(0.1)
Out[1]: 230584300921369408
In [2]: (int(1*(2**61-1)/10)) % (2**61 -1)
Out[2]: 230584300921369408
```

• Si la clé est une chaîne de caractères la valeur de hachage est calculée avec une part de hasard à chaque nouvelle session et a pour valeur un entier qui s'écrit sur 64 bits.

Étape 2 Il faut ensuite une fonction qui réduise chaque entier pour obtenir un entier appartenant à $[\![0, n-1]\!]$, où n est la taille du tableau (la capacité du dictionnaire).
En Python, on utilise `hash(c) % n`

■ Gestion des collisions

En Python deux objets distincts de même valeur ont la même valeur de hachage :

```
In [1]: a=(3,2)
In [2]: b=(3,2)
In [3]: a is b
Out[3]: False
In [4]: hash(a) == hash(b)
Out[4]: True
```

Mais deux objets dont les valeurs ne vérifient pas le critère d'égalité peuvent aussi avoir la même valeur de hachage :

```
In [1]: hash(0)
Out[1]: 0
In [2]: hash(2**61 -1)
Out[2]: 0
```

Il faut donc trouver des solutions à ce problème de collision.
Pour gérer une collision, on a plusieurs méthodes. En cas de collision, on peut calculer une autre place ou prendre la première place libre qui suit. On commence par prendre un dictionnaire plus long. Une méthode de redimensionnement est utilisée par Python qui prévoit des dictionnaires dont pas plus des deux tiers de la capacité est utilisée. Si l'on atteint les deux tiers de la capacité et qu'un élément doit être ajouté, la capacité du dictionnaire est multipliée par 2.
Par exemple :
de 1 à 5 éléments, la capacité est 8 car $(3/2) \times 5 = 7.5$ et $(3/2) \times 6 = 9$.
Jusqu'à 10 éléments, la capacité est 16 car $(3/2) \times 10 = 15$ et $(3/2) \times 11 = 16.5$
etc.
La capacité n est ainsi une puissance de 2.
Si la taille est $n = 2^p$, alors $h \,\%\, n$ et $h \,\&\, (n-1)$ ont la même valeur.

```
In [1]: 556 % 16
Out[2]: 12
In [2]: 556 & 15
Out[2]: 12
```

En Python, avec un dictionnaire de capacité 8, en cas de collision à une place i, la nouvelle place est calculée avec la formule :

$$i = (5 * i + (h // (2 * *5)) + 1) \,\%\, 8$$

où h a pour valeur `hash(c)` si `c` est la clé.
Si les clés sont des entiers i strictement inférieurs à 32, l'entier h est i et `h // (2**5)` vaut 0 et la formule se réduit à : $i = (5 * i + 1) \,\%\, 8$.
Ainsi si $i = 4$, les 7 valeurs successives sont :

```
In [1]: ind=4
In [2]: for k in range(1,8) :
            ind=(5*ind+1)%8; print(ind)
Out[2]: 5
        2
        3
        0
        1
        6
        7
```

On parcourt ainsi les huit places en sachant que quatre au moins sont disponibles.

Exemple 4

Illustrons un problème de collision. Soit le dictionnaire :

```
In [1]: d = { "i": 19, "n": 14, "f": 6 , " o ": 15}
```

Nous utilisons une liste de longueur 8 pour stocker un dictionnaire de cinq éléments maximum.

```
In [1]: def dic(couples):
          liste = 8 * [None]
          for cle, val in couples:
              liste[hash(cle)%8]=(cle,val)
          return liste
```

```
In [1]: dic((("i",19),("n",14),("f",6), ("o",15)))
    [None, None, ('n', 14), None, None, None, ('o', 15), None]
```

On voit que le couple `("f",6)` par exemple a disparu car `("o",15)` a été le dernier affecté en liste[6] car `hash("f")%8` et `hash("0")%8` sont les mêmes.
On modifie la fonction `dic` pour gérer les collisions.

```
In [1]: def dic(couples):
            liste = 8* [None]
            for cle, val in couples :
                i = hash(cle) % 8
                while liste[i] is not None :
                    i = (5*i +1) % 8
                liste[i] = (cle,val)
            return liste

In[2]:  dic((("i",19),("n",14),("f",6), ("o",15)))
Out[2]: [None,None,('n',14),None,('o', 15),None,('i', 19),('f',
    6)]
```

La place d'indice 7 est occupée après l'insertion de `("f",6)` et la nouvelle place calculée pour insérer `("o",15)` est

```
In [1]: (5*7+1)%8
Out[1]: 4
```

C'est bien l'indice 4.

■ Manipulation

Construction

Pour l'instant, on a crée un dictionnaire en plaçant entre des accolades des couples (clé,valeur) séparés par des virgules, chaque clé et sa valeur associée étant séparées par deux points. On peut construire un dictionnaire par **compréhension** :

```
In [1]: d  = {x : x**2 for x in range(1,5)}
In [2]: d
Out[2]: {1: 1, 2: 4, 3: 9, 4: 16}
```

Ou aussi par **insertion**. On crée un dictionnaire vide puis on insère les éléments par une boucle.

```
In [1]: d = {}
In [2]: for x in range(1,5):
            d[x] = x**2
In [3]: d
Out[3]: {1: 1, 2: 4, 3: 9, 4: 16}
In [4]: d[3]
Out[4]: 9
```

Au passage `d[x]` renvoie la valeur dont `x` est la clé.
De même que les éléments d'une liste peuvent être des listes, les éléments d'un dictionnaire peuvent être des dictionnaires :

```
In [1]: pays = {"France":{"capitale": "Paris",
                          "population": 68014000,
                          "superficie": 643800.0},
                "Portugal" : {"capitale": "Lisbonne",
                              "population": 10302674,
                              "superficie":92300.0},
                "Italie" :{"capitale": "Rome",
                           "population" : 60359546,
                           "superficie" : 301336.0}}
In [2]: pays["France"]["population"]
Out[2]: 68014000
In [3]: pays["France"]
Out[3]: {'capitale': 'Paris', 'population': 68014000, 'superficie
    ': 643800.0}
```

Utilisation

Accès aux éléments

Deux méthodes donnent accès aux clés ou aux valeurs, ce sont les méthodes `keys` et `values`. Et la méthode `items` donne accès à l'ensemble des couples. Le mieux c'est le faire fonctionner.

```
In [1]: d = {"true":"vrai", "false" : "faux", "and" : "et", "or "
    : "ou"}
In [2]: d.keys()
Out[2]: dict_keys(['true', 'false', 'and', 'or '])
In [3]: d.values()
Out[3]: dict_values(['vrai', 'faux', 'et', 'ou'])
In [4]: d.items()
Out[4]: dict_items([('true', 'vrai'), ('false', 'faux'), ('and',
   'et'), ('or ', 'ou')])
```

L'accès à une valeur s'obtient comme avec les listes. La différence est qu'il faut préciser la clé à la place de l'indice.

```
In [1]: d["true"]
Out[1]: 'vrai'
```

Appartenance

le mot clé permet de tester l'appartenance d'une clé à un dictionnaire par l'appartenance d'une valeur.

```
In [1]: "vrai" in d, "and" in d
Out[1]: (False, True)
```

Boucles

On peut itérer avec une boucle `for` sur un dictionnaire, la variable d'itération est une clé.

```
In [1]: cles=[]
In [2]: for obj in d:
            cles.append(obj)

In [3]: cles
Out[3]: ['true', 'false', 'and', 'or ']
```

Il est possible avec une boucle `for` d'itérer sur les clés, sur les valeurs ou sur les couples (clé, valeur) à l'aide des objets `d.keys()`, `d.values()` et `d.items()`

```
In [1]: val = []
In [2]: for newobj in d.values():
            val.append(newobj)
In [3]: val
Out[3]: ['vrai', 'faux', 'et', 'ou']
```

Propriété

On ne peut pas modifier la taille d'un dictionnaire durant une itération sans un message d'erreur.

```
In [1]: dd = {0:0, 1:-5, 2:4}
In [2]: for c in dd:
            dd[c+1] = dd[c] + 1

Traceback (most recent call last):
File "<ipython-input-22-04e6fc6c3560>", line 1, in <module>
    for c in dd:
RuntimeError: dictionary changed size during iteration
```

Pourtant, l'opération se fait quand-même.

```
In [1]: dd
Out[1]: {0: 0, 1: 1, 2: 2}
```

Nombre d'éléments

La fonction `len` renvoie la longueur d'un dictionnaire c'est-à-dire le nombre de couples (clé, valeur).

```
In [1]: len(d),len(dd)
Out[1]: (4, 3)
```

Suppression

Pour supprimer un élément, nous utilisons la fonction `del`

Copie

On use de `d.copy()`. La vigilance s'impose, comme avec les listes, puisque les dictionnaires sont des objets mutables.

```
In [1]: d2 = d.copy()
In [2]: d2
Out[2]: {'true': 'vrai','false': 'faux','and': 'et', 'or ': 'ou'}
In [3]: del d2["false"]
In [4]: d
Out[4]: {'true': 'vrai','false': 'faux', 'and': 'et','or ': 'ou'}
In [5]: d2 = d
In [6]: del d2["false"]
In [7]: d
Out[7]: {'true': 'vrai', 'and': 'et', 'or ': 'ou'}
```

Donc `d` a perdu `"false"` alors qu'on l'a enlevé qu'à `d2`.
Autre exemple où le même pb survient.

```
In [1]: d3 = {"true": ["vrai", "vraie"]}
In [2]: d4 = d3.copy()
In [3]: d4
Out[3]: {'true': ['vrai', 'vraie']}
In [4]: d4["true"][1] = "vrais"
In [5]: d3
Out[5]: {'true': ['vrai', 'vrais']}
In [6]: d4
Out[6]: {'true': ['vrai', 'vrais']}
```

■ Applications dans le langage Python

Les dictionnaires sont des objets au coeur du fonctionnement du langage Python. Lorsqu'on exécute un programme, plusieurs dictionnaires sont mobilisés en arrière plan, même si le programme ne contient aucune définition explicite de dictionnaire.

Espace de noms

Les noms appartenant à un espace de noms quelconque sont les clés d'un dictionnaire. Au démarrage de l'interpréteur, le module `__builtins__` est chargé. Ce module contient tous les objets que nous pouvons utiliser directement dans le programme et le dictionnaire `__builtins__.__dict__` lié à ce module contient tous les identifiants liés à ces objets comme `help,len,True,max` etc. Donnons un exemple (la barre `__` devant et après `builtins` et `dict` est constituée de deux fois le tiret sous la touche 8.

```
In [1]:  __builtins__.__dict__['min'](5,2,3)
Out[1]:  2
```

Le contenu des modules que nous importons est représenté par un dictionnaire.
Si nous importons `math` avec `import math` alors nous obtenons le contenu, comme les fonctions disponibles par `dir(math)`. Ceci revient à demander la liste des clés du dictionnaire `math.__dict__`
Si l'on tape ces deux commandes, la dernière écrit les mêmes fonctions que la première mais avec la syntaxe d'un dictionnaire.

```
In [1]: dir(math)
Out[1]: ['__doc__',
         '__loader__',
         '__name__',
         '__package__',
         '__spec__',
         'acos',
         'acosh',
         'asin',
         'asinh',
          ...
         'tau',
         'trunc']
```

```
In [1]: math.__dict__
Out[1]:  {'__name__': 'math',
   '__doc__': 'This module provides access to the mathematical
      functions\ndefined by the C standard.',
 '__package__': '',
 '__loader__': _frozen_importlib.BuiltinImporter,
 '__spec__': ModuleSpec(name='math', loader=<class '
    _frozen_importlib.BuiltinImporter'>, origin='built-in'),
 'acos': <function math.acos(x, /)>,
 'acosh': <function math.acosh(x, /)>,
 'asin': <function math.asin(x, /)>,
 ...
  'e': 2.718281828459045,
 'tau': 6.283185307179586,
 'inf': inf,
 'nan': nan}
```

Tous les identifiants des variables et des fonctions que nous définissons sont stockés dans un dictionnaire que nous obtenons avec `globals()`. Tapez le et vous verrez apparaître tout notre passé du jour.
Les paramètres d'une fonction et les variables définies dans le corps d'une fonction sont stockés dans un dictionnaire créé à l'exécution de la fonction et supprimé à la fin de l'exécution. On peut observer son contenu en ajoutant dans le corps de la fonction `print(locals()`

```
In [1]: def f(x,y):
          z = x+y
          print(locals())
In [2]: f(3,4)
Out[2]: {'x': 3, 'y': 4, 'z': 7}
```

Notion de portée
Définissons une fonction `abs` dans un programme en cours. Le nom `abs` appartient alors à l'espace de nom local de ce programme. Donc si l'on utilise `abs` dans ce programme, c'est l'espace local qui est examiné en premier et donc cette fonction est utilisée en premier. Mais la fonction `abs` connue (la valeur absolue) du module `__builtins__` n'a pas été modifiée.

```
In [1]:def abs(x):
          return x
In [2]: abs(-5)
Out[2]: -5
In [3]:__builtins__.abs(-5)
Out[3]:  5
```

Notion d'affectation
Écrire par exemple $x = 300$ et `globals()["x"]`$= 300$ est équivalent. On ajoute au dictionnaire la clé `x` associée à la valeur 300. Autrement dit, on ajoute au dictionnaire le couple (identité ou adresse de l'objet de type `str` et de valeur $'x'$, identité de l'objet de type `int` et de valeur `300`). Si l'on écrit ensuite $x = 500$, on remplace l'identité de l'objet de type `int` et de valeur 300 par celui de valeur 500. Le premier objet n'est pas modifié mais `x` est lié à un nouvel objet.

Exécution d'une fonction
Tapons :

```
In [1]: def f(f):
          return 2*f
In [2]: f(5)
Out[2]: 10
```

La clé `f` (le nom de la fonction) est ajoutée au dictionnaire `globals()` et liée à l'objet de type `function` et de valeur le code de la fonction.
Lors de l'exécution, un dictionnaire local lié à la fonction est créé. Ce dictionnaire contient pendant l'exécution le nom `f` (celui du paramètre), associé à la valeur 5.

```
In [1]: def f(f):
          return 2*f, locals()
In [2]: f(5)
Out[2]: (10, {'f': 5})
```

Pour finir, les dictionnaires sont aussi utilisés dans des opérations de comptage et pour l'implémentation des graphes. Ils sont utilisés aussi pour la modélisation des jeux ou le traitement de données en tables etc.

Chapitre 3

Graphs

■ Objectifs

Les incontournables :

savoir

Résumé de cours

VOIR FEUILLES PDF

diagbox

Chapitre 4

Database

■ Objectifs

Les incontournables :

savoir

Résumé de cours

■ Organisation des données. Modèle relationnel

I. Introduction. Type de données

La quantité de données utilisables dans notre vie est gigantesque et se mesure en gigaoctet $= 10^6$ octets, téraoctet $= 10^9$ octets ou pétaoctet $= 10^{12}$ octets. L'utilisation de bases de données relationnelles permet le stockage de données ainsi que l'organisation des mises à jour et des consultations par un grand nombre d'utilisateurs.

Parlons un peu du type de données. Les formats peuvent être différents : des nombres, des textes, des images, des vidéos. Pour les stocker, il faut choisir comment les coder afin de les enregistrer sur un support numérique, donc définir un type de données par exemple des types numériques, des types textes, des types dates.
Chaque système de gestion des bases de données, ou SGBD, présente quelques différences sur les types disponibles mais en propose suffisamment afin de pouvoir utiliser dans chaque cas le plus économique en besoin au niveau de la mémoire et limiter ainsi la taille de la base de données.
Les **types numériques** se décomposent en **types entiers** ou non et dans ce cas, on se limite au **type flottant**. Une représentation des nombres en machine a été étudié en TSI1 avec des mots de taille fixe $(8, 16, 32, 64$ bits). Les SGBD peuvent utiliser des représentations différentes pour certains types de nombres afin de pouvoir les stocker de manière exacte et effectuer des calculs exacts. Par exemple, pour un nombre décimal, on peut coder en binaire chaque chiffre de son écriture. Avec des mots de même taille pour chaque chiffre, nous aurons besoin de 4 bits par mots, en effet le plus grand nombre est 9 et s'écrit 1001 en binaire et par exemple $68,3$ est codé 0110 1000 0011.
Pour les textes également, on se restreint en général à un **type chaîne de caractères**.
Concernant les dates, il existe des types précis mais on se limite à des chaînes de caractères du type `aaaa-mm-jj`, ce qui est suffisant pour pouvoir effectuer des comparaisons avec l'ordre lexicographique. Une autre façon est d'utiliser des nombres entiers représentant le nombre de secondes écoulées depuis l'origine (souvent 1970-01-01) et la date souhaitée.

II. Schéma relationnel. Vocabulaire

Des notions mathématiques comme le langage ensembliste, les relations, et la logique sont à la base du modèle.

Considérons par exemple dans un lycée des données concernant les él'èves, les professeurs et les administratifs. On reprèsente l'ensemble des données, comme le nom, le prénom, la date de naissance, le numéro de téléphone de chaque personne dans un tableau à deux dimensions appelé **relation ou table**.

Une colonne du tableau est appelée **attribut ou champ** par exemple la colonne nommée `nom`. L'ensemble des colonnes caractérisent la relation. On peut noter l'analogie avec une loi conjointe d'un couple (X, Y) de variables aléatoires discrètes finies. Chaque personne correspondrait à une valeur de X et chaque attribut correspondrait à une valeur de Y.

Chaque ligne de la table ou relation est un **enregistrement**, on dit aussi une **instance** ou `tuple` ou encore `n-uplet` (à condition qu'il y ait n colonnes). Chacun de ces tuples est unique et caractérise donc une personne donnée. On part du principe que c'est impossible d'avoir le même nom, le même prénom, la même date de naissance et le même numéro de phone pour deux personnes différentes (à moins qu'il y ait eu piratage des données mais on suppose être dans un monde honnête).

Chaque valeur dans une colonne a un type unique pour cette colonne avec des valeurs limites (par exemple 8 caractères max etc.). On dit que les valeurs permises dans une colonne appartiennent à un **domaine**.
L'obligation d'appartenance d'une valeur à un domaine s'appelle une **contrainte d'intégrité**. Plus précisément, il s'agit ici d'une **intégrité de domaine**.
Le **produit cartésien** des domaines $D_1, D_2, ..., D_n$ noté $D_1 \times D_2 \times ... \times D_n$ est l'ensemble des `n-uplets` $(v_1, v_2, ..., v_n)$ tels que, pour tout i, $v_i \in D_i$.
Une relation est donc un sous-ensemble du produit cartésien de domaines.
Exemple : supposons les domaines $D_1 = \{a, b, c\}$ et $D_2 = \{x, y\}$.

D_1 \ D_2	x	y
a	(a, x)	(a, y)
b	(b, x)	(b, y)
c	(c, x)	(c, y)

Une relation est par exemple $R = \{(a, x), (b, x), (c, y)\}$ ou encore

C_1	C_2
a	x
b	x
c	y

, où la colonne C_1 est constituée de valeurs du domaine D_1 et la colonne C_2 est constituée de valeurs du domaine D_2.

III. Notion de clé dans un schéma relationnel

• Dans toute relation, un attribut (c'est-à-dire une colonne) ou un groupe d'attributs, doit permettre d'identifier de manière unique un enregistrement (c'est-à-dire une ligne).
On l'appelle une **clé candidate**
Par exemple, si l'on reprend l'exemple plus haut, la colonne C_2 ne peut pas être une clé car il y a deux fois x et par contre la colonne C_1 est bien une clé candidate.

• S'il y a plusieurs clés candidates, on en privilégie une, nommée la **clé primaire**. Choisir une clé primaire est une contrainte d'intégrité imposée à la relation. Dans l'exemple précédent, on n'a pas le choix, la colonne C_1 est la seule clé primaire possible.

• Certaines relations peuvent contenir un attribut dont les valeurs permises sont uniquement celles prises par la clé primaire d'une autre relation. Cet attribut est alors appelé **clé étrangère**. Une clé étrangère est une **contrainte d'intégrité référentielle** (elle fait référence à un attribut d'une autre table). Elle relie deux tables de manière cohérente.

Les clés primaires et les clés étrangères permettent de spécifier des contraintes dans les bases de données et d'effectuer des opérations de jointures entre les tables, opérations présentées dans le chapitre suivant.

La présence d'une clé primaire est capitale pour assurer que les lignes sont toutes distinctes. Une contrainte d'unicité est donc imposée : chaque valeur prise par l'attribut ou le groupe d'attributs, sur lequel est posé une contrainte de clé primaire est unique.

IV. Exemples détaillés

Reprenons l'exemple d'un établissement scolaire et considérons une base de données qui sert à stocker des informations utiles pour l'établissement scolaire. Plusieurs relations sont possibles.

Pour représenter le concept d'élève, nous utilisons une relation **Élève**. Les attributs (ou les colonnes) de cette relation par exemple sont **INE** (numéro d'identification), **Nom**, **Prénom**, **Âge**, **Sexe**, **Classe**. Chaque élève est représenté par un enregistrement (une ligne).

Pour représenter le concept de classe, nous utilisons une relation **Classe** avec pour attributs **Nom**, **Effectif**, **Salle**.
Il y a bien entendu des liens entre l'ensemble des élèves et l'ensemble des classes.
Un élève donné fait parti d'une classe particulière.
Ce lien fonctionne dans les deux sens : une classe contient plusieurs élèves.

Les données peuvent être modélisées d'une autre manière.
On ajoute une relation **Personne** pour représenter les personnes appartenant à la communauté scolaire. Si cette relation dispose d'attributs comme le nom, le prénom, etc., on peut supprimer ces atrributs de la relation **Élève** et ne garder que le numéro INE et la classe. Pour des relations comme **Professeurs**, **Administratif**, **Service**, on peut remplacer l'attribut **INE** par un attribut ayant pour valeur le numéro de sécurité sociale par exemple.

Un schéma de relation se présente sous cette forme :

Relation(attribut 1, attribut 2, ..., attribut N)
Clé primaire : **attribut 1**
Clé étrangère : **attribut N** en référence à la clé primaire d'une autre relation.

Par exemple :
Élève (INE, Nom, Prénom, Âge, Sexe, Classe)
Les attributs **INE** et **Âge** sont de type entier, les autres de type chaîne de caractères.
Clé primaire : **INE**
Clé étrangère : **Classe** en référence à la clé primaire d'une autre relation.

Ce schéma peut aussi se noter :
Élève (INE, Nom, Prénom, Àge, Sexe, #Classe)
L'attribut sur lequel est posé une contrainte de clé primaire est souligné et l'attribut sur lequel est posé une contrainte de clé étrangère est précédé du caractère #

On peut représenter le schéma ainsi :

INE
Nom
Prénom
Âge
Sexe
#Classe

Conception d'une relation

Si à une valeur d'un attribut A correspond une et une seule valeur d'un attribut B, on dit que l'attribut B est en **dépendance fonctionnelle** de l'attribut A.
Par exemple, à un identifiant d'élève correspond un unique nom d'élève.

De manière générale, tous les attributs ne contiennent qu'une seule information et sont en dépendance fonctionnelle de la clé primaire.
Si une clé primaire est constituée de plusieurs attributs, les autres attributs sont en dépendance fonctionnelle de l'intégrabilité de la clé primaire et pas seulement d'une partie de celle-ci. De plus, tous les attributs sont en d¡'ependance fonctionnelle uniquement de la clé primaire.

Organisation des données
Les données peuvent être représentées dans un tableau comme ci-après.

INE	NOM	Prénom	Âge	Sexe	Classe
1	Parker	Bonnie			2
2	Barrow	Clyde			4
3	Laurel	Stanley			1
...					
50	Hardy	Oliver			
...					

Remarque : Ce tableau ne doit pas être interprété comme celui obtenu dans un tableur.
Un tableur permet d'effectuer diverses actions sur ces données : des calculs, des schémas, des graphiques.
Ici l'attribut **INE** est un identifiant numérique qui permet d'identifier chaque ligne de manière unique. Cet attribut **INE** peut donc être défini comme clé primaire.

Pour la deuxième relation nommée **Classe**, on a le tableau :

Nom	Effectif	Salle
PSI1	45	$B307$
PSI2	42	C206
PC	38	$B311$
TSI2	29	C204
PT	41	B309
...	...	...

Pour distinguer les attributs **Nom** des relations **Élève** et **Classe**, nous pouvons utiliser la notation pointée comme **Élève.Nom** et **Classe.Nom**
L'attribut **Nom** de la relation **Classe** est une clé primaire. Nous souhaitons poser une contrainte de clé étrantère sur l'attribut **Classe** de la relation **Élève**.
Or, pour effectuer des recherches et pour procéder à d'éventuelles modifications, il est plus efficace d'utiliser un nombre entier plutôt qu'une chaîne de caractères. On ajoute donc une colonne supplémentaire, un nouvel attribut nommé **Id**. Il s'agit d'une clé artificielle ou clé de substitution (surrogate key) qui permet d'identifier une ligne quelconque.

Id	Nom	Effectif	Salle
1	PSI1	45	$B307$
2	PSI2	42	C206
3	PC	38	$B311$
4	TSI2	29	C204
5	PT	41	B309
...	...	...	

Les valeurs de la clé étrangère **Classe** de la relation **Élève** sont alors des valeurs de la clé primaire **Id** de la relation **Classe**.

Avec cette représentation, les modifications sont beaucoup plus simple à gérer que que si toutes les données étaient dans un tableau unique. Ainsi si par exemple, les élèves de la classe PSI1 changent de salle, il n'y a qu'une seule valeur à modifier dans le tableau **Classe**.

La notion de contrainte est capitale. Supposons que dans la table **Classe** un enregistrement, ou une ligne, contient la valeur **7** pour l'identifiant **Id** et que cette valeur **7** est celle d'un ou plusieurs enregistrements pour l'attribut **Classe** dans la table **Élève**.
Si par erreur, la suppression dans la table **Classe** de l'enregistrement avec l'identifiant **7** est demandée, cela va provoquer une erreur car cela conduirait à une violation de la contrainte de référence ou contrainte de clé étrangère, de **Élève.Classe** vers **Classe.Id**
En pratique, on peut contourner ce problème en imposant une suppression en cascade. Dans ce cas toutes les lignes concernées de la table **Élève** sont supprimées. On peut aussi imposer que les lignes concernées ne soient pas supprimées mais qu'une valeur par défaut remplace la valeur **7**.

V. Bases de données relationnelles et SGBD

Un SGBD est un système de gestion de base de données. L'informaticien Charles Bachman est le premier concepteur d'un SGBD moderne.

Pour accéder à une base de données, on utilise donc un SGBD. Ce peut être un composant logiciel, comme SQLite qui est une bibliothéque avec une interface directement intégrable dans un programme. D'autres SGBD comme MySQL et PostgreSGL fonctionnent avec un serveur.

Il y a deux types de systèmes de gestion de base de données :

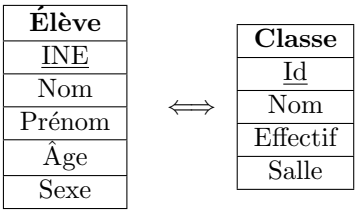
- des systèmes libres comme MySQL, PostgreSQL, SQLite ;
- des systèmes propriétaires comme Oracle Database ou encore SQL Server de Microsoft ou encore Access de la suite Microsoft Office

Fonction d'un SGBD

Le système doit assurer la persistance des données. Cela consiste à garder en mémoire des versions antérieures lorsque des modifications sont effectuées.
Le système doit gérer des accès concurrents et la sécurité.
Il gère les droits et privilèges d'un utilisateur.
L'un des attendus principaux est un accès aux données efficace. Cette efficacité est conditionnée par la manière de stocker les différentes clés primaires. L'utilisation d'une clé de substitution (surrogate key) apporte une simplification pour la recherche. On peut utiliser plutôt qu'une structure linéaire (pour la recherche des données) une structure d'arbre, en particulier un arbre binaire (Btree en anglais). C'est la structure utilisée par les SGBD. Enfin, les SGBD utilisent aussi des tables de hachage pour classifier les données.

VI. Entités-Associations

Reprenons l'exemple plus haut. Un élément qui joue un rôle dans une relation par exemple **Élèves** est appelé **entité**. C'est ici un élève particulier qui possède des attributs servant à le décrire. Des liens ou relations existent entre **Élève** et **Classe**. La relation entre deux entités de ces schémas s'appelle une **association**. Le type d'association est décrit par un titre comme **Appartenir** puisqu'un élève appartient à une classe ou **Contenir** puisqu'une classe contient un élève.



Le lien entre une entité et une association est caractérisé par un couple qui indique le nombre minimum de fois et le nombre maximum de fois qu'une entité peut participer à l'association.

Le couple est **1-1** si chaque entité participe exactement une fois à la relation.
Le couple est **1-n** si chaque entité participe au moins une fois à la relation.
Le couple est **0-n** si certaines entités ne participent pas une fois à la relation.

Pour notre exemple, on a le couple **1-1** entre **Élève** et **Appartenir** signifie que chaque élève appartient à exactement une classe. On a le couple **1-n** entre **Appartenir** et **Classe** qui signifie que chaque classe contient au moins un élève.
On peut présenter le modèle entité-association en parlant d'associations **1-***, **1-1**, ***-*** entre deux entités. Par exemple, nous avons une association **1-*** entre **Élève** et **Classe**.

■ Langage SQL (Structured Query Language)

I. Introduction

Le langage SQL (Structured Query Language = langage de requêtes structurées) permet aux utilisateurs de communiquer avec une base de données. C'est un langage déclaratif qui décrit ce que l'on souhaite faire avec une requête : une demande adressée à un SGBD.
Plus précisément, une instruction en SQL traduit une interrogation, une mise à jour, une insertion, une suppressio, une création etc.

Les instructions en SQL ont une syntaxe assez simple, plus libre qu'en Python.

Déjà contrairement à Python, SQL ne fait pas la différence entre majuscule et minuscule. Ceci dit, la convention veut que les mots prédéfinis du langage soient en majuscule (par exemple SELECT qui est le premier de tous) et les noms de tables ou de colonnes (de la table) soient en minuscule (à part éventuellement la première lettre). Les espaces, les retours à la ligne et l'indentation n'ont aucune valeur syntaxique mais participent à la clarté de la requête (donc à utiliser dans une copie en particulier le jour du concours).

L'écriture des valeurs doivent respecter des règles (liées aux types ou domaines) : un flottant s'écrit comme en Python avec un point pour la virgule, une chaîne de caractères s'écrit comme en Python avec des guillemets ou des apostrophes et par exemple la chaîne de caractères `'Ta␣gueule!'` est différent de `'ta␣gueule!'`.
Enfin, on termine une requête ou instruction par un point-virgule qui sert de séparateur entre deux instructions. Ce n'est pas obligatoire s'il n'y a qu'une instruction. Là aussi cela rappelle Python où dans une ligne, on peut mettre plusieurs instructions indépendantes separées par un point-virgule.

II. Les requêtes d'interrogation dans une table

Nous allons nous intéresser à cinq cas : **projection** (c'est-à-dire récupérer et sortir de la table des colonnes ou attributs), **champ calculé** (c'est-à-dire faire des calculs dans des colonnes donnés), **renommage de colonne**, **tri** (c'est-à-dire ordonner les lignes de la table dans un sens imposé), **selection de lignes**.

1. La projection

Reprenons un des exemples plus haut avec la table **Classe**

Id	Nom	Effectif	Salle
1	PSI1	45	$B307$
2	PSI2	42	C206
3	PC	38	$B311$
4	TSI2	29	C204
5	PT	41	B309
...	...	...	

Tapons **SELECT Nom FROM Classe**, on obtient :

Nom
PSI1
PSI2
PC
TSI2
PT
...

On peut demander le contenu de plusieurs colonnes en séparant les noms par des virgules :

Tapons **SELECT Nom, Salle FROM Classe**, on obtient :

Nom	Salle
PSI1	B307
PSI2	C206
PC	B311
TSI2	C204
PT	B309
...	...

Pour obtenir toutes les colonnes, on utilise le symbole *****

Retenir les syntaxes suivantes qui permettent de récupérer une ou plusieurs colonnes. :

SELECT Attribut1 FROM Table récupère la colonne **Attribut1** de **Table**

SELECT Attribut1 , Attribut2 FROM Table récupère les colonnes **Attribut1** et **Attribut2** de **Table**

SELECT * FROM Table récupère toutes les colonnes de **Table**

SELECT DISTINCT Attribut1 FROM Table récupère la colonne **Attribut1** de **Table** en supprimant les doublons.

Remarque La commande **SELECT ... FROM ...** s'appelle une projection car on peut faire l'analogie avec une projection vectorielle, l'espace vectoriel E est la table et l'image de la projection est constitué des colonnes récupérées qu'on peut voir comme un sous-espace vectoriel de E.

2. Le champ calculé

Un champ est un synonyme d'attribut ou de colonne ici. Le mieux est un exemple.

Relation : **Notes (Id, Maths, Physique, Informatique, SI)** avec **Id** pour clé primaire.
Les autres champs sont de type flottant et représentent des notes.

Supposons que la commande **SELECT Maths FROM Notes** renvoie :

Maths
14
12.5
8

Alors la commande **SELECT Maths * 3 FROM Notes** renvoie :

Maths*3
42
37.5
24

On peut aussi faire des opérations entre plusieurs champs (à condition qu'ils soient de type flottant)

La requête **SELECT Id, (Maths + Physique) / 2 FROM Notes** renvoie deux colonnes de **Notes**, la première est la colonne des identifiants des élèves et la seconde fait la moyenne ligne par ligne des notes de maths et de physique de chaque élève défini par son identifiant.

La requête **SELECT Id, 2*3 FROM Notes ;** renvoie deux colonnes, la première est la colonne **Id** de **Notes** et la seconde colonne (qui conserve le même nombre de lignes que la colonne **Id** est constituée de la valeur 6 sur chaque ligne.

Remarque : Ces opérations sur les champs utilisent les opérateurs classiques +, -, * et /
On verra plus loin qu'on peut les prolonger par les fonctions d'agrégation.

3. Renommage des colonnes

Dans l'exemple de la relation **Notes**, on a crée un nouveau champ appelé **Maths*3**, on peut décider de lui donner un autre nom par exemple, on décide de l'appeler **Total_Maths** et de même la colonne **(Maths + Physique)/2**, on veut la nommer **Moyenne**.

On tape alors les requêtes SQL :

SELECT Maths * 3 AS Total_Maths FROM Notes ;
SELECT (Maths + Physique) / 2 AS Moyenne FROM Notes ;

On a donc introduit une nouvelle fonction prédéfinie **AS** qui ressemble à `as` de Python.

4. Le tri

L'opération **SELECT** renvoie des colonnes dont les lignes respectent l'ordre initial. Ainsi une table est un ensemble d'enregistrements qui ne sont pas ordonné à moins que l'on impose un ordre. C'est l'objet de la suite. Le tri est une opération qui permet de classer les enregistrements selon un ou plusieurs critères. On utilise la commande prédéfinie **ORDER BY** qui a deux arguments le premier est **ASC** pour un tri dans l'ordre croissant et le second est **DESC** pour un tri dans l'ordre décroissant. Par défaut l'ordre est croissant et donc **ASC** n'est pas obligatoire (maintenant on peut le mettre pour la clarté). Enfin, l'ordre pris en compte est celui des lignes d'une ou de plusieurs colonnes données. Encore une fois, c'est beaucoup plus clair avec notre exemple que nous vous rappelons.

Relation : **Notes (Id, Maths, Physique, Informatique, SI)** avec **Id** pour clé primaire.
Les autres champs sont de type flottant et représentent des notes.

Nous désirons obtenir les identifiants des élèves classés suivant l'ordre décroissant des notes de maths. On tapera :

SELECT Id FROM Notes ORDER BY Maths DESC

On remarquera l'emplacement des différentes commandes, c'est une chronologie de syntaxe à respecter.

Nous désirons obtenir les identifiants des élèves classés suivant l'ordre croissant des notes de maths. On tapera :

SELECT Id FROM Notes ORDER BY Maths ;

Ici par défaut, c'est bien **ASC**

Plus subtil : nous désirons obtenir les identifiants des élèves classés suivant l'ordre croissant des notes de maths et s'il y a égalité entre deux notes de maths, on classera selon l'ordre croissant des notes de physique. On tapera :

SELECT Id FROM Notes ORDER BY Maths, Physique ;

C'est ce que l'on appelle l'ordre lexicographique (qui permet de créer un ordre dans $\mathbb{C}$ au passage mais c'est un autre sujet). On peut pousser ce process jusqu'au bout.

SELECT Id FROM Notes ORDER BY Maths, Physique, Informatique, SI, Id ;

5. La selection de lignes de la table

Avec la fonction **SELECT**, on obtient toute une colonne.

Supposons que l'on ne désire que certaines lignes de cette colonne.

α) Supposons d'abord que l'on veuille la partie d'une colonne qui va d'une ligne donnée à une autre ligne donnée. Ce sont les clauses **LIMIT** et **OFFSET** qui permettent de préciser les numéros des lignes voulues. On place ces fonctions à la fin de la requête SQL. Attention comme en Python on commence l'indice de la ligne par **0** et ainsi la commande **OFFSET 0** signifie à partir de la première ligne et par exemple la commande **LIMIT 5 OFFSET 8** signifie que l'on prend 5 lignes à partir de la neuvième ligne. Tapons la requête SQL :

SELECT Id, Maths FROM Notes LIMIT 2 OFFSET 1 ;

Elle renvoie la deuxième et troisième ligne des colonnes **Id** et **Maths**

On peut demander des lignes triées. Par exemple :

SELECT Maths FROM Notes ORDER BY Physique LIMIT 2 OFFSET 1 ;

Cette requête renvoie la deuxième et troisième ligne de la colonne **Maths** dont les lignes ont été préalablement triées selon les notes de physique.
On comprend que **ORDER BY** doit être logiquement avant **LIMIT OFFSET**

β) Supposons que l'on veuille non pas un certain nombre de lignes consécutives mais choisies selon un critère précis, on utilise alors la clause **WHERE** Donnons un exemple.
Considérons la relation **table** suivante.

x	y	z
-1	-1	-2
-1	0	-1
-1	1	0
0	-1	1
0	0	0
0	1	1
1	-1	-1
1	0	0
1	1	2

Tapons **SELECT x , y FROM Table WHERE z $>$ 0** On obtient :

x	y
0	-1
0	1
1	1

Il faut retenir que **WHERE** est suivi d'une expression booléenne qui est formée de :

de noms de colonnes ;

d'opérateurs mathématiques $+$, $-$, $*$, $/$;

d'opérations de comparaison $<$, $<=$, $>=$, $<>$;

d'opérateurs logiques **NOT**, **AND**, **OR**

Les opérateurs **AND** et **OR** traduisent une intersection et une réunion respectivement.
Par exemple supposons le schéma relationnel **Table** constitué de trois colonnes **col1**, **col2** et **col3** toutes de type flottant. Il s'agit de renvoyer cette table en ne gardant que les lignes telles que sur chaque ligne de **Table**, on ait la contrainte :

2*col1 + col2 <100 et (col 3 <> 1000 ou col3 = col1)

On doit donc taper :

SELECT col1, col2, col 3 FROM Table WHERE (2*col1 + col2 < 100) AND (col3<>1000 OR col3 = col1) ;

Remarque : attention, l'égalité est = et non == comme en Python et la négation de l'égalité est <> et non ! = comme en Python.

Autre exemple, prenons la relation **Élève (Id, Nom, Prénom, Adresse, Ville, Tel)** avec la clé primaire **Id** et écrivons une requête pour avoir les noms et prénoms des élèves qui habitent à Nice. On tapera :

SELECT Nom, Prénom FROM Élève WHERE Ville = "Nice" ;

Écrivons une requête pour avoir les noms et prénoms des élèves qui habitent à Nice ou à Cannes.

On tapera :

SELECT Nom, Prénom FROM Élève WHERE Ville = "Nice" OR Ville = "Cannes" ;

Écrivons une requête pour avoir le nom et le numéro de téléphone des élèves dont la première lettre du nom est comprise entre 'A' et 'M' bornes comprises.

On tapera :

SELECT Nom, Tel FROM Élève WHERE Nom >= "A" AND Nom < "N" ;

On admettra que l'ordre lexicographique est valable pour les chaînes de caractères.

γ) On peut faire des opérations logiques appliquées à des morceaux de tables : **UNION** pour l'union, **INTERSECT** pour l'intersection et **EXCEPT** pour la différence. Les morceaux de tables doivent être compatibles (même nombre et mêmes types d'attributs).

Exemple : Soit la relation **Élève (Id, Nom, Prénom, Adresse, Ville, Tel, #Numprof)** qui donne la liste des élèves ayant cours une heure précise.

La clé étrangère **Numprof** permet de savoir quel est le professeur de l'élève à cette ligne. On veut récupérer le nom et le prénom des élèves qui ont cours avec le professeur ayant pour **Numprof** la valeur **7** et dont l'identifiant **Id** est inférieur à **30**. On tapera :

SELECT Nom, Prénom FROM Élève WHERE Numprof = 7
INTERSECT
SELECT Nom , Prénom FROM Élève WHERE Id < 30 ;

Comme les opérations portent sur des ensembles, les doublons sont automatiquement éliminés.

III. Fonctions d'agrégation

Nous nous bornons aux cinq principales : **COUNT, SUM, AVG, MIN, MAX**

Ce sont des fonctions qui s'appliquent à des colonnes d'un schéma relationnel et qui se placent juste après **SELECT** dans la requête SQL.
Elles permettent de faire des calculs statistiques basiques.

Prenons pour tout le paragraphe le schéma suivant :

Météo (Id, Lieu, Année, Mois, Tmin, Tmax, Précipitations)

1. La fonction COUNT

Cette fonction sert à compter des enregistrements, donc les colonnes ne sont pas nécessairement de type flottant ou int pour l'utiliser.

Par exemple, écrivons la requête SQL pour obtenir le nombre de mois en 2021 pour lesquels les précipitations sont supérieures à 20 mm à Antibes.

Ce nombre de mois doit être renommé **Nombre_de_mois** dans la requête.

SELECT COUNT(Mois) AS Nombre_de_mois FROM Météo WHERE Lieu = "Antibes" AND Année = 2021 AND Precipitations > 20 ;

Remarque : On peut remplacer **COUNT(Mois)** par **COUNT(*)** car il s'agit simplement de compter le nombre de lignes qui vérifient la condition **WHERE**

2. La fonction SUM

Cette fonction s'applique à une colonne de valeurs numériques et fait la somme des lignes sélectionnées de cette colonne.

Par exemple, écrivons la requête SQL pour obtenir la quantité de pluies tombée en 2021 à Antibes.
Cette quantité de pluie doit être renommé **Précipations_annuelles** dans la requête.

SELECT SUM(Précipitations) AS Précipations_annuelles FROM Météo WHERE Lieu = "Antibes" AND Année = 2021 ;

3. La fonction AVG

Cette fonction sert à calculer la moyenne (en anglais average) des valeurs d'un champ numérique.

Par exemple, écrivons la requête SQL pour obtenir la moyenne des températures maximales en 2021 à Antibes.
Cette moyenne doit être renommée **Moyenne_Tmax** dans la requête.

SELECT AVG(Tmax) AS Moyenne_Tmax FROM Météo WHERE Lieu = "Antibes" AND Année = 2021

4. Les fonctions MIN et MAX

Elles permettent d'obtenir la valeur minimale et la valeur maximale d'un champ numérique.

Par exemple, écrivons la requête SQL pour obtenir la valeur maximale des précipitations en 2021 à Antibes.
Cette valeur maximale doit être renommé **Max_précipitations** dans la requête.

SELECT MAX(Précipitations) AS Max_précipitations FROM Météo WHERE Lieu = "Antibes" AND Année = 2021 ;

IV. Jointure

La jointure permet de mettre en relation des tables, par l'intermédiaire des liens qui existent en particulier entre la clé primaire de l'une et la clé étrangère de l'autre.
La jointure est une opération de selection car elle permet de ne retenir par exemple que les enregistrements pour lesquels la valeur de la clé primaire d'une table est égal à la valeur de la clé étrangère d'une autre table.

Au niveau de la syntaxe, l'opérateur **JOIN** exprime la jointure entre deux tables et la clause **ON** permet de préciser le critère de jointure. Par exemple :

table1 JOIN table2 ON table1.attribut_a = table2.attribut_b

Attention, les tables concernées peuvent contenir des colonnes dont les noms sont identiques. On évite de se mélanger les pinceaux en utilisant la syntaxe **table.colonne** qu'on conseille d'utiliser systématiquement pour la clarté visuelle de la requête.

Exemple : Utilisons les deux tables suivantes.

Relation 1 : **Professeur (Id, Nom, Prénom, Tel, Salle)**

Relation 2 : **Élève (Id, Nom, Prénom, Adresse, Ville, Tel, #Numprof)**

On constate que **Numprof** est une clé étrangère de la table **Élève** en référence avec la clé primaire **Id** de la table **Professeur**

La requête est d'obtenir le nom des élèves qui ont cours avec Madame Hopper (connue sous le nom de "Hopper" dans la colonne **Nom** de **Professeur** et la salle où a lieu ce cours.
Comme plusieurs colonnes ont la même dénomination dans les deux tables, on va utiliser aussi la syntaxe **table.colonne**

On tape la requête SQL :

SELECT Élève.Nom , Professeur.Salle FROM Élève JOIN Professeur ON Élève.Numprof = Professeur.Id WHERE Professeur.Nom = "Hopper" ;

Cet exemple est assez complet.
Notez que l'on selectionne ici une colonne dans chaque table avec **SELECT** puis on signale que l'on a **JOIN** les deux tables et que l'on assimile les colonnes **Numprof** d'une table et **Id** de l'autre avec **ON** et enfin **WHERE** permet de ne garder que les lignes concernant Madame Hopper.

V. Renommage de tables

Le renommage permet de raccourcir et de simplifier l'écriture d'une requête. On renomme les tables et on utilise le nouveau nom (l'alias) devant les champs dans la requête. On use du mot **AS** (comme pour le renommage de colonnes). Il est facultatif mais pour la clarté de l'écriture autant l'utiliser.
Par exemple si l'on veut que **Table** devienne son alias **T**, on tape **Table AS T** ou **Table T**
On peut créer l'alias au moment de **JOIN**. Par exemple :

Table1 AS t1 JOIN Table2 As t2 ON t1.attribut_a = t2.attribut_b

Reprenons l'exemple précédent des aventures de Madame Hopper.

On décide de renommer **e** la table **Élève** et de renommer **p** la table **Professeur** et on veut obtenir le nom des élèves qui ont cours avec Madame Hopper et la salle où a lieu ce cours. Et tout ça dans la même requête.

SELECT e.Nom , p.Salle FROM Élève AS e JOIN Professeur AS p ON e.Numprof = p.Id WHERE p.Nom = "Hopper" ;

Remarque : On remarque que le choix des tables arrivant en premier dans l'ordre d'exécution, les alias sont crées au niveau de **JOIN** et s'utilisent donc au moment de **SELECT** qui arrive ensuite dans l'exécution. Attention, l'ordre d'écriture donne l'impression que l'on utilise les alias avant de les définir. C'est normal. C'est simplement que l'ordre d'écriture n'est pas l'ordre d'exécution.

VI. Groupement et filtrage

Les fonctions d'agrégation peuvent être utilisées sur un groupe de données.
On utilise deux clauses **GROUP BY** et **HAVING**

1. La clause GROUP BY

Le mieux c'est un exemple.

Considérons le schéma relationnel : **Frais (Id, nom, prix, #id_marque)** qui donne des données sur une liste de produits frais (leur identification (clé primaire), leur nom, leur prix et leur marque (clé étrangère), les champs **Id** et **id_marque** sont de type entier, le champ **nom** est une chaîne de caractères et le champ **prix** est de type flottant.

On a par ailleurs le schéma relationnel **Marque (id, nom)** avec **id** pour clé primaire qui est un tableau donnant les identifiants des marques et les noms des marques.

On va écrire une requête donnant pour chaque marque le nom de la marque et le nombre de produits en vente (qui sont sur le schéma relationnel **Frais** bien entendu).

Comme **nom** apparaît pour **Marque** et **Frais**, on va utiliser les notations **Table.attribut**
De plus, comme deux tables entrent en jeu on va sortir un **JOIN** comme on dit.

La requête SQL est :

```
SELECT Marque.nom, COUNT(Frais.id) FROM Frais JOIN Marque
ON id_marque = Marque.id GROUP BY Marque.nom;
```

Ainsi **GROUP BY** précise le (ou les) attribut(s) utilisés pour le groupement. Ici c'est l'attribut **nom** de **Marque**

On va écrire une requête donnant pour chaque marque le nom de la marque et le prix de son produit le plus cher (qui sont sur le schéma relationnel **Frais** bien entendu).
On modifie un peu la requête précédente.

La requête SQL est :

```
SELECT Marque.nom, Max(Frais.prix) FROM Frais JOIN Marque
ON id_marque = Marque.id GROUP BY Marque.nom;
```

On va écrire maintenant une requête donnant pour chaque marque le nom de la marque et le prix moyen de leurs produits coûtant strictement plus de 10 euros (qui sont sur le schéma relationnel **Frais** bien entendu).
On modifie encore un peu la requête précédente.

La requête SQL est :

```
SELECT Marque.nom, AVG(Frais.prix) FROM Frais JOIN Marque
ON id_marque = Marque.id WHERE Frais.prix > 10 GROUP BY Marque.nom;
```

On remarque ici que **WHERE** est avant **GROUP BY**

2. La clause HAVING

Le mieux encore c'est de reprendre l'exemple précédent avec pour nouvelle requête d'obtenir le nom des marques et le prix moyen de leurs produits si ce prix moyen est supérieur strictement à 10 euros
On modifie la requête précédente. La clause **WHERE** disparaît et est remplacée par **HAVING** mais qui se place maintenant après **GROUP BY**

La requête SQL est :

```
SELECT Marque.nom, AVG(Frais.prix) FROM Frais JOIN Marque
ON id_marque = Marque.id GROUP BY Marque.nom HAVING AVG(Frais.prix)>10;
```

Pour distinguer **WHERE** et **HAVING**, on dira que la clause **WHERE** filtre avant une agrégation et la clause **HAVING** filtre après les groupements.

On peut utiliser **WHERE** et **HAVING** ensemble.

Par exemple prenons la relation **Notes (Date, Exercice , Points)**

```
SELECT Date, Exercice, AVG(Points) FROM Notes
WHERE Date >= "2022-01-01" GROUP BY Date, Exercice
HAVING AVG(Points)>10 ORDER BY Date;
```

VII. Ordre de la composition d'une requête

Il faut bien distinguer l'ordre d'écriture d'une requête et l'ordre d'exécution qui ne sont pas du tout les mêmes.

1. Ordre d'écriture d'une requête SQL

Les seules clauses obligatoires sont **SELECT** et **FROM**, les autres sont à placer à la demande selon la requête voulue. Dans l'ordre d'apparition sur la scène :

SELECT expressions (composées avec des noms de colonnes, séparées par des virgules)

FROM tables (séparées par le mot **JOIN** ou par des virgules)

ON conditions (si jointure avec **JOIN**)

WHERE conditions

GROUP BY attributs (séparés par des virgules)

HAVING conditions

ORDER BY attributs

LIMIT

OFFSET

2. Ordre d'exécution d'une requête

Cet ordre est important car il permet de batir la requête sur le brouillon (avant de la mettre en ordre d'écriture sur votre copie).

▷ **Étape 1 : choix des tables**

FROM : où trouver les informations, dans quelles tables ;

JOIN : s'il y a plusieurs tables à traiter.

▷ **Étape 2 : choix des lignes**

ON : sélection des lignes à traiter si jointure avec **JOIN** ;

WHERE : sélection des lignes à traiter ;

GROUP BY : pour effectuer des regroupements sur les lignes à traiter ;

HAVING : pour ajouter des conditions sur les groupes obtenus.

▷ **Étape 3 : choix des colonnes**

SELECT : choisir les colonnes.

▷ **Étape 4 : Traitement des colonnes**

MIN, MAX, ..., +, -, ... : traiter les colonnes avec des fonctions ou des opérateurs ;

DISTINCT : supprimer les doublons ;

ORDER BY : ordonne les résultats ;

LIMIT OFFSET : ne conserver que certaines lignes de la réponse.

Conséquence de l'ordre d'execution sur le renommage d'une table ou une colonne

Si l'on renomme une table dans la clause **FROM** qui s'execute en premier, ce renommage est disponible pour toute la suite.

Si l'on renomme une colonne dans l'instruction **SELECT**, il faut faire attention.

On peut écrire sans problème **SELECT Id AS n FROM Classe ;**

Par contre ce renommage n'est pas disponible pour **WHERE** par exemple qui est traitée avant.

Ainsi la requête **SELECT Id AS n FROM Classe WHERE n>5 ;** ne fonctionnera pas car **n** est inconnu au moment de la clause **WHERE** qui intervient après **FROM** mais avant **SELECT** dans l'ordre d'exécution.
La solution pour résoudre ce problème est d'écrire **WHERE Id>5** dans la requête précédente.

Chapitre 5

Lagrange polynomials and applications

■ Objectifs

Les incontournables :

savoir

Cours

■ Polynômes interpolateurs de Lagrange

Problématique : Soient $n+1$ couples $(x_0, y_0), ..., (x_n, y_n)$ dans $\mathbb{R}^2$ tel que pour tout $i \in [\![0, n]\!]$, les réels x_i soient **tous distincts**. Il s'agit de construire un polynôme P de degré au plus n tel que pour tout $i \in [\![0, n]\!]$, on ait l'égalité : $P(x_i) = y_i$.

Définition : On appelle **polynômes interpolateurs de Lagrange** associés à la famille $(x_0, x_1, ..., x_n)$ les $n+1$ polynômes définis pour tout $i \in [\![0, n]\!]$ par :

$$L_i = \prod_{j=0, j\neq i}^{n} \frac{X - x_j}{x_i - x_j}.$$

Vocabulaire : La famille $(L_0, L_1, ..., L_n)$ est appelée **base d'interpolation de Lagrange** associée à $(x_0, x_1, ..., x_n)$.

Exercice 01

Déterminer la base d'interpolation de Lagrange pour la famille $(0, 1, 2)$.

Propriétés des polynômes $L_0, ..., L_n$

- ☞ Pour tout $i \in [\![0, n]\!]$, L_i est de degré n.
- ☞ Pour tout $i \in [\![0, n]\!]$ et pour tout $j \in [\![0, n]\!]$, $L_i(x_j) = \delta_{i,j} = \begin{cases} 1 & \text{si} \quad j = i \\ 0 & \text{si} \quad j \neq i \end{cases}$
- ☞ La famille $(L_0, L_1, ..., L_n)$ est une famille libre de $\mathbb{K}_n[X]$ et en est donc une base.

Exercice 02 : *Faisons les preuves de ces trois propriétés*

Comme $(L_0, ...L_n)$ est une base de $\mathbb{R}_n[X]$, tout polynôme de $\mathbb{R}_n[X]$ s'exprime dans cette base et on a en particulier le résultat suivant :

Théorème : Il existe un seul polynôme de degré au plus n (donc dans $\mathbb{R}_n[X]$) qui vérifie pour tout $i \in [\![0, n]\!]$, $P(x_i) = y_i$. C'est le polynôme $P = \sum_{i=0}^{n} y_i L_i$.

Exercice 03 : *Faisons la preuve de ce théorème*

Exercice 04 : *On suppose que* $L_0, ..., L_n$ *est la base d'interpolation de Lagrange associée à* $(x_0, x_1, ..., x_n)$. *Justifier que* $1 = \sum_{i=0}^{n} L_i$ *et que* $X = \sum_{i=0}^{n} x_i L_i$

Exercice 05 : *Déterminer l'unique polynôme* P *de degré* 2 *tel que* $P(0) = 1$, $P(1) = -3$ *et* $P(2) = 2$ *en utilisant la base d'interpolation trouvée dans l'exercice 01*

■ Mise en place d'un algorithme pour trouver une base d'interpolation de Lagrange

On part de $Xabs = [x_0, x_1, ..., x_n]$. L'idée est de créer une liste $L = [L_0, L_1, ..., L_n]$ constituée des $n+1$ polynômes de la base d'interpolation de Lagrange associés à $Xabs$

On commence par initialiser `L=[]`

On doit créer une boucle de variable `i` avec `i` variant de `0` à `n`

Prenons `i=0` Alors on veut calculer $L_0 = \prod_{j=0, j\neq 0}^{n} \frac{X - x_j}{x_0 - x_j}$.

On initialise `Pol=1` puis on fait une boucle intérieure d'indice `j` variant de `0` à `n` avec `j!=0` et dans cette boucle intérieure, on fait :

$$Pol \leftarrow Pol \times \frac{X - x_j}{x_0 - x_j}$$

Et la valeur finale de Pol est bien L_0 que l'on rentre dans L.

Puis on prend `i=1` Alors on veut calculer $L_1 = \prod_{j=0, j\neq 1}^{n} \frac{X - x_j}{x_1 - x_j}$.

On initialise `Pol=1` puis on fait une boucle intérieure d'indice `j` variant de `0` à `n` avec `j!=1` et dans cette boucle intérieure, on fait :

$$Pol \leftarrow Pol \times \frac{X - x_j}{x_1 - x_j}$$

Et la valeur finale de Pol est bien L_1 que l'on rentre dans L.

Ainsi de suite... À la fin, on retourne `L`

L'algorithme est donc :

```
Donnees : la liste Xabs
n <- longueur(Xabs) - 1  et L  <- []
pour i variant de 0 -> n faire :
     Pol <- 1
     pour j variant de 0 -> n faire :
          si j different de i faire :
               Pol  <- Pol . (X-Xabs[j])/(Xabs[i]-Xabs[j])
     L <- Pol
On renvoie L
```

À partir de $L = [L[0], ..., L[n]]$, on peut donc calculer le polynôme d'interpolation `P` à partir des valeurs $y_0, ..., y_n$. Si l'on pose $Xord = [y_0, y_1, ..., y_n]$ on peut faire la boucle :

```
Donnees : la liste Xord et la liste L
n <- longueur(Xord) - 1 ; P <- 0
pour k variant de 0 -> n faire :
     P  <- P + Xord[k]*L[k]
On renvoie P
```

On peut aussi utiliser la fonction `sum` en Python, c'est plus rapide.

Chapitre 6

Machine learning

■ Objectifs

Les incontournables :

savoir

Résumé de cours

■ Apprentissage automatique

Notion d'apprentissage

L'apprentissage automatique est un domaine de l'intelligence artificielle. Il s'agit de permettre à une machine de progresser, d'apprendre par elle-même à améliorer son fonctionnement dans un cadre de résolution d'un problème, mais sans jamais la programmer explicitement pour résoudre ce problème. Des outils mathématiques principalement des outils statistiques, sont appliqués sur des données et les résultats obtenus permettent à la machine d'améliorer peu à peu son efficacité dans la résolution du problème.

À partir de données, un modèle est construit. Par exemple, pour apprendre à reconnaître un élément dans une image, on fournit à la machine des images avec la présence de l'élément ou pas et un modèle est élaboré. Ce modèle permet à la machine, lorsqu'on lui fournit une image nouvelle, de dire si l'élément est présent ou pas. Ceci constitue la phase d'apprentissage. Ensuite, lorsque le modèle est suffisamment correct, on peut l'utiliser en poursuivant ou pas l'apprentissage. Pour pousuivre un apprentissage, il est nécessaire d'avoir un moyen de vérifier une réponse et de la certifier ou pas, afin que la machine puisse prendre en compte dans son modèle les nouvelles informations.

On distingue plusieurs formes d'apprentissage, parmi lesquelles ***l'apprentissage supervisé*** et ***l'apprentissage non supervisé***. Pour la première forme, des données sont fournies à la machine avec un classement. On dit que les données sont étquetées ou que les données sont classées. Pour la seconde forme, les données ne sont pas étiquetées et c'est la structure de ces données que la machine doit essayer de déterminer.

D'autres formes d'apprentissage existent, par exemple ***l'apprentissage par renforcement*** ou ***l'apprentissage en profondeur***.

Dans un apprentissage supervisé, à partir des caractéristiques fournies, deux dans le cas le plus simple, soit 0 ou 1 par exemple, on procède à un classement en décidant des caractéristiques d'un nouvel élément après l'examen des caractéristiques de ces k plus proches voisins. On utilise dans ce cadre ***l'algorithme des k plus proches voisins***, en anglais ***K Nearest Neighbours***.

Noter qu'une méthode de classement est différente d'une méthode de régression. Dans une méthode de regression, des données sont fournies et utilisées dans un calcul qui permet de construire une relation, ou une équation, entre les différentes caractéristiques. On peut alors par exemple estimer pour un nouvel élément la valeur cherchée.
Pour une méthode de classement, on essaie de déterminer l'appartenance à une classe en observant les classes des éléments proches.

Dans un apprentissage non supervisé, on dispose de données non classées et on essaye d'établir une classification, une structure de ces données. Les données sont séparées en plusieurs groupes avec dans chaque groupe des données qui présentent un certain degré de similarité. Ceci revient à créer une partition de l'ensemble des données. On dispose pour cela de ***l'algorithme des k-moyennes***, en anglais ***k-means***.

Notion de partition

Pour classer une donnée parmi des données appartenant à un ensemble E, il est nécessaire de disposer d'une classification, c'est-à-dire de sous-ensembles de E qui constituent une partition de cet ensemble. Chaque élément de E appartient à une et une seule classe. La question est de déterminer à quelle classe appartient la nouvelle donnée.
Établir une classification consiste à écrire une partition de E.

Définition

une partition d'un ensemble E est une famille $(E_i)_i$ de sous-ensembles non vides de E qui vérifient les deux conditions :

- ☞ $\bigcup_i E_i = E$.
- ☞ $E_i \cap E_j = \emptyset$ pour tout couple (i, j) avec $i \neq j$.

Un sous-ensemble appartenant à cette famille est une partie de E.
Le nombre de parties d'un ensemble E à n éléments est 2^n.

Écrivons un programme récursif qui génère toutes les parties d'un ensemble E.

```
In [1]: def parties(liste, resultats):
            if len(liste) == 0 :
                resultats.append([])
            else :
                liste2 = liste[1:len(liste)]
                parties(liste2,resultats)
                for k in range(len(resultats)):
                    elt = resultats[k] + [liste
                        [0]]
                    resultats.append(elt)
```

```
In [2]: def ens_parties(ensemble):
            rep = []
            parties(ensemble,rep)
            return rep
```

```
In [3]:  E = [1,2,3,4]
In [4]:  print(ens_parties(E))
Out[4]:
[[], [4], [3], [4, 3], [2], [4, 2], [3, 2], [4,
    3, 2], [1], [4, 1], [3, 1], [4, 3, 1], [2,
    1], [4, 2, 1], [3, 2, 1], [4, 3, 2, 1]]
```

Explication de cette procédure.
En effet transformons `parties` en terminant par `return resultats`

```
In [5]: def parties(liste, resultats):
            if len(liste) == 0 :
                resultats.append([])
            else :
                liste2 = liste[1:len(liste)]
                parties(liste2,resultats)
                for k in range(len(resultats)):
                    elt = resultats[k] + [liste[0]]
                    resultats.append(elt)
            return resultats
In [6]: liste = [1,2,3,4]
In [7]: liste2 = liste[1:len(liste)]
In [8]: liste2
Out[8]: [2,3,4]
In [9]: parties(liste2,[])
Out[9]: [[], [4], [3], [4, 3], [2], [4, 2], [3, 2], [4, 3, 2]]
```

On a `liste2` est `liste` privée de `1`.
Puis comme c'est le premier `return` qui s'affiche, on affiche `[1:len(liste)]` qui est l'ensemble des parties de `liste2`.
Puis la boucle `for k in range(len(resultats)):` n'est autre ici que `for k in range(0):` et dans `elt`, on ajoute `resultats[k]` qui est le vide et `[liste[0]` qui est `1`. Ainsi dans `resultats`, on reprend `parties(liste2,[])` où on ajoute `1`.

■ Apprentissage supervisé

Algorithmes des k plus proches voisins

Un ensemble de données est connu et chacun des données appartient à une classe ou une partie bien déterminée. Cette classe est une des caractéristiques de chaque donnée liée aux autres caractéristiques.
Dans un apprentissage supervisé, un algorithme doit permettre d'émettre une prévision sur cette caractéristique à propos d'une donnée dont on ne connaît que les autres caractéristiques, donc de prédire la partie dans laquelle la donnée peut être classée.
Une méthode consiste à baser cette prédiction sur la détermination des classes de ses k plus proches voisins, où k est à préciser, en retenant la classe majoritaire.
On dispose donc d'un ensemble E de n points représentant des données classées ou étiquetées. L'ensemble E est l'ensemble d'apprentissage. On choisit un entier k, plus petit que n, et on dispose d'un point x qui n'est pas dans E. Il s'agit de trouver parmi les points de E les k plus proches de x pour classer x ou l'étiqueter. Le mot « proche » sous-entend une notion de distance. Ce peut être une distance sur les couleurs par exemple sur la quantité de rouge ou sur le niveau de gris. Dans la reconnaissance des caractères, ce peut être une distance sur les formes (tailles, boucles). Ainsi des caractères d'imprimerie comme le b ou le h peuvent être considérés comme proches.

Dans tous les cas, les caractéristiques sont exprimées par des valeurs numériques et nous pouvons utiliser une distance dans un espace de dimension d quelconque (les d coordonnées correspondent à d caractéristiques).

Le cas le plus simple est la recherche du plus proche voisin, c'est-à-dire pour $k = 1$. On l'utilise par exemple pour trouver un chemin reliant des points dans un espace. À partir d'un point origine, on cherche le point le plus proche, et ainsi de suite jusqu'à atteindre un point extrémité. Il s'agit donc d'un algorithme qui permet dans certains cas de produire le chemin le plus court entre les deux extrémités.

En pratique, nous utilisons la distance euclidienne ou plutôt le carré de la distance euclidienne. Donnons cette fonction en dimension n.

```
In [1]: def d(x,y):
            n=len(x)
            s=0
            for i in range(n):
               s = s + (x[i] - y[i])**2
            return s
```

Exemple

Nous considérons un ensemble E dont un élément est représenté par une liste de deux flottants. Un élément est donc considéré comme un point dans un espace de dimension 2.
On construit une liste appelée `voisins` qui contient les k plus proches voisins d'un point x quelconque n'appartenant pas à E, les voisins étant des éléments de E.
On commence par écrire une fonction `d` qui calcule le carré de la distance euclidienne entre deux points du plan.

```
In [2]: def d(x,y):
            return (x[0] - y[0])**2 + (x[1] - y[1])**2
```

La liste des points de E est triée selon l'ordre croissant des distances à un point P. On écrit donc une fonction nommée `tri` en utilisant la fonction `sorted` de Python. Commençons par apprivoiser cette fonction `sorted`.
Faisons un cas simple pour commencer.

```
In [1]: sorted([5,2,3,1,4])
Out[1]: [1, 2, 3, 4, 5]
```

`sorted` trie dans une liste. Le paramètre `key` que l'on peut ajouter comme attribut dans `sorted` permet de spécifier une fonction qui peut être appelée sur chaque élément de la liste avant d'effectuer des comparaisons. Donnons des exemples.

```
In [1]: sorted("This is a test string from Walter".split(),key=str.
   lower)
Out[1]: ['a','from','is','string','test','This','Walter']
In [2]: studenten = [('john','A',15),('jane','B',12),('dave','B',10)]
In [3]: sorted(studenten,key = lambda studenten : studenten[2])
Out[3]: [('dave', 'B', 10), ('jane', 'B', 12), ('john', 'A', 15)]
In [4]: sorted(studenten,key = lambda studenten : studenten[1])
Out[4]: [('john', 'A', 15), ('jane', 'B', 12), ('dave', 'B', 10)]
In [5]: sorted(studenten,key = lambda studenten : studenten[0])
Out[5]: [('dave', 'B', 10), ('jane', 'B', 12), ('john', 'A', 15)]
```

Passons à une fonction `tri` d'arguments la liste `E`, le point `P` et la fonction distance `d`.
La fonction `choix(elt)` permet de trier suivant les valeurs d'indices 1. On peut créer E comme une liste de points `(a,b)` (à la place de `[a,b]` qui marche aussi).

```
In [1]: def tri(E,P,d):
            def choix(elt):
                return elt[1]
            distances = [ (p,d(p,P)) for p in E]
            return sorted(distances, key=choix)

In [2]: E=[(1,2),(-1,0),(1,1),(1,-1)]

In [3]: P=(0,0)

In [4]: tri(E,P,d)
    [((-1, 0), 1), ((1, 1), 2), ((1, -1), 2), ((1, 2), 5)]
```

Donnons la fonction `knn` qui renvoie les k premiers points de la liste triée.

```
In [1]: def knn(E,P,d,k):
            pts = tri(E,P,d)
            return [elt[0] for elt in pts[0:k]]

In [2]: knn(E,P,d,3)
Out[2]: [(-1, 0), (1, 1), (1, -1)]
```

Créons une liste E un peu plus élaborée que la liste E précédente. On l'appellera `points`.
Dans la syntaxe, `random()` fournit un flottant aléatoire entre 0 et 1 et donc `round(100*random(),2)` fournit un flottant avec deux chiffres après la virgule entre 0 et 100.

```
In [1]: from random import random
In [2]: points = []
In [3]: for i in range(1000):
            x = round(100 * random(), 2)
            y = round(100 * random(), 2)
            points.append((x,y))
```

Si l'on tape `points` alors 1000 points apparaissent.

```
In [1]: P=(0,0)
In [2]: voisins = knn(points,P,d,4)
In [3]: voisins
Out[3]: [(0.96, 2.5), (4.87, 3.16), (3.86, 4.85), (5.32, 4.06)]
```

Supposons maintenant que l'on dispose d'une classification des données de E. Afin de classer la donnée P, l'algorithme lui assigne l'étiquette de la classe majoritaire parmi les voisins qui ont été déterminés.
Plus précisément, si par exemple une classe est représentée par un niveau de gris (gris clair, gris moyen, gris amiral), on prédit le niveau de gris du point en choisissant le niveau de gris qui prédomine parmi les voisins du point.
La manière de déterminer la classe prédominante parmi les voisins repose sur un comptage et une recherche de maximum.
Par exemple, créons une fonction `classe_maj` de premier argument `pts` qui sont les voisins de `P` choisis, de second argument `nb_classes` qui est en fait `len(classes)` si l'on a définit la liste `classes` et de troisième argument `partition` qui est un dictionnaire dont les clès sont les points de E sous forme de couples et les valeurs sont les numéros de classe.
Dans la procédure, on crée `cpts` qui fait le comptage, la variable `maxi` permet de rechercher le maximum, `randrange(4)` par exemple renvoie aléatoirement un entier pris parmi $0, 1, 2$ et 3.

```
In [1]: def classe_maj(pts,nb_classes,partition):
            cpts = [0] * nb_classes
            for v in pts :
                classe = partition[v]
                cpts[classe] += 1
            maxi = 0
            for n in cpts:
                if n > maxi :
                    maxi = n
            choix = [i for i in range(nb_classes) if cpts[i] == maxi]
            estime = choix[randrange(len(choix))]
            return estime
```

Puis on crée des classes. Ici il y aura trois classes : $0, 1$ et 2. De plus `randrange(0,len(classes)` donne aléatoirement un entier compris entre 0 et `len(classes)` c'est-à-dire 2. La procédure va créer `partition` sous forme d'un dictionnaire.

```
In [1]: classes = [0,1,2]
In [2]: from random import randrange
In [3]: partition = {point : randrange(0,len(classes)) for point in
   points}
```

Si l'on demande de retourner `partition` (On n'a affiché que les premiers et derniers par mesure d'économie de place) :

```
In [1]: partition
Out[1]: {(29.69, 20.92): 1, (53.65, 41.9): 2, (63.62, 80.06): 2,
... (82.74, 11.6): 1, (5.5, 44.53): 0, (28.68, 51.55): 0}
```

On exécute alors la fonction `classe_maj` avec la liste des 4 plus proches voisins de P.

```
In [1]: voisins = knn(points, P,d,4)

In [2]: voisins
Out[2]: [(0.15, 1.26), (1.46, 1.13), (3.14, 2.27), (3.32, 3.39)]

In [3]: classe_maj(voisins,3,partition)
Out[3]:  0
```

Remarque : La notion de plus proches voisins est présente dans de nombreux domaines. Simplement dans l'espace par exemple, elle est utile pour gérer des risques de collisions ou des interactions à distance entre des objets.
Si des objets peuvent être identifiés par des caractéristiques mesurables, il est possible de définir une distance entre ces objets. On peut alors trouver les plus proches voisins d'un nouvel objet dans des questions de reconnaissance ou d'identification.
Des applications pour smartphones permettent de reconnaître une musique, une chanson, son interprète, simplement avec quelques secondes d'écoute. Cet extrait n'est pas comparé dans sa globalité avec les extraits figurant dans une énorme base de données (des millions de titres), afin d'y trouver des « voisins » et de choisir le plus proche. Ce serait beaucoup trop lent. La réussite est basée sur la définition d'un petit nombre de points caractéristiques, des marqueurs, qui permettent de définir de manière unique chaque titre. C'est à partir des marqueurs de l'extrait que la recherche est effectuée dans la base de données.

L'algorithme des k plus proches voisins permet d'émettre des prédictions. Il reste à mesurer la qualité de ces prédictions.

Matrice de confusion

Exemple 1

Considérons un test censé prédire si une personne est porteuse ou non d'une maladie. Ce test fait la bonne prédiction dans 90% des cas. Cela signifie que dans 90% des cas, si une personne est porteuse de la maladie, le test est positif et sinon le test est négatif. On parle de « vrais positifs » et de « vrais négatifs ». Dans 10% des cas, le test ne fait pas la bonne prédiction. Le résultat est donc soit positif alors que la personne n'est pas porteuse de la maladie, c'est un « faux positif » et soit négatif alors que la personne est porteuse de la maladie, on parle de « faux négatif ».
Il est intéressant de connaître la répartition de ces 4 types de résultats.

On peut schématiser avec un nuage de points. Les points gris foncés représentent la classe + et les points gris clairs représentent la classe −.
Nous allons visualiser avec le programme suivant à taper :

```
In [1]: from random import randint
In [2]: def creer_points(nb,dim,coul):
            pts=[]
            while len(pts) <nb:
                 x = randint(0,dim)
                 y = randint(0,dim)
                 c = coul[randint(0,1)]
                 if (x,y,c) not in pts :
                     pts.append((x,y,c))
            return pts
```

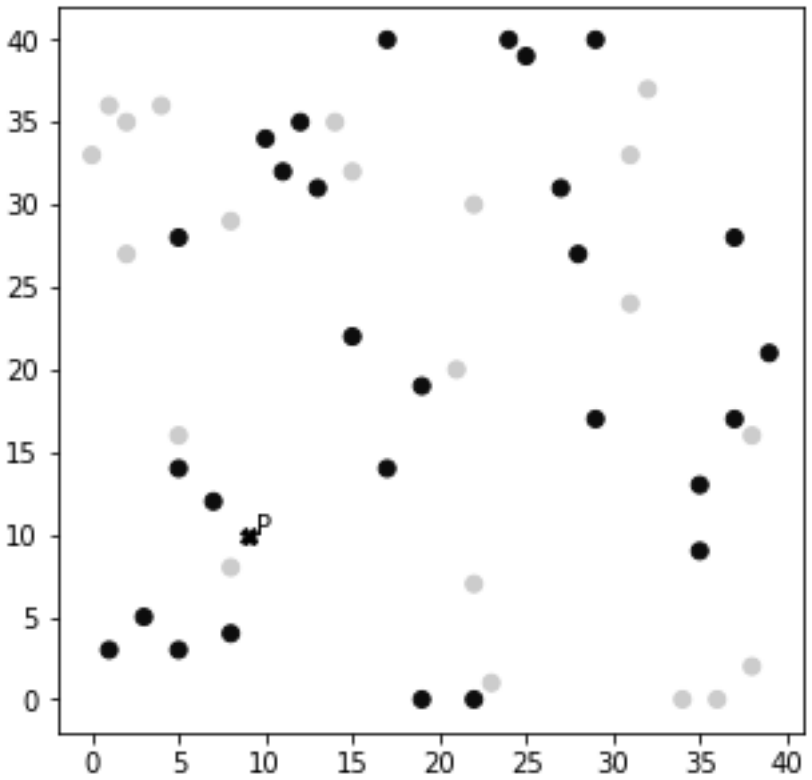
On a créé une liste de points dans le plan avec $0 \leqslant x \leqslant$ `dim` et $0 \leqslant y \leqslant$ `dim` et `c` désigne la couleur de (x, y) qui sera un niveau de gris (ici clair ou foncé). Tapez alors :

```
In [1]: couleurs=["0.05", "0.8"]
In [2]: points = creer_points(50,40,couleurs)
In [3]: points
Out[3]: [(10, 34, '0.05'), (24, 40, '0.05'), (34, 0, '0.8'),
  ...
  (39, 21, '0.05'), (5, 16, '0.8'), (2, 35, '0.8')]
```

On va créer un point P aléatoire et tracer le nuage `points` en visualisant P. On tape :

```
In [1]: import matplotlib.pyplot as plt
In [2]: P = (randint(0,40), randint(0,40),"k")
In [3]: x = [p[0] for p in points]
In [4]: y = [p[1] for p in points]
In [5]: c = [p[2] for p in points]
In [6]: plt.scatter(x,y,linestyle='None',color=c, marker="o");
In [7]: plt.plot(P[0],P[1],P[2]+"X"); plt.text(P[0]+0.4,P[1],"P");plt.
   show()
```

Le point P n'appartient pas au nuage de points mais on peut l'y associer en affectant par exemple à P la couleur dominante (gris clair ou gris foncé) parmi ses k voisins les plus proches par l'algorithme des k plus proches voisins. C'est ce que l'on fera en TD. Ainsi si la couleur dominante est le gris foncé, on attribuera le gris foncé à P par prédiction. Mais P est peut-être en réalité gris clair.



On ajoute ainsi des points dont on connaît la couleur, foncée ou claire. On vérifie alors pour chaque point ajouté si sa couleur est bien celle qui est prévue ou pas par le modèle de prédiction.
On obtient quatre sortes de points : des vrais foncés et des vrais clairs (la prédiction est alors exacte) et des faux foncés et des faux clairs (la prédiction est inexacte).

		Prédiction de foncés	**Prédiction de clairs**
		Foncé	Clair
Foncés réels	Foncé	*vrais foncés*	*faux clairs*
Clairs réels	Clair	*faux foncés*	*vrais clairs*

Avec 200 points ajoutés, par exemple 150 points foncés et 50 points clairs, on peut obtenir une matrice comme : $\begin{pmatrix} 137 & 13 \\ 4 & 46 \end{pmatrix}$, appelée *matrice de confusion*.

Dans un problème de prédiction avec plus de deux classes, le principe est le même.
De manière générale, une matrice de confusion est construite ainsi :

- ☞ chaque ligne correspond à une classe réelle ;
- ☞ chaque colonne correspond à une classe estimée.

Donc à l'intersection d'une ligne i et d'une colonne j, on trouve le nombre d'éléments de la classe réelle i qui ont été estimés comme appartenant à la classe j.

Exemple 2

Donnons un exemple de matrice de confusion à trois classes : $\begin{pmatrix} 178 & 13 & 9 \\ 7 & 137 & 6 \\ 3 & 5 & 92 \end{pmatrix}$.
Il y a $178 + 13 + 9 = 200$ de la classe `c = 0` réellement, $7 + 137 + 6 = 150$ de la classe `c = 1` réellement, $3 + 5 + 92 = 100$ de la classe `c = 2` réellement.
Puis, il y a 178 points réellement de la classe `c = 0` à qui on prédit qu'ils sont de la classe `c = 0`.
Le taux de prédictions correctes de la classe `c = 0` est donc $178/200 = 0.89$.

Le programme suivant `tpc` permet de renvoyer ce taux de prédictions correctes de la classe `c` avec `c` entier entre 0 et 2. Si `mc` est la matrice de confusion alors `mc[c][c]` donne le nombre de prédictions correctes de la classe `c`. On tape :

```
In [1]: def tpc(mc,c):
            n = len(mc)
            nbc = 0
            nbpc = mc[c][c]
            for j in range(n):
                nbc = nbc + mc[c][j]
            return nbpc/nbc
```

On l'applique à notre matrice de confusion et aux différentes classes.

```
In [2]: mc = [[178,13,9],[7,137,6],[3,5,92]]

In [3]: tpc(mc,2), tpc(mc,1), tpc(mc,0)
Out[3]: 0.92   0.9133333333333333  0.89

In [4]: tpc(mc,3)
Out[4]:
Traceback (most recent call last):
    File "<ipython-input-53-bbaca20b13f2>", line 1, in <module>
    tpc(mc,3)
    File "<ipython-input-48-fdb413bcaa44>", line 4, in tpc
    nbpc = mc[c][c]
IndexError: list index out of range
```

Le `tpc(mc,3)` c'est juste pour voir.
Enfin, $3+5+92 = 100$ et $92/100 = 0.92$ puis $7+137+6 = 150$ et $137/150 = 0.9133333333333333$.
Donc c'est OK.

Intéressons nous now plutôt aux taux de prédictions incorrectes. Intéressons nous pour fixer les idées à la classe `c = 0` par exemple. Il y a 7 points pour lesquels on a prédit `c = 0` alors que c'est réelement `c = 1` et 3 points où on a prédit `c = 0` alors que c'est réelement `c = 2`. Donc cela fait 10 points où on prédit `c = 0` de façon éronée. Le nombre de points qui sont réellement de classe `c = 1` est $7 + 137 + 6 = 150$ et le nombre de points qui sont réellement de classe `c = 2` est $3 + 5 + 92 = 100$. Le nombre de points qui ne sont pas réellement de la classe `c = 0` est donc 250. Le rapport $10/250 = 0.04$ donne un taux de prédiction incorrecte de la classe `c = 0`.
Le programme suivant permet de faire ce rapport pour toutes les classes.

```
In [1]: def tpi(mc,c):
            n = len(mc)
            nbi = 0 # total classes i differentes de c
            nbpi = 0 # total predictions incorrectes de c
            for i in range(n):
                if i != c :
                    for j in range(n):
                        nbi = nbi + mc[i][j]
                    nbpi = nbpi + mc[i][c]
            return nbpi/nbi

In [2]: tpi(mc,0), tpi(mc,1), tpi(mc,2)
Out[2]:   (0.04, 0.06, 0.04285714285714286)

In [3]: 18/300
Out[3]: 0.06
In [4]: 15/350
Out[4]: 0.04285714285714286
```

Vérifier à la main dans le cas `c = 0` ce programme.

On a d'autres instruments de mesure dans une matrice de confusion. Par exemple :

- ☞ **Le taux d'erreurs**. C'est le nombre de prédictions incorrectes sur le nombre total de prédictions. On vise une valeur la plus proche de 0. Dans le cas de l'exemple 2, ce taux vaut $(7 + 3 + 13 + 5 + 9 + 6)/450 = 43/450 = 0.095555$.
- ☞ **La précision**. C'est le nombre de prédictions correctes sur le nombre total de prédictions. On vise une valeur la plus proche de 1. Si TE est le taux d'erreurs et P la précision, alors $P = 1 - TE$. Dans le cas de l'exemple 2, P vaut $1 - 43/450 = 0.9044444$.

Revenons au cas général.
Pour calculer la matrice de confusion, il faut disposer d'un ensemble de données à tester et de l'ensemble des résultats attendus. On compare les résultats attendus avec les résultats prédits sur les données à tester. Pour cela, on utilise un ensemble de données classées. Cet ensemble est partagé en deux-sous ensembles, l'un constituant **l'ensemble d'apprentissage** et l'autre **l'ensemble de test**. On peut choisir comme dans l'exemple 3 qui suit, les proportions 3/4 et 1/4. L'ensemble d'apprentissage est utilisé pour prédire la classe de chaque élément de l'ensemble de test avec l'algorithme des k plus proches voisins.

Exemple 3

Étape 01

On crée un ensemble E de points du plan appartenant à deux classes 0 et 1 visualisées par des disques noirs ou des carrés noirs en utilisant le code suivant. On rappelle que `random()` renvoie un flottant aléatoire compris entre 0 et 1. Et `random() < val` crée une fonction booléene.

```
In [5]: from random import random
In [6]: random()
Out[6]: 0.0976439354494626
In [7]: random() <0.9
Out[7]: True
In [8]: random()> 0.9
Out[8]: False
```

Puis, on crée un ensemble de points sous forme d'une liste de triplets (x, y, c), où x et y sont des réels pris aléatoirement entre `-dim` et `dim` puis c est la classe 0 ou 1 choisi alátoirement avec `random()`.

```
In [9]: from random import randint

In[10]: nbpoints, dim = 800, 50
In[11]: classes = [0,1]

In[12]: def points(n,dim) :
            ens = []
            while len(ens) < n:
                x = randint(-dim,dim)
                y = randint(-dim, dim)
                if x <= 0:
                    cl = classes[random() < 0.9]
                else :
                    cl = classes[random() >= 0.9]
                ens.append((x,y,cl))
            return ens
```

Faisons tourner.

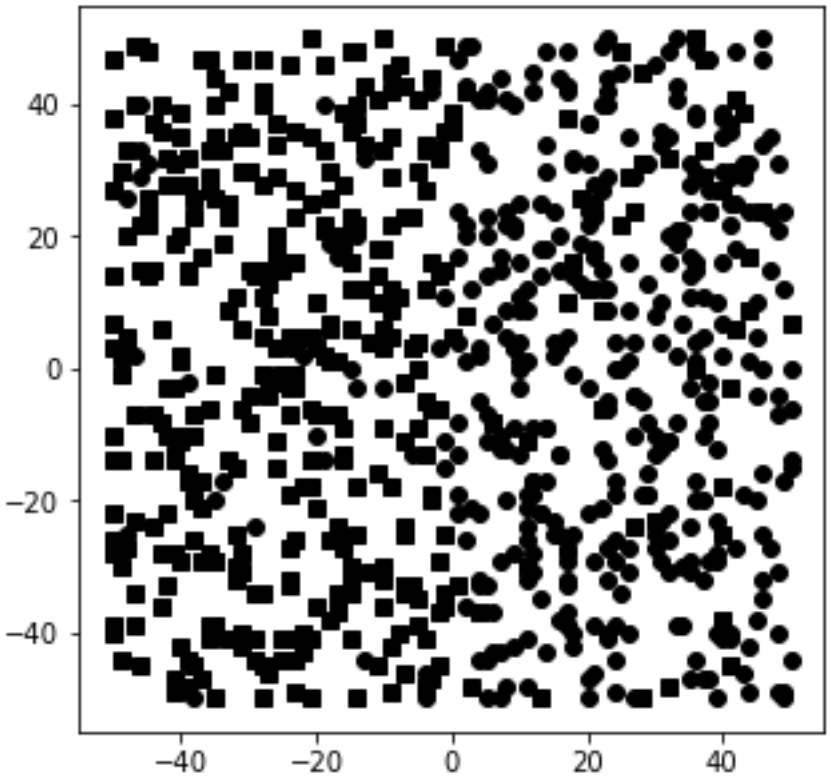
```
In[13]: E = points(nbpoints,dim)
In[14]: E
Out[14]:  [(-35, -27, 1), (-32, 26, 1), ...
 (39, 42, 0), (28, 30, 0), (-13, -11, 1), (-33, 44, 1)]
```

Visualisons tout ça. Le code `plt.recParams` qui est un objet de type dictionnaire en passant, permet d'ajuster des paramètres. Ici, on définit la taille des points que l'on va dessiner. Puis dans `plt.plot`, l'attribut `color = "k"` fait dessiner en noir les points, puis `marker = "o"` fournit des disques et `marker = "s"` fournit des carrés. Donc ici les points de classe 0 seront des disques noirs et les points de classe 1 seront des carrés noirs.

```
In[15]: import matplotlib.pyplot as plt

In[16]: plt.rcParams["figure.figsize"] = (5,5)

In[17]: for p in E:
            if p[2] == 0 :
                plt.plot(p[0],p[1],color="k", marker= "o")
            else :
                plt.plot(p[0],p[1],color="k", marker= "s")
```



Étape 2.
On va scinder E en deux ensembles, l'ensemble de points d'apprentissage `ens_appr` et un ensemble de test `ens_test`. On va respecter la proportion 75% et 25%. L'idée est de reprendre `points` et de le retoucher. On l'appellera `new_points` pour commencer.

On scinde `ens` en deux ensembles `ens_appr` et `ens_test` dans le corps de la procédure.
Puis on remplace `len(ens)` par `len(ens_appr) + len(ens_test)`.
Puis dans la boucle `while`, après la création de `cl`, on tape `choix = randint(0,3)` et donc dans `choix`, on met aléatoirement $0, 1, 2$ ou 3. Le code `choix == 0` donne donc le quart des choix. Et on remplit `ens_test` avec ce choix et sinon on remplit `ens_appr`.
À la fin, on renvoie `ens_appr` et `ens_test`, donc deux listes distinctes.

```
In[18]: def new_points(n,dim):
            ens_appr = []
            ens_test = []
            while len(ens_appr) + len(ens_test) < n:
                x = randint(-dim, dim)
                y = randint(-dim,dim)
                if x <= 0 :
                    cl = classes[random() < 0.9]
                else :
                    cl = classes[random() >= 0.9]
                choix = randint(0,3)
                if choix == 0 and len(ens_test) < n/4:
                    ens_test.append((x,y,cl))
                elif choix > 0 and len(ens_appr) < 3*n/4:
                    ens_appr.append((x,y,cl))
            return ens_appr, ens_test
```

Faisons un exemple avec $n = 10$ pour bien voir.

```
In[19]:new_points(10,50)

Out[19]:
   ([(-29, -18, 1), (-50, -26, 1), (-17, -23, 0), (-15, -2, 1),
   (14, 16, 1), (13, -32, 0), (-37, 10, 1), (-4, -22, 0)],
   [(11, 4, 0), (-11, 4, 1)])
```

On remarque que `ens_appr` a 8 termes et `ens_test` a 2 termes. Ce n'est pas tout à fait $3/4$ et $1/4$ mais il ne faut pas oublier que `randint` c'est aléatoire et donc plus n est grand, plus le $3/4$, $1/4$ est à peu près respecté.

Étape 3.
Création de la matrice de confusion.
On utilise les fonctions `d`, `tri` et `knn` inspirées de ce que l'on a plus haut pour l'algorithme des k plus proches voisins. On les rappelle si vous devez les retaper.

```
In [1]: def d(p1,p2):
            (x1,y1,c1) = p1
            (x2,y2,c2) = p2
            return (x1-x2) ** 2 + (y1-y2) ** 2
In [2]: def tri(E,P,d):
             def choix(elt):
                 return elt[1]
             distances = [ (p,d(p,P)) for p in E]
             return sorted(distances, key = choix)
In [3]:  def knn(E,p,d,k):
             pts = tri(E,p,d)
             return [elt[0] for elt in pts[0:k]]
```

Puis on tape la procédure `matrice(A,T,k,d)`, où `A` est l'ensemble d'apprentissage et `T` l'ensemble de tests. On utilisera `A,T = new_points(nbpoints, dim)` pour récupérer `A` et `T`.

```
In [4]: def matrice(A,T,d,k):
             conf = [[0,0],[0,0]]
             for p in T :
                 vs = knn(A,p,d,k)
                 cpt = [0,0]
                 for v in vs :
                     if v[2] == 0:
                         cpt[0] += 1
                     else :
                         cpt[1] += 1
                 if cpt[0] > cpt[1] :
                     indice = 0
                 elif cpt[0] < cpt[1] :
                     indice = 1
                 else :
                     indice = randint(0,1)
                 estime = classes[indice]
                 if estime == p[2] :
                     if estime == 0 : # VRAI POSITIF
                          conf[0][0] += 1
                     else : # VRAI NEGATIF
                          conf[1][1] += 1
                 else :
                     if estime == 0 : # FAUX POSITIF
                          conf[1][0] += 1
                     else : # FAUX NEGATIF
                          conf[0][1] += 1
             return conf
```

Comprenons cette procédure :
1) au départ, on prend pour matrice de confusion appelée `conf` la matrice nulle d'ordre 2.
2) Puis pour chaque point `p` fixé de `T`, l'ensemble des tests, on tape `vs = knn(A,p,k,d)` qui donne les k voisins les plus proches de $p \in T$ pris parmi les éléments de `A`, l'ensemble d'apprentissage.
On obtient donc une liste de k triplets du type (x, y, c) avec $c = 0$ ou $c = 1$.
3) On compte les $c = 0$ et les $c = 1$ de cette liste `vs` en rentrant dans la liste `cpt` le nombre de $c = 0$ et le nombre de $c = 1$.
4) Puis on récupère dans `indice` l'indice de la couleur la plus représentée.
Dans le cas où `cpt[0] == cpt[1]`, on s'en remet au dieu random avec `indice = randint(0,1)`
5) On met dans `estime` la valeur `classes[indice]` qui est donc la couleur la plus représentée parmi nos k voisins de `p`.
La matrice `conf` est de la forme $\begin{pmatrix} \text{Vrai} + & \text{Faux} - \\ \text{Faux} + & \text{Vrai} - \end{pmatrix}$ (voir l'exemple 1).
On suppose que la classe `c=0` est celle des positifs et la classe `c=1` est celle des négatifs.
Si `estime == p[2]` et `estime == 0`, cela signifie que l'on a un vrai positif. En effet, `p[2]` fournit la vraie valeur qui est la classe 0. Et `estime` donne aussi 0.
Si par contre, `estime == p[2]` et `estime == 1`, cela signifie que l'on a un vrai négatif. En effet, la vraie classe est 1 (donc un négatif) et `estime` donne aussi 1.
Si par contre `estime not== p[2]` et `estime == 0`, cela signifie que l'on a un faux positif. En effet, `p[2]` est alors 1, et on a un négatif et pourtant `estime` dit qu'il est positif.
Si `estime not== p[2]` et `estime == 1`, cela signifie que l'on a un faux négatif.
En effet, `p[2]` est alors 0, on a affaire à un positif et pourtant le classe parmi les négatifs.

```
In [5]: A,T = new_points(nbpoints, dim)

In [6]: mat_conf = matrice(A,T,d,3)

In [7]: mat_conf
Out[7]:
   [[85, 12], [11, 92]]
```

On commence par remarquer que $85 + 12 + 11 + 92 = 200$. C'est le quart de `nbpoints=800` donc c'est bien la proportion voulue car tous les points comptés dans cette matrice sont uniquement les points de `ens_test`.
Tapons pour finir la procédure `taux_erreur`.

```
In [8]: def taux_erreur(m):
            erreurs = m[0][1] + m[1][0]
            predictions = erreurs + m[0][0] + m[1][1]
           return erreurs/predictions

In [9]: m = [[85,12],[11,92]]

In[10]: taux_erreur(m)
Out[10]:    0.115
```

■ Apprentissage non supervisé

Algorithme des k-moyennes : de quoi s'agit-t-il ?

On dit aussi **clustering algorithm**. On regroupe des données possédant des éléments de similarité dans des groupes. L'ensemble des données étudiées doit pouvoir être encodé dans un ensemble de vecteurs ou une matrice.
À partir d'un jeu de données, on définit un nombre k de groupes ou classes. L'algorithme permet d'analyser le jeu de données et d'affecter chaque donnée à une classe, des données similaires ou proches appartenant alors à une même classe. Dans un apprentissage supervisé, on essaie de trouver une corrélation entre les valeurs des données d'un ensemble et une valeur à prédire. Ici, on essaie de trouver des ressemblances entre les données pour constituer une partition.
Pour déterminer la proximité entre deux données, on utilise une mesure de similarité ou une distance. Nous utiliserons la distance euclidienne $d(X,Y) = \sqrt{\left(\sum_{i=1}^{n} (x_i - y_i)^2\right)}$.

L'objectif est de partitionner l'ensemble des données en groupes (ou clusters) distincts, chaque groupe étant représenté par une donnée centrale dont les autres données du groupe sont proches. Ces centres peuvent être choisis au hasard parmi les données à l'initialisation. Chaque itération consiste à affiner le partitionnement. Pour cela, le centre de chacun des k groupes est ré-évalué par le calcul du barycentre de l'ensemble des données du groupe. C'est de là que vient le nom d'algorithme des k moyennes puisqu'à chaque itération k moyennes sont calculées. Les données sont alors regroupées en utilisant les nouveaux centres.

Warning : les centres calculés à chaque itération ne sont plus en général des éléments de l'ensemble des données.

Pour résumer :

- ☞ Un ensemble de données est fourni sous forme de matrice, le nombre de groupes k est choisi.
- ☞ Initialisation de manière aléatoire des centres de chaque groupe.
- ☞ Répéter les deux étapes :
  regrouper chaque donnée dans la partie définie par le centre le plus proche ;
  remplacer chaque centre par le barycentre des données de son groupe.

Soit le nombre d'itérations est fixé à l'avance, soit les centres ne sont plus modifiés lors d'une itération et l'algorithme a alors trouvé une partition stable de l'ensemble des données : on dit que l'algorithme a convergé.

Le choix du nombre de groupes k est important. Il est nécessaire de tester plusieurs valeurs. Si k est trop petit, les données dans chaque groupe risque de ne pas avoir une similarité forte, et si k est trop grand, chaque groupe ne permet pas de découvrir des caractéristiques intéressantes.

Exemple

Un fruit de forme arrondie possède de nombreuses caractéristiques dont le poids et le rapport grand axe/petit axe. Les valeurs utilisées dans le programme sont aléatoires et les cerises sont plutôt grosses.

Étape 01 : création du jeu de données

Un traitement des données est effectué. Il est nécessaire afin que la caractéristique poids ne soit pas trop prévalente. Ici l'ensemble des données est noté `E` et il y a `k=3` groupes. Dans les étapes suivantes, pour conserver la généralité et adapter dans d'autres exemples, dans les procédures `genere_centres`, `partitionne` et `k_moyennes`, on prendra pour arguments les lettres `E` et `k` et non leurs valeurs.

```
In [1]: from random import randrange, choice, random
In [2]: E = [] # INITIALISATION LISTE
In[3] : classes_fruits = {} # INITIALISATION dictionnaire
    classes_fruits
In [4] : k = 3 # nombre de classes
In [5] : for i in range(50):
             poids = 65 + 20* random()
             rapport = 1.2 + 0.20 * random()
             E.append([poids,rapport])
             classes_fruits[i] = "mandarine"
IN [6]: for i in range(50,100):
             poids = 85 + 30* random()
             rapport = 1.4 + 0.40 * random()
             E.append([poids,rapport])
             classes_fruits[i] = "kiwi"
In [7]: for i in range(100,400):
             poids = 10 + 5* random()
             rapport = 1+ 0.20 * random()
             E.append([poids,rapport])
             classes_fruits[i] = "cerise"
In [8]: for elt in E:
             elt[1] = 100 * elt[1]
             # permet que rapport a une echelle
             # proche de poids
In [9]: def d(p1,p2):
             return (p1[0] -p2[0])**2 + (p1[1] -p2[1])**2
In [10] E[0:5]
# ce sont les poids et rapports des 5 premieres mandarines
Out[10]: [[75.86078914576923, 134.14448987615725],
 [83.67408121188569, 121.96558970927552],
 [76.30788312031973, 120.46897920304187],
 [82.83900713275025, 136.02710420467895],
 [74.84143226834124, 130.96360604974902]]
# puis on fait les derniere mandarines et premiers kiwis
In [11]: E[48:52]
Out[11]:[[75.03393068205956, 127.49644192598353],
 [66.36588321546508, 122.29631153910394],
 [100.68243818256286, 172.53653599760003],
 [88.43714662901907, 163.1368537164303]]
# puis les derniers kiwis et premieres cerises
In[12]: E[98:104]
Out[12]: [[92.78301461131325, 149.97338776724902],
 [100.42416911047346, 172.89241817528114],
 [14.150911251377163, 119.50627300346821],
 [11.630382309749942, 110.665021051174],
 [10.058103764037662, 106.28150059271017],
 [11.842895971664955, 114.60364946718724]]
```

Étape 02 : création des centres des classes

On rappelle que `randrange(n)` renvoie un nombre entier entre 0 et $n-1$.

```
In [1]: def genere_centres(E,k):
            n = len(E)
            liste = [] # L est liste vide initialement
            for i in range(k):
                x = randrange(n)
                # on met dans x un entier entre 0 et len(E)-1
                while x in liste :
                     x = randrange(n)
                    # permet de ne pas prendre deux fois le meme x
                liste.append(x)
            return liste
In [2]: centres = genere_centres(E,k)
In [3]: centres
Out[3]: [324,96,194]
In [4]: classes_fruits[324],classes_fruits[96], classes_fruits[194]
Out[4]: "cerise", "kiwi", "cerise"
```

Étape 03 : création des barycentres pour transformer les centres

Une fonction **barycentre** détermine les nouveaux centres à calculer à chaque itération. Le point appelé barycentre d'un ensemble de points est plus précisement l'isobarycentre de ces points.

```
In [4]: def barycentre(points):
            n = len(points)
            # n est le nombre de points ici 3 listes
            m = len(points[0])
            # m est la taille des listes ici taille 2
            s = [0] * m
            # initialement ici s = [0,0]
            for p in points :
                for k in range(m):
                    s[k] = s[k] + p[k]
                    # on somme les listes qui composent points
            s = [s[k]/n for k in range(m)]
            return s
In [5]: barycentre([[3,2] ,[4,1],[-1,0]])
Out[5]:    [2.0, 1.0]
In [6]: (3+4-1)/3 , (2+1+0)/3
Out[6]: (2.0, 1.0)
# C'est OK !
```

Étape 04 : création d'une fonction détection des centres les plus proches

Ici, on crée une fonction `ind_minimum` prend en paramètre une liste de distances entre un point et les centres de chaque groupe. Elle renvoie une liste contenant l'indice du centre le plus proche ou les indices des centres les plus proches (un point peut être équidistant de plusieurs centres).

```
In [1]: def ind_minimum(listes_distances):
            mini = min(listes_distances)
            # dans mini on met le minimum des distances
            choix = [] # on initialise avec liste vide
            for i in range(len(listes_distances)) :
                if listes_distances[i] == mini :
                    choix.append(i)
                    # on met dans choix les indices des distances = mini
                    # Warning choix peut contenir plusieurs indices
            return choix

In [2]: ind_minimum([1.5,3,0.45,4]), ind_minimum([1.5,3,0.45,4,0.45])
Out[2]: [2],  [2, 4]
```

On remarque donc que `ind_minimum` peut renvoyer une liste de plusieurs entiers.

Étape 05 : création de la partition initiale

La fonction `partitionne` qui va suivre crée la partition initiale. On retournera 3 données, la première `partition` donne une liste de trois listes (car $k = 3$ ici) qui donne une partition des fruits par leurs numéros. La seconde `new_centres` donne les centres sous forme de liste de 3 listes de longueurs 2, calculés avec la fonction **barycentre**. Enfin, la troisième est `classes` et qui donne la classe (entier $0, 1, 2$) de chaque numéro de fruit.

Noter l'usage de `choice(liste)` qui renvoie un élément de `liste` choisi aléatoirement. Ainsi `choix = choice(ind_minimum(di))` donne un des entiers de `ind_minimum(di)` choisi aléatoirement. En général, `di` a une seule valeur et le choix est vite fait !

```
In [3]: def partitionne(ens, centres, distance):
            n = len(ens) # ici n = 400 car ens est E
            k = len(centres) # ici centres =[324,96,194] donc k=3
            partition = [[c] for c in centres]
            # En begin partition =[[324],[96],[194]]
            classes = {centres[i]: i for i in range(k)}
            # classes est un dictionnaire et initialement
            # classes={324:0,96:1,194:2}
            for i in range(n):
                if i not in centres :
                   # on ne prend que des points de ens differents
                   # des points de centre
                   di = [distance(ens[i],ens[k]) for k in centres]
                   choix = choice(ind_minimum(di))
                   # si ind_minimum(di) a plusieurs valeurs
                   classes[i] = choix
                   # si par exemple choix retourne 2 et donc ens[194]
                   # est le centre le plus proche, on met ens[i]
                   # dans partition [2] qui contient 194 en premier elt
                   partition[choix].append(i)
            new_centres = [None] * k
            # On cree les nouveaux centres
            for i in range(k):
                points = [ens[j] for j in partition[i]]
                new_centres = barycentres (points)
            return partition, new_centres, classes
```

Faisons now tourner `partitionne(E, centres, d)` mais pour rendre plus visible, on va séparer `partition`, `c` et `classes`.
On tape : `partition, c, classes = partitionne(E,centres,d)` puis on tape `c`, on fait Enter, puis on tape `partition[0]`, on fait Enter, puis on tape `partition[1]`, on fait Enter, puis on tape `partition[2]`, on fait Enter et enfin on tape `classes` et on fait Enter.

Remarquons que `centres=[324,96,194]` et on voit que `324` et `194` sont les indices d'éléments de E ayant un petit poids car ce sont des cerises et par contre `96` est l'indice d'un kiwi donc de poids plus grand, `partition[1]` va regrouper en gros les kiwis et les mandarines et `partition[0]` et `partition[2]` auront en gros des cerises.

```
In [4]:  partition, c, classes = partitionne(E,centres,d)
# Donnons les centres
In [5]:  c
Out[5]:  [[12.291546972365632, 115.18165824251905],
 [87.31490977424329, 144.48256947424764],
 [12.717679188352552, 105.42253229808598]]

# Cela correspond bien a une cerise,
# plus un kiwi ou mandarine et une cerise

In [6]:  partition[0]
Out[6]: [324,100,101, 103,107, ... ,388,389,390,391,394,396,398]

In [7]:  partition[1]
Out[7]: [96,0,1,2, 3,...,95,97,98,99]

In [8]:  partition[2]
Out[8]:[194,102,104,105, 106,110, 111,115, 119, ..., 393,395,397, 399]

# pour l'instant dans chacune des partitions
# on a les kiwis et mandarines ensemble
# et les cerises dans les deux autres partitions

In [9]:  classes
Out[9]:  {324: 0, 96: 1, 194: 2, 0: 1,1: 1,2: 1,3: 1, ...,100: 0,101:
    0,102: 2,
 103: 0,104: 2,105: 2,..., 391: 0,392: 2, 393: 2,394: 0,395: 2,
 396: 0,397: 2,398: 0,399: 2}

# c'est bien un dictionnaire
```

Maintenant passons à la fonction qui va à chaque étape modifier nos partitions et quelques fruits vont se balader, surtout des kiwis et mandarines car le poids est assez similaire.

Étape 06 : création et exécution de la fonction algorithme des k-moyennes

Notons là `k_moyennes`.

```
In [50]: def k_moyennes(ens,centres, distance, max_iter):
    ...:     n = len(E)
    ...:     k = len(centres)
    ...:     partition,c,classes = partitionne(ens,centres,distance)
    ...:     iteration = 0 # Initialisation de iteration
    ...:     evolution = True  # par defaut evolution est True
    ...:     while evolution and(iteration < max_iter):
    ...:         # comme evolution est True initialement on commence
    ...:         iteration = iteration + 1
    ...:         # on ajoute 1 au compteur de iteration
    ...:         for i in range(n):
    ...:             di = [distance(ens[i],k) for k in c]
    ...:             # on calcule les distances de chaque ens[i] a chaque
    ...:             # elt c du centre actuel
    ...:             choix = choice(ind_minimum(di))
    ...:             if choix != classes[i] : # MODIFICATIOn
    ...:                 # On transporte ens[i] dans sa nouvelle classe
    ...:                 partition[classes[i]].remove(i)
    ...:                 classes[i] = choix
    ...:                 partition[choix].append(i)
    ...:                 evolution = True
    ...:         for i in range(k) : # CALCUL DES NEW CENTRES
    ...:             points = [ens[j] for j in partition[i]]
    ...:             c[i] = barycentre(points)
    ...:     return partition,c
    ...:     # retourne la partition finale et le new centre final
    ...:
```

```
# Avec la partition initiale
In [51]: k_moyennes(E,centres,d,0)
Out[51]:
([[324, 100, 104, 106,107,110,...,394,395,396],
  [96, 0, 1, 2, 3,...,95, 97, 98, 99],
  [194,11,16,32,47,101,102, ...,397,398,399]],
 [[12.45931373496306, 103.39213272271405],
  [88.71892694048223, 144.78640753801096],
  [13.621889815740184, 113.57118263237996]])
```

```
# Interessons nous aux centres obtenus apres une iteration
In [53]: partition, c = k_moyennes(E,centres,d,1)
In [54]: c
Out[54]:
[[12.47669179569801, 104.30609341256614],
 [87.84092588734582, 143.8832383519534],
 [12.59200997071767, 114.294706016
```

```
# Apres 5 iterations
In [55]: partition, c = k_moyennes(E,centres,d,5)
In [56]: c
Out[56]:
[[12.412711839744585, 105.07036697410766],
 [87.84092588734582, 143.8832383519534],
 [12.667331211151817, 114.99787322440679]]
```

```
# Apres 15 iterations
In [57]: partition, c = k_moyennes(E,centres,d,15)
In [58]: c
Out[58]:
[[12.400052961824144, 105.1045802812521],
 [87.84092588734582, 143.8832383519534],
 [12.68081609289292, 115.02991565482199]]
```

```
# Apres 25 iterations
In [59]: partition, c = k_moyennes(E,centres,d,25)
In [60]: c
Out[60]:
[[12.400052961824144, 105.1045802812521],
 [87.84092588734582, 143.8832383519534],
 [12.68081609289292, 115.02991565482199]]
# On voit que cela ne bouge plus
```

graphicx colortbl pst-tree,pst-node,pstricks
tikz
graphics graphicx pstricks,pst-node tikz

Chapitre 7

Algorithms and games

■ Objectifs

Les incontournables :

savoir

Résumé de cours

■ Introduction

Dans la recherche sur les jeux, on y étudie en particulier des notions de stratégie et d'équilibre. La recherche informatique s'est aussi développée avec les jeux sur les graphes qui ont des applications dans plusieurs domaines d'informatique théorique et des mathématiques en permettant de modéliser certains problèmes.
À partir d'une situation donnée, des joueurs à tour de rôle prennent une décision parmi un ensemble fini de décisions possibles, chaque décision amenant à une nouvelle situation. Nous considérons principalement dans la suite des **jeux à deux joueurs** satisfaisant certaines conditions :

- ☞ chaque joueur a la même vue d'ensemble de la situation (jeu à information complète) ;
- ☞ une décision est prise en fonction de la situation présente et pas des situations passées (jeu sans mémoire) ;
- ☞ cette décision ne dépend que de la situation et pas du joueur (jeu impartial) ;
- ☞ dans une situation donnée, une décision amène toujours à la même nouvelle situation (jeu sans hasard).

Les jeux sont orientés par des graphes orientés finis. Une situation, ou une position, est un sommet du graphe. Une décision est le choix d'un arc amenant à une nouvelle position. Dans un **jeu d'accessibilité**, chaque joueur souhaite atteindre un sous-ensemble de sommets particulier en se déplaçant à tour de rôle sur le graphe. Un jeu d'accessibilité à deux joueurs est modélisé par un **graphe biparti**.

■ Graphes bipartis

Définition Un graphe G est biparti si l'ensemble de ses sommets peut être divisé en deux sous-ensembles disjoints G_1 et G_2 tels que chaque arête de G a une extrémité dans G_1 et une extrémité dans G_2.

Exemple 1

Figure 1

On remarque dans la **figure 1** que $G_1 = \{1, 2, 3\}$ et $G_2 = \{4, 5, 6, 7\}$.

On peut employer aussi une technique de coloration ou utiliser la notion de cycle.

Exemple 2

FIGURE 2 + FIGURE 3

En effet, on démontre (théorème admis) qu'un graphe est biparti si et seulement s'il ne contient pas de cycles de longueurs impaires.

Dans la **figure 2**, la coloration du graphe est possible avec seulement deux couleurs sans que deux sommets voisins n'aient la même couleur. Le graphe possède un cycle pair $ABFEA$ (le nombre d'arcs constituant le cycle est pair).

Dans la **figure 3**, une coloration avec deux couleurs sans que deux sommets voisins n'aient la même couleur) est impossible. Et le graphe possède un cycle impair $ABEA$.

Donnons un programme pour tester si un graphe possède un cycle impair. On utilise une file. Et en particulier un `deque` du package `collection`. Le mot `deque` vient de double-ended queue, c'est une généralisation des piles et des files. On peut faire des ajouts et retraits à chaque extremité contrairement aux files qui sont FIFO (first in first out) et piles qui sont LIFO (last in first out). La commande `popleft()` retourne l'élément le plus à gauche.

```
In [1]: from collections import deque

In [2]: def cycle_impair(graphe, sommet):
            niveaux = {s : None for s in graphe}
            niveaux[sommet] = 0
            file = deque()
            file.append(sommet)
            while len(file) > 0 :
                sommet = file.popleft()
                for s in graphe[sommet]:
                    if niveaux[s] is None :
                        niveaux[s] = niveaux[sommet] + 1
                        file.append(s)
                    elif niveaux[s] == niveaux[sommet]:
                        return True
            return False
```

On teste les graphes de la **figure 2** et **figure 3**. Les graphes sont représentés par des dictionnaires.

```
In [1]: g1 = {'A' : ['B','E'] , 'B' : ['A', 'C', 'D', 'F' ], 'C': ['B'],
'D' : ['B'] , 'E' : ['A', 'F'], 'F' : ['B', 'E' ]}

In [2]: cycle_impair(g1,'A')
Out[2]: False
```

```
In [3]: g2 = {'A' : ['B','E'] , 'B' : ['A', 'C', 'D', 'E', 'F' ], 'C' :
['B'], 'D' : ['B'] , 'E' : ['A', 'B', 'F'], 'F' : ['B', 'E' ] }

In [4]: cycle_impair(g2,'A')
Out[4]: True
```

Essayons de tracer l'algorithme. J'ai rajouté plein de print
On va reprendre `g1` et `g2` avec le sommet `'A'`

```
In [3]: g1 = {'A' : ['B','E'], 'B': ['A','C','D','F'], 'C': ['B'], 'D' :
    ['B'], 'E' : ['A','F'], 'F' : ['B','E']}

In [4]: cycle_impair(g1,'A')
niveaux initial est {'A': 0, 'B': None, 'C': None, 'D': None, 'E': None,
    'F': None}
file initial est deque(['A'])
le sommet est A
file vaut now deque(['B'])
niveaux vaut now {'A': 0, 'B': 1, 'C': None, 'D': None, 'E': None, 'F':
    None}
file vaut now deque(['B', 'E'])
niveaux vaut now {'A': 0,'B': 1,'C': None,'D': None,'E': 1,'F': None}
file apres parcours graphe est deque(['B', 'E'])
niveaux apres parcours graphe est {'A': 0, 'B': 1, 'C': None, 'D': None,
    'E': 1, 'F': None}
le sommet est B
file vaut now deque(['E', 'C'])
niveaux vaut now {'A': 0, 'B': 1, 'C': 2, 'D': None, 'E': 1, 'F': None}
file vaut now deque(['E', 'C', 'D'])
niveaux vaut now {'A': 0, 'B': 1, 'C': 2, 'D': 2, 'E': 1, 'F': None}
file vaut now deque(['E', 'C', 'D', 'F'])
niveaux vaut now {'A': 0, 'B': 1, 'C': 2, 'D': 2, 'E': 1, 'F': 2}
file apres parcours graphe est deque(['E', 'C', 'D', 'F'])
niveaux apres parcours graphe est {'A': 0, 'B': 1, 'C': 2, 'D': 2, 'E':
    1, 'F': 2}
le sommet est E
file apres parcours graphe est deque(['C', 'D', 'F'])
niveaux apres parcours graphe est {'A': 0, 'B': 1, 'C': 2, 'D': 2, 'E':
    1, 'F': 2}
le sommet est C
file apres parcours graphe est deque(['D', 'F'])
niveaux apres parcours graphe est {'A': 0, 'B': 1, 'C': 2, 'D': 2, 'E':
    1, 'F': 2}
le sommet est D
file apres parcours graphe est deque(['F'])
niveaux apres parcours graphe est {'A': 0, 'B': 1, 'C': 2, 'D': 2, 'E':
    1, 'F': 2}
le sommet est F
file apres parcours graphe est deque([])
niveaux apres parcours graphe est {'A': 0, 'B': 1, 'C': 2, 'D': 2, 'E':
    1, 'F': 2}
Out[4]: False
```

On a donc décortiqué la bestiole.

```
In [5]: g2 = {'A' : ['B','E'], 'B': ['A','C','D','E','F'], 'C': ['B'],
'D' : ['B'], 'E' : ['A','B','F'], 'F' : ['B','E']}

In [6]: cycle_impair(g2,'A')
niveaux initial est {'A': 0, 'B': None, 'C': None, 'D': None, 'E': None,
    'F': None}
file initial est deque(['A'])
le sommet est A
file vaut now deque(['B'])
niveaux vaut now {'A': 0, 'B': 1, 'C': None, 'D': None, 'E': None, 'F':
    None}
file vaut now deque(['B', 'E'])
niveaux vaut now {'A': 0,'B': 1,'C': None,'D': None, 'E': 1,'F': None}
file apres parcours graphe est deque(['B', 'E'])
niveaux apres parcours graphe est {'A': 0, 'B': 1, 'C': None, 'D': None,
    'E': 1, 'F': None}
le sommet est B
file vaut now deque(['E', 'C'])
niveaux vaut now {'A': 0, 'B': 1, 'C': 2, 'D': None, 'E': 1, 'F': None}
file vaut now deque(['E', 'C', 'D'])
niveaux vaut now {'A': 0, 'B': 1, 'C': 2, 'D': 2, 'E': 1, 'F': None}
Out[6]: True
```

On a décortiqué encore la bestiole. Ici, on boucle moins car on retourne `True` à un moment donné.

■ Jeux à deux joueurs

Définitions, vocabulaire

On considère un jeu à deux joueurs J_1 etJ_2 sur un graphe orienté fini $G = (S, A)$, où S est l'ensemble fini des sommets et A l'ensemble des arcs (ou arêtes). Chaque joueur possède son ensemble de sommets, respectivement S_1 et S_2. On appelle **graphe du jeu** ou **arène** le triplet (G, S_1, S_2). Les ensembles S_1 et S_2 constituent une partition de S : $S_1 \cap S_2 = \emptyset$ et $S_1 \cup S_2 = S$.

On dit que les sommets de S_1 sont contrôlés par J_1 et les sommets de S_2 sont contrôlés par J_2.

Le graphe du jeu est évidemment biparti : une arête joint un sommet contrôlé par un joueur avec un sommet contrôlé par l'autre joueur.

On marque chaque sommet par un entier (chaque sommet est étiqueté) $0, 1, 2, ...$ Une partie est alors un chemin a priori infini dans le graphe (on suppose les deux joueurs immortels). Un jeton est placé sur un sommet initial. Le joueur qui contrôle ce sommet déplace le jeton selon une arête qu'il a choisit jusqu'à un sommet voisin et c'est l'autre joueur qui prend le relai dans le déplacement du jeton et ainsi de suite.

Dans un jeu d'accessibilité, un joueur J_1 gagne lorsqu'il atteint un sous-ensemble de sommets $F \subset S$. Un sommet de F est un sommet final. L'ensemble des sommets de F sont les états gagnants pour le joueur J_1. Il y a bien entendu un sous-ensemble de sommets constitué des états gagnants pour l'autre joueur et un sous-ensemble de sommets consitué des états de partie nulle (aucun des joueurs ne perd ni ne gagne).
Un jeu est dit **à somme nulle** si toute partie gagnée par l'un des deux joueurs est perdue par l'autre joueur.

Un jeu d'accéssibilité à somme nulle est alors un quadruplet (G, S_1, S_2, F), où (G, S_1, S_2) est une arène et F l'ensemble des sommets gagnants pour J_1. Si W est l'ensemble des chemins contenant un sommet de F, on peut définir le jeu par le quadruplet (G, S_1, S_2, W). L'ensemble W est appelé **condition de gain**.

Exemple 1

FIGURE 4

Ici, on a des cercles pour les sommets contrôlés par J_1 et des carrés pour les sommets contrôlés par J_2.
Un jeton est déposé sur le sommet 0 qui est contrôlé par J_1. C'est donc J_1 qui commence et il peut déplacer le jeton soit jusqu'au sommet 1, soit jusqu'au sommet 2. Ces deux sommets sont contrôlés par J_2 qui commence alors.
Une partie est par exemple le chemin

0 1 3 1 4 1 3 1 4 1 3 1 4 ...

On remarque que sur ce chemin, on reste sur un cycle pair.

Le joueur J_1 gagne la partie s'il atteint un sommet d'un certain sous-ensemble F de S. Dans ce cas, le joueur J_2 perd la partie. Pour que J_2 ne perde pas la partie, il faut que le chemin ne passe pas par un des sommets de F.

Exemple 2 : jeu de Nim

Le jeu de Nim remonte à la chine antique où on l'appelait sous le nom de **fan-fan** et en Afrique sous le nom de **tiouk-tiouk**. Le nom actuel est une déformation de Nimm ! = Prends !

Le jeu de Nim se joue avec plusieurs tas d'objets identiques, classiquement des allumettes (ou des jetons, des pièces etc.) Chaque joueur à tour de rôle choisit un tas et le nombre d'objets, au moins un, qu'il retire de ce tas. Le dernier joueur à retirer des objets a gagné la partie.
Il existe différentes variantes, par exemple prenons le cas d'un seul tas de 15 objets et chaque joueur peut retirer 1, 2 ou 3 objets du tas.

On peut faire le cheminement suivant :

15 objets dans le tas à cette étape : J_1 joue et retire 3 objets ;

12 objets dans le tas à cette étape : J_2 joue et retire 2 objets ;

10 objets dans le tas à cette étape : J_1 joue et retire 2 objets ;

8 objets dans le tas à cette étape : J_2 joue et retire 3 objets ;

5 objets dans le tas à cette étape : J_1 joue et retire 1 objet ;

4 objets dans le tas à cette étape : J_2 joue et retire 1 objet ;

3 objets dans le tas à cette étape : J_1 joue et retire 3 objets et gagne la partie.
La stratégie qui permet à J_1 de gagner est de laisser à J_2 un multiple de 4 objets dans le tas.

Exemple 3 : jeu de Marienbad

Ce jeu est en fait la somme de quatre jeux de Nim à un seul tas mais en version inverse, c'est-à-dire que le joueur qui prend le dernier objet a perdu la partie.
Exercice. *Les quatre tas t_A, t_B, t_C et t_D sont constitués respectivement de* 1, 3 , 5 *et* 7 *allumettes. Chaque joueur peut retirer à tour de rôle des allumettes, au moins une, dans un seul tas. Le joueur qui prend en dernier a perdu.*
Illustrer par des dessins où à chaque étape on visualise les quatre tas. Au départ du jeu, le joueur J_1 prend 3 *allumettes dans t_D. Puis ensuite J_2 prend* 2 *allumettes dans t_B. Puis ensuite J_1 prend* 3 *allumettes dans t_C. Puis ensuite J_2 prend* 2 *allumettes dans t_D. Puis J_1 prend les* 2 *dernieres allumettes de t_D. C'est à J_2 de jouer. Peut-il encore gagner ou est-ce cuit pour lui ?*

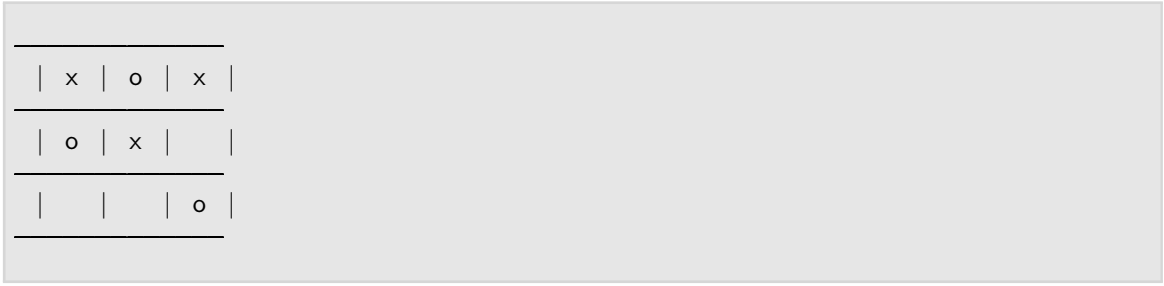
Exemple 4 : jeu tic-tac-toe ou encore jeu du morpion ou jeu oxo

Un symbole **x** ou **o** est attribué à chaque joueur. Les deux joueurs tracent alternativement dans une case vide d'une grille carrée de neuf cases leur symbole. Le gagnant est le premier à obtenir une ligne horizontale, verticale ou en diagonale de trois symboles identiques.
Ce jeu a plusieurs variantes. La grille peut être plus grande, un joueur doit aligner 5 symboles identiques pour marquer un point etc.

Donnons un programme affichant le jeu de base à 9 cases. On fait la barre verticale avec Alt Gr 6

```
In [1]: jeu = {i: " " for i in range(9)}
In [2]: def affiche(jeu):
            print(13 * "-")
            print(" | "+ jeu[0]+" | "+jeu[1] + " | " +jeu[2] + " | ")
            print(13* "-")
            print(" | "+jeu[3] + " | " +jeu[4] + " | " +jeu[5] + " | ")
            print(13 * "-")
            print(" | "+jeu[6] + " | " +jeu[7] + " | " +jeu[8] + " | ")
            print(13* "-")
```

```
In [3]: jeu[0] ="x";jeu[3] ="o" ; jeu[4] ="x" ; jeu[8]="o" ; jeu[2]="x"
    ; jeu[1]="o"
In [4]: affiche(jeu)
```



Exemple 5 : jeu de chomp ou du chocolat

Le jeu a été inventé en 1952 par Fred Schuh en termes de choix de diviseurs à partir d'un entier donné puis réinventé par David Gate sous sa forme actuelle. to chomp = mâcher.

Une tablette de chocolat est composée de n rangées horizontales et m rangées verticales. Si la tablette a pour dimension n et m, chaque carré est représenté par le couple (p, q) avec $1 \leqslant p \leqslant n$ et $1 \leqslant q \leqslant m$.
On suppose que le carré $(1, 1)$ ne doit pas être mangé car il est empoisonné. On le met en gris. L'idée est que chaque joueur à tour de rôle mange des carrés, ce qui réduit la tablette à chaque étape. Maintenant il y a plusieurs façons de réduire la tablette en enlevant des carrés. Donnons pour commencer le protocole classique du jeu de chomp.

Protocole 1 (le classique) : Un joueur choisit un carré et retire tous les carrés qui se trouvent à droite et en dessous au sens large, c'est-à-dire qu'il retire aussi le carré (a, b). On dit qu'il mange (ou chomp) les carrés enlevés. Ainsi si le carré choisi est (a, b), le joueur retire tous les carrés (p, q) tels que $p \geqslant a$ et $q \geqslant b$. Si après le coup d'un joueur, il ne reste que le carré $(1, 1)$, comme on ne peut pas manger un carré empoisonné, ce joueur est déclaré gagnant car l'autre joueur est bloqué et a donc perdu.

***Exercice** : dessiner une tablette de 20 carrés avec $n = 4$ et $m = 5$. Le joueur J_1 choisit le carré $(3, 5)$. Retirer les carrés adéquats et dessiner la tablette restante. Puis le joueur J_2 choisit le carré $(2, 2)$. Retirer les carrés adéquats. Puis J_1 choisit $(1, 2)$. Retirer les carrés adéquats. Puis J_2 joue et choisit $(2, 1)$. Conclusion ?*

Protocole 2 On fait une autre variante du jeu précédent. Un joueur coupe suivant une verticale (respectivement une horizontale) et mange les rangées à droite de la verticale (respectivement en dessous de l'horizontale). Le premier joueur qui obtient après sa coupe une tablette carrée a gagné dans cette variante.

■ Stratégie

Une stratégie permet de préciser ce qui doit être joué dans chaque situation. Élaborer une stratégie peut être très compliqué et on se contente de stratégies qui ne prennent en compte que le sommet où se trouve le jeton et qu'on appelle **stratégie positionnelle**.

Sur un graphe, une stratégie pour un joueur J_i est une fonction γ qui a tout sommet s de S_i associe un sommet de S avec $(s, \sigma(s)) \in A$. Cette fonction précise le sommet à atteindre depuis la position s. Un joueur J_i respecte une stratégie si pour toute partie $s_0 s_1 s_2 ...$ et tout $k \geqslant 0$, si $s_k \in S_i$ alors $s_{k+1} = \sigma(s_k)$.

Exercice : *reprenons le jeu de chomp avec le protocole 1 et supposons que J_1 a pour stratégie de choisir au départ le carré $(2, 2)$. En déduire les choix possibles pour J_2 puis J_1 quand c'est de nouveau son tour. Qui va gagner ?*

■ Stratégie gagnante

Dans un jeu d'accessibilité, le joueur J_1 gagne une partie si la partie contient un sommet d'un ensemble F donné. On dit qu'une stratégie est gagnante pour J_1 depuis un sommet s_k si pour toute partie contenant s_k la partie est gagnée par J_1 s'il respecte la stratégie depuis s_k, et ceci, quelle que soit la stratégie suivie par l'autre joueur. On dit alors que s_k est une **position gagnante** pour J_1.
Une stratégie est gagnante pour J_1 si la stratégie est gagnante pour J_1 depuis s_0.
Voir l'exercice précédent où le choix de $(2, 2)$ comme s_0 est une stratégie gagnante pour J_1.

Donnons un autre exemple.
Exemple Considérons la figure suivante dans lequel les sommets en forme de cercle sont controlés par J_1 et les sommets en forme de carré par J_2.

FIGURE 5

On pose un jeton sur le sommet marqué 0, contrôlé par J_1 et la partie débute sur ce sommet. Chacun son tour, les deux joueurs déplacent le jeton en suivant une arête. Une partie est gagnée par J_1 si le jeton arrive sur le sommet marqué 6.
Une stratégie gagnante pour J_1 est de déplacer le jeton sur le sommet 1. Le joueur J_2 ne peut alors déplacer le jeton que sur l'un des sommets 3 ou 4. Le joueur J_1 peut alors tranquillement aller au sommet 6 au tour suivant.

■ Algorithme min-max

Introduction de l'algorithme mini-max : position du problème

Considérons l'exemple de jeu de Nim à un tas. On suppose ici que le tas initial a 5 objets. Chaque joueur peut retirer un, deux ou trois objets du tas. Le premier qui ne peut plus retirer un objet perd la partie. Il n'y a pas de partie nulle.
On peut représenter le jeu par le graphe biparti qui suit avec à gauche les sommets contrôlés par J_1 et ¨ a droite ceux contrôlés par J_2.
Les sommets sont étiquetés par le nombre d'objets restants.

FIGURE 6

Une autre manière de représenter le jeu est de faire un arbre de décision. Les sommets de l'arbre appelés **noeuds** sont des positions et les arêtes sont les coups joués par les joueurs. Un noeud particulier est la **racine** de l'arbre. Les **enfants** d'un noeud n sont les noeuds correspondant aux positions qui suivent la position correspondant à n. Ce sont les racines des sous-arbres de n. Les feuilles de l'arbre sont les noeuds correspondants aux positions finales. Les feuilles n'ont pas d'enfants, pas de sous-arbres. Le nombre de feuilles est égal au nombre de parties différentes. Pour le jeu en exemple, l'arbre obtenu est représenté avec les sommets contrôlés par J_2 en gris.

FIGURE 7

Remarque : On remarque qu'une case blanche avec 0 signifie que J_1 ne peut plus rien retirer et il a perdu (il y a 6 telles cases) et qu'une case grise avec 0 signifie que J_2 ne peut plus rien retirer et il a perdu (il y a 7 telles cases). Cela signifie que J_1 a plus de chance de gagner que J_2. Et cela sans que J_1 démarre avec une stratégie gagnante. Donc J_1 a un gros avantage d'être en premier.

On peut parcourir l'arbre, comme on parcourt un graphe, pour chercher à partir de chaque position le meilleur coup à jouer.
On utilise pour cela un algorithme comme l'algorithme **min-max** ou encore appelé **minimax**.

Principe de l'algorithme mini-max

On se place du côté d'un joueur, par exemple J_1 (car on a vu qu'il a plus une tête de winner que le joueur J_2).
Chaque feuille de l'arbre indique si J_1 est gagnant ou perdant. On ajoute donc une marque ou une valeur au noeud, $+1$ s'il est gagnant et -1 s'il est perdant.
Dans un jeu où il peut y avoir des parties nulles, on ajoute 0. Ensuite, on définit les valeurs pour les autres noeuds de manière récursive.

- ☞ Si le noeud est contrôlé par J_1, sa valeur est le maximum des valeurs de ses sous-arbres.
- ☞ Si le noeud est contrôlé par J_2, sa valeur est le minimum des valeurs de ses sous-arbres.

Le but est pour J_1 de maximiser ses gains et pour J_2 de minimiser ses pertes.

Exercice : *Reprendre l'arbre de la figure* 7*, les feuilles étant les cases numérotés* 0*, commencer par affecter les feuilles blanches de* -1 *(*J_1 *a perdu) et on écrit dans la case alors* $0 : -1$ *et les feuilles grises de* $+1$ *(*J_1 *a gagné) et on écrit alors dans la case* $0 : +1$*. Puis suivre l'algorithme définit plus haut en remplaçant un noeud numérotée* n *par* $n : -1$ *ou* $n : +1$ *puis en remontant jusqu'à la racine qui est une case blanche numérotée* 5*. Dans cette case met-on* $5 : +1$ *ou* $5 : -1$ *?*

L'objectif est de déterminer une stratégie ou des positions gagnantes. Si un noeud a une valeur positive, alors J_1 dispose à partir de cette position d'une stratégie pour gagner la partie. Sur l'exemple J_1 a une stratégie gagnante depuis la position initiale.