

Trier et rechercher

INFORMATIQUE COMMUNE - TP n° 1.4 - Olivier Reynet

À la fin de ce chapitre, je sais :

- ✍ coder un algorithme de tri simple et explicite
- ✍ évaluer le temps d'exécution d'un algorithme avec la bibliothèque `time`
- ✍ rechercher un élément dans un tableau séquentiellement ou par dichotomie itérative
- ✍ générer un graphique légendé avec la bibliothèque `matplotlib`

A Trier un tableau

- A1. On souhaite trier des listes Python, considérées ici comme des tableaux, avec des algorithmes différents (cf. algorithmes 1, 2 et 3). Chaque algorithme de tri est implémenté par une fonction Python. Le prototype de ces fonctions est `my_sort(t)`, où `t` est un paramètre formel qui représente le tableau à trier.

```
1  def my_sort(t):
2      # tri du tableau
3      # for i in range(len(t))
4      #     t[i] = ...
```

Cette fonction, une fois réalisée, trie le tableau `t` passé en paramètre mais ne renvoie rien (i.e. pas de `return`). Expliquer pourquoi.

Solution : Une liste Python est une variable muable, donc on peut modifier `t` sans pour autant changer la référence de `t` en mémoire. Les modifications apportées par l'algorithme sur `t` le sont directement sur la liste passée en paramètre effectif par référence (toute variable Python étant une référence vers une case mémoire). C'est pourquoi on n'a pas besoin d'effectuer un `return` pour récupérer le travail de la fonction.

- A2. Coder les algorithmes de tri par sélection, par insertion et par comptage en respectant le prototype défini à la question précédente¹.

Solution :

```
1  def selection_sort(t):
2      for i in range(len(t)):
3          min_index = i
```

1. On a le droit de collaborer, de se répartir les algorithmes et de s'échanger les codes s'ils sont corrects!

```

4     for j in range(i + 1, len(t)):
5         if t[j] < t[min_index]:
6             min_index = j
7     swap(t, i, min_index)
8
9
10    def insertion_sort(t):
11        for i in range(1, len(t)):
12            to_insert = t[i]
13            j = i
14            while t[j - 1] > to_insert and j > 0:
15                t[j] = t[j - 1]
16                j -= 1
17            t[j] = to_insert
18
19
20    def counting_sort(t):
21        if len(t) > 0:
22            v_max = t[0]
23            for i in range(1, len(t)):
24                if t[i] > v_max:
25                    v_max = t[i] # trouver le maximum
26            count = [0 for _ in range(v_max + 1)]
27            for i in range(len(t)):
28                count[t[i]] += 1 # compter les occurrences de chaque t[i] de t.
29                # Certaines cases de count sont nulles : pas de valeur
30                # correspondante dans t
31            for i in range(1, v_max+1):
32                count[i] += count[i - 1] # cumul des effectifs
33            output = [None for _ in range(len(t))]
34            for i in range(len(t)):
35                count[t[i]] -= 1 # on compte à partir de 0
36                output[count[t[i]]] = t[i] # à la bonne place !
37                #print("elem :", t[i], "place :", count[t[i]], output)
38            return output
39        else:
40            return None

```

- A3. Tester ces algorithmes sur une **même** liste Python de longueur 20 et contenant de types `int` choisis aléatoirement entre 0 et 100.

Solution :

```

1  N = 20
2  M = 100
3
4  t = [randint(0, M) for _ in range(N)]
5  copy_t = t[:] # création d'une copie du tableau t
6              # pas comme copy_t = t, référence seulement
7  selection_sort(copy_t)
8  print(copy_t)
9  copy_t = t[:]
10 insertion_sort(copy_t)

```

```

11 print(copy_t)
12 copy_t = t[:]
13 counting_sort(copy_t)
14 print(copy_t)

```

- A4. Peut-t-on trier des listes de chaînes de caractères avec ces mêmes codes? Tester cette possibilité à l'aide de la liste ["Zorglub", "Spirou", "Fantasio", "Marsupilami", "Marsu", "Samovar", "Zantafio"]. Analyser les résultats. Pourquoi est-ce possible? Pourquoi n'est-ce pas possible?

Solution : Les tris par sélection ou par insertion sont des tris comparatifs : ils utilisent la comparaison de deux éléments du tableau pour trier les éléments. Lorsque les éléments sont des entiers, cela fonctionne grâce à l'ordre sur les entiers naturels. Lorsque les éléments sont des chaînes de caractères, c'est l'ordre lexicographique qui est utilisé. Cet ordre s'appuie sur l'ordre alphabétique et la position du caractère dans la chaîne pour comparer deux chaînes. Python utilise implicitement l'ordre lexicographique lorsque on compare deux chaînes de caractères avec l'opérateur <. C'est pourquoi les tris par insertion, sélection ou bulles fonctionnent aussi dans ce cas.

Par contre, le tri par comptage n'est pas un tri comparatif : c'est un tri par dénombrement de valeurs entières. Il ne porte donc que sur des tableaux contenant des entiers et ne peut pas fonctionner pour des chaînes de caractères.

Algorithme 1 Tri par sélection

```

1: Fonction TRIER_SELECTION(t)
2:   n ← taille(t)
3:   pour i de 0 à n - 1 répéter
4:     min_index ← i                                ▷ indice du prochain plus petit
5:     pour j de i + 1 à n - 1 répéter              ▷ pour tous les éléments non triés
6:       si t[j] < t[min_index] alors
7:         min_index ← j                            ▷ c'est l'indice du plus petit non trié!
8:     échanger(t[i], t[min_index])                 ▷ c'est le plus grand des triés!

```

Algorithme 2 Tri par insertion

```

1: Fonction TRIER_INSERTION(t)
2:   n ← taille(t)
3:   pour i de 1 à n-1 répéter
4:     à_insérer ← t[i]
5:     j ← i
6:     tant que t[j-1] > à_insérer et j>0 répéter
7:       t[j] ← t[j-1]                                ▷ faire monter les éléments
8:       j ← j-1
9:     t[j] ← à_insérer                                ▷ insertion de l'élément

```

Algorithme 3 Tri par comptage

```

1: Fonction TRIER_COMPTEGE( $t, v_{max}$ )                                ▷  $v_{max}$  est le plus grand entier à trier
2:    $n \leftarrow \text{taille}(t)$ 
3:    $c \leftarrow$  un tableau de taille  $v_{max} + 1$  initialisé avec des zéros
4:   pour  $i$  de 0 à  $n - 1$  répéter
5:      $\text{elem} \leftarrow t[i]$                                 ▷ compter les occurrences de chaque élément du tableau.
6:      $c[\text{elem}] \leftarrow c[\text{elem}] + 1$ 
7:   pour  $i$  de 1 à  $v_{max}$  répéter
8:      $c[i] \leftarrow c[i] + c[i-1]$                                 ▷ cumuler les effectifs pour faire le lien avec la position
9:    $\text{résultat} \leftarrow$  un tableau de taille  $n$ 
10:  pour  $i$  de 0 à  $n$  répéter
11:     $\text{elem} \leftarrow t[i]$                                 ▷ On va placer au bon endroit une occurrence de l'élément elem
12:     $c[\text{elem}] \leftarrow c[\text{elem}] - 1$                                 ▷ Un de moins, et on compte à partir de 0
13:     $\text{résultat}[c[\text{elem}]] \leftarrow \text{elem}$ 
14:  renvoyer  $\text{résultat}$ 

```

La bibliothèque `matplotlib` permet de générer des graphiques à partir de données de type `list` qui constituent les abscisses et les ordonnées associées. La démarche à suivre est de :

- importer la bibliothèque `from matplotlib import pyplot as plt`
- créer une figure `plt.figure()`
- tracer une courbe `plt.plot(x,y)` si x et y sont les listes des abscisses et des ordonnées associées. La bibliothèque trace les points $(x[i], y[i])$ sur le graphique.
- ajouter les éléments de légende et de titre,
- montrer la figure ainsi réalisée `plt.show()`.

La bibliothèque `time` permet notamment de mesurer le temps d'exécution d'un code. Un exemple de code utilisant ces deux bibliothèques est donné ci-dessous. Le graphique qui en résulte est montré sur la figure 1.

Code 1 – Exemple d'utilisation des bibliothèques `time` et `matplotlib`

```

1 import time
2 from matplotlib import pyplot as plt
3
4
5 def to_measure(d):
6     time.sleep(d) # Do nothing, wait for d seconds
7
8
9 # Simple use
10 tic = time.perf_counter()
11 to_measure(0.1)
12 toc = time.perf_counter()
13
14 print(f"Execution time : {toc - tic} seconds")
15
16 # Plotting results
17 timing = []
18 delay = [d / 1000 for d in range(1, 100, 5)]

```

```
19 for d in delay:
20     tic = time.perf_counter()
21     to_measure(d)
22     toc = time.perf_counter()
23     timing.append(toc - tic)
24
25 plt.figure()
26 plt.plot(delay, timing, color='cyan', label='fonction to_measure')
27 plt.xlabel('Delay', fontsize=18)
28 plt.ylabel("Execution time", fontsize=16)
29 plt.legend()
30 plt.show()
```

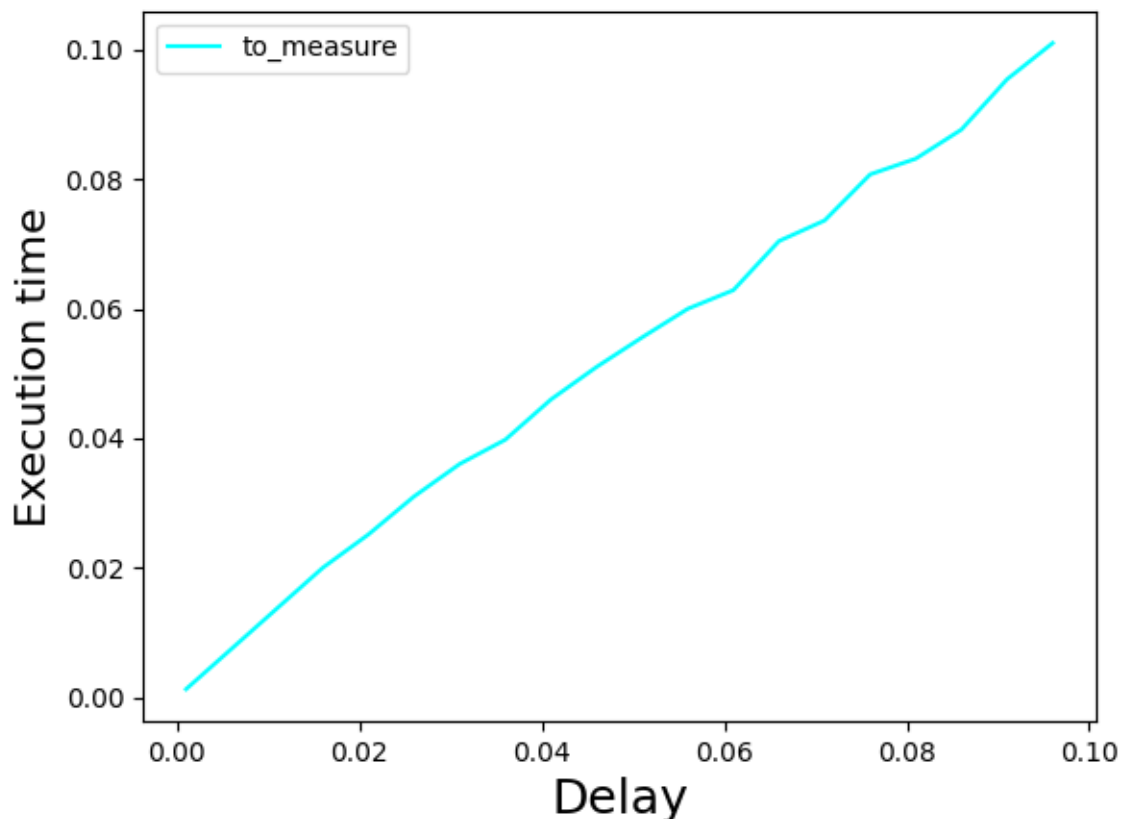


FIGURE 1 – Figure obtenue à partir des bibliothèques matplotlib et time et du code 1

- A5. À l'aide de la bibliothèque matplotlib, tracer les temps d'exécution nécessaires au tri d'un même tableau d'entiers par les algorithmes implémentés. On pourra également les comparer à la fonction `sorted` de Python. Analyser les résultats. Essayer de qualifier les coûts des algorithmes en fonction de la taille du tableau d'entrée.

Solution :**Code 2 – Trier des tableaux**

```
1 import time
2 from random import randint
3
4
5 def swap(t, i, j):
6     t[i], t[j] = t[j], t[i]
7
8
9 def selection_sort(t):
10     for i in range(len(t)):
11         min_index = i
12         for j in range(i + 1, len(t)):
13             if t[j] < t[min_index]:
14                 min_index = j
15         swap(t, i, min_index)
16
17
18 def insertion_sort(t):
19     for i in range(1, len(t)):
20         to_insert = t[i]
21         j = i
22         while t[j - 1] > to_insert and j > 0:
23             t[j] = t[j - 1]
24             j -= 1
25         t[j] = to_insert
26
27
28 def bubble_sort(t):
29     for i in range(len(t) - 1):
30         for j in range(len(t) - 1):
31             if t[j] > t[j + 1]:
32                 swap(t, j, j + 1)
33
34
35 def complexity_counting_sort(t):
36     v_max = max(t)
37     count = [0] * (v_max + 1)
38     for i in range(len(t)):
39         count[t[i]] += 1
40     for i in range(1, v_max + 1):
41         count[i] += count[i - 1]
42     output = [None for i in range(len(t))]
43     for i in range(len(t)):
44         output[count[t[i]] - 1] = t[i]
45         count[t[i]] -= 1
46     return output
47
48
49 def counting_sort(t):
50     if len(t) > 0:
51         v_max = t[0]
52         for i in range(1, len(t)):
```

```

53         if t[i] > v_max:
54             v_max = t[i] # trouver le maximum
55     count = [0 for _ in range(v_max + 1)]
56     for i in range(len(t)):
57         count[t[i]] += 1 # compter les occurrences de chaque t[i] de t.
58         # Certaines cases de count sont nulles : pas de valeur
59         # correspondante dans t
60     for i in range(1, v_max+1):
61         count[i] += count[i - 1] # cumul des effectifs
62     output = [None for _ in range(len(t))]
63     for i in range(len(t)):
64         count[t[i]] -= 1 # on compte à partir de 0
65         output[count[t[i]]] = t[i] # à la bonne place !
66         print("elem :", t[i], "place :", count[t[i]], output)
67     return output
68 else:
69     return None
70
71 def sort_timing():
72     sizes = [i for i in range(10, 5000, 500)]
73     M = 5 * max(sizes)
74     results = []
75     for i in range(len(sizes)):
76         t = [randint(0, M) for _ in range(sizes[i])]
77         mem_t = t[:]
78         results.append([])
79         for method in [selection_sort, bubble_sort, insertion_sort,
80                        counting_sort, sorted]:
81             t = mem_t[:]
82             tic = time.perf_counter()
83             method(t)
84             toc = time.perf_counter()
85             results[i].append(toc - tic)
86         print("#", i, " - ", sizes[i], " -> ", results)
87
88     sel = [results[i][0] for i in range(len(sizes))]
89     bub = [results[i][1] for i in range(len(sizes))]
90     ins = [results[i][2] for i in range(len(sizes))]
91     cnt = [results[i][3] for i in range(len(sizes))]
92     p = [results[i][4] for i in range(len(sizes))]
93
94     from matplotlib import pyplot as plt
95
96     plt.figure()
97     plt.plot(sizes, sel, color='red', label='selection sort')
98     plt.plot(sizes, bub, color='green', label='bubble sort')
99     plt.plot(sizes, ins, color='blue', label='insertion sort')
100    plt.plot(sizes, cnt, color='cyan', label='counting sort')
101    plt.plot(sizes, p, color='black', label='sorted python')
102    plt.xlabel('n', fontsize=18)
103    plt.ylabel('time', fontsize=16)
104    plt.legend()
105    plt.show()

```

```
106
107 # MAIN PROGRAM
108 N = 20
109 M = 100
110
111 t = [randint(0, M) for _ in range(N)]
112
113 print(t, max(t))
114 print(counting_sort(t))
115
116 # for method in [selection_sort, insertion_sort, bubble_sort, counting_sort]:
117 #     tmp = t[:]
118 #     if method in [selection_sort, bubble_sort, insertion_sort]: # in-place
119 #         method(tmp)
120 #     else:
121 #         tmp = method(tmp) # not in-place
122 #     print("--> Method", method.__name__, ":\n ", t, "\n", tmp)
123 #
124 # t = ["Zorglub", "Spirou", "Fantasio", "Marsupilami", "Marsu", "Samovar", "
125 #       Zantafio"]
126 # for method in [selection_sort, insertion_sort, bubble_sort]: # ,
127 #     counting_sort:
128 #         method(t)
129 #         print("--> Method", method.__name__, ":\n ", t, "\n", t)
130 #
131 # sort_timing()
```



B Tri bulle (Facultatif)

Le tri bulle est un tri par comparaisons et échanges, stable et en place. Il n'est pas efficace sauf dans le cas où le tableau à trier est quasiment trié. Sa vocation reste essentiellement pédagogique.

Le principal avantage du tri bulle est qu'il est simple à expliquer. Son principe est le suivant : on prend chaque élément du tableau et on le fait monter d'une place dans le tableau si celui-ci est plus grand que son successeur en échangeant leurs places. Sinon, on fait prendre le successeur et on opère à l'identique.

Algorithme 4 Tri bulle

```

1: Fonction BUBBLE_SORT(t)
2:    $n \leftarrow \text{taille}(t)$ 
3:   pour  $i$  de 0 à  $n-2$  répéter
4:     pour  $j$  de 0 à  $n-i-2$  répéter
5:       si  $t[j] > t[j+1]$  alors
6:         échange( $t[j]$ ,  $t[j+1]$ )
```

▷ faire monter la bulle

C Recherche d'un élément dans un tableau

On considère une liste L contenant des éléments de type `int`. Cette liste est **triée** par ordre croissant de ses éléments. On veut savoir si un élément x est présent dans L et comparer les performances des approches séquentielles et la dichotomiques. On dispose des algorithmes 5, 6 et 7.

Algorithme 5 Recherche séquentielle d'un élément dans un tableau

```

1: Fonction RECHERCHE_SÉQUENTIELLE(t, elem)
2:   n ← taille(t)
3:   pour i de 0 à n - 1 répéter
4:     si t[i] = elem alors
5:       renvoyer i                                ▷ élément trouvé, on renvoie sa position dans t
6:   renvoyer l'élément n'a pas été trouvé

```

Algorithme 6 Recherche d'un élément par dichotomie dans un tableau trié

```

1: Fonction RECHERCHE_DICHOTOMIQUE(t, elem)
2:   n ← taille(t)
3:   g ← 0
4:   d ← n-1
5:   tant que g ≤ d répéter                                ▷ ≤ cas où valeur au début, au milieu ou à la fin
6:     m ← (g+d)//2                                          ▷ Division entière : un indice est un entier!
7:     si t[m] < elem alors
8:       g ← m + 1                                          ▷ l'élément devrait se trouver dans t[m+1, d]
9:     sinon si t[m] > elem alors
10:      d ← m - 1                                          ▷ l'élément devrait se trouver dans t[g, m-1]
11:    sinon
12:      renvoyer m                                          ▷ l'élément a été trouvé
13:  renvoyer l'élément n'a pas été trouvé

```

- C1. Coder l'algorithme de recherche séquentielle d'un élément dans un tableau. Lorsque l'élément n'est pas présent dans le tableau, la fonction Python renvoie `None`. Sinon, elle renvoie l'indice de l'élément trouvé dans le tableau. Vérifier que cet algorithme fonctionne sur un tableau d'entiers de 20 éléments rempli aléatoirement. Dans le pire des cas, quel est le coût d'une recherche séquentielle en fonction de la taille du tableau?
- C2. Coder l'algorithme 6. Lorsque l'élément n'est pas présent dans le tableau, la fonction Python renvoie `None`. Sinon, elle renvoie l'indice de l'élément trouvé dans le tableau. Vérifier que cet algorithme fonctionne sur un tableau d'entiers de 20 éléments rempli aléatoirement et trié.
- C3. Coder l'algorithme 7. Vérifier que cet algorithme fonctionne sur un tableau d'entiers de 20 éléments rempli aléatoirement et trié et que l'indice renvoyé est bien l'indice minimal de la première occurrence de l'élément recherché.
- C4. On suppose que la longueur de la liste est une puissance de 2, c'est à dire $n = 2^p$ avec $p \geq 1$. Combien d'étapes l'algorithme 7 comporte-t-il? En déduire le nombre de comparaisons effectuées, dans le cas où l'élément est absent, en fonction de p puis de n , et comparer avec l'algorithme de recherche séquentielle.

Algorithme 7 Recherche d'un élément par dichotomie dans un tableau trié, renvoyer l'indice minimal en cas d'occurrences multiples.

```

1: Fonction RECHERCHE_DICHOTOMIQUE_INDICE_MIN(t, elem)
2:   n ← taille(t)
3:   g ← 0
4:   d ← n-1
5:   tant que g < d répéter                                ▷ attention au strictement inférieur!
6:     m ← (g+d)//2                                          ▷ Un indice de tableau est un entier!
7:     si t[m] < elem alors
8:       g ← m + 1                                          ▷ l'élément devrait se trouver dans t[m+1, d]
9:     sinon
10:      d ← m                                              ▷ l'élément devrait se trouver dans t[g, m]
11:   si t[g] = elem alors
12:     renvoyer g
13:   sinon
14:     renvoyer l'élément n'a pas été trouvé

```

Solution : Supposons que la taille de la liste soit une puissance de 2 : $n = 2^p$. Soit k, le nombre de tours de boucle nécessaires pour terminer l'algorithme. À la fin de l'algorithme, le tableau ne contient qu'un seul élément. Comme on divise par deux la taille du tableau à chaque tour de boucle, à la fin de l'algorithme on a nécessairement :

$$1 = \frac{n}{2^k} = \frac{2^p}{2^k}$$

On en déduit que $k = \log_2 n$. On dit que cet algorithme est de complexité logarithmique en $O(\log n)$ (cf. cours du deuxième semestre)

- C5. Tracer le graphique des temps d'exécution des algorithmes précédent en fonction de n. Les tracés sont-ils cohérents avec les calculs des coûts effectués précédemment?
- C6. La recherche dichotomique fonctionne-t-elle sur les listes non triées? Donner un contre-exemple si ce n'est pas le cas.

Solution :

Code 3 – Recherche un élément dans un tableau

```

1 import time
2 from random import randint
3
4
5 def seq_search(t, elem):
6     for i in range(len(t)):
7         if t[i] == elem:
8             return i
9     return None

```

```
10
11
12 def dichotomic_search(t, elem):
13     g = 0
14     d = len(t) - 1
15     while g <= d:
16         m = (d + g) // 2
17         # print(g, m, d)
18         if t[m] < elem:
19             g = m + 1
20         elif t[m] > elem:
21             d = m - 1
22         else:
23             return m
24     return None
25
26
27 def dichotomic_search_left_most(t, elem):
28     g = 0
29     d = len(t) - 1
30     while g < d:
31         m = (d + g) // 2
32         # print(g, m, d)
33         if t[m] < elem:
34             g = m + 1
35         else:
36             d = m
37     if t[g] == elem:
38         return g
39     else:
40         return None
41
42
43 def search_timing():
44     sizes = [i for i in range(10, 100000, 500)]
45     M = 5 * max(sizes)
46     results = []
47     for i in range(len(sizes)):
48         t = sorted([randint(0, M) for _ in range(sizes[i])])
49         mem_t = t[:]
50         results.append([])
51         for method in [seq_search, dichotomic_search,
52                        dichotomic_search_left_most]:
53             t = mem_t[:]
54             tic = time.perf_counter()
55             method(t, M // 4)
56             toc = time.perf_counter()
57             results[i].append(toc - tic)
58         print(f"# {i} - {sizes[i]} -> {results[i]}")
59
60     seq = [results[i][0] for i in range(len(sizes))]
61     dics = [results[i][1] for i in range(len(sizes))]
62     dicslm = [results[i][2] for i in range(len(sizes))]
63
64     from matplotlib import pyplot as plt
```

```

64
65 plt.figure()
66 plt.plot(sizes, seq, color='cyan', label='Sequential search')
67 plt.plot(sizes, dics, color='blue', label='Dichotomic search')
68 plt.plot(sizes, dicslm, color='black', label='Dichotomic search left most')
69 plt.xlabel('n', fontsize=18)
70 plt.ylabel('time', fontsize=16)
71 plt.legend()
72 plt.show()
73
74 # MAIN PROGRAM
75 t = [0, 1, 2, 4, 7, 7, 9, 13, 17, 65, 99, 99, 99, 99, 101, 111, 111, 111, 113]
76
77 print(t)
78 for value in t:
79     print(value, seq_search(t, value), dichotomic_search(t, value),
80           dichotomic_search_left_most(t, value),
81           dichotomic_search(t, value) == dichotomic_search_left_most(t, value))
82 search_timing()

```

