Лабораторна робота № 5.
Алгоритми пошуку підрядків в рядках.

Мета роботи: Ознайомитись з різними методами пошуку підрядків в рядках.

Задача роботи: Навчитись реалізовувати алгоритми пошуку Рабіна-Карпа (РК) та Кнута-Морриса-Пратта (КМП).

Теоретичні відомості.

Алгоритм Рабіна-Карпа.

Ідея, запропонована Рабіном і Карпом, полягає в тому, щоб поставити у відповідність кожному рядку деяке унікальне число, і замість того, щоб порівнювати самі рядки, порівнювати числа, що набагато швидше.

Задано: текст *T=abcdeabfgfcakmbddaf;*

підрядок, який треба знайти *P=bfgfc (m*=5);

алфавіт {0,1,2,3,4,5,6,7,8,9}, *d*=10, $h=d^{m-1}=10^4=10000$;

просте число *q*=13.

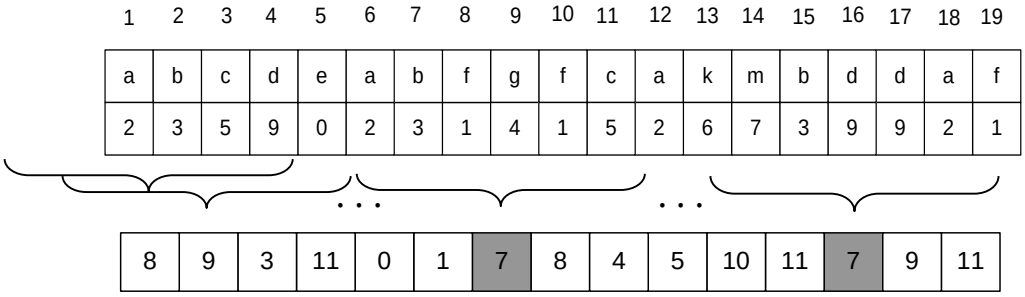
b	f	g	f	c
3	1	4	1	5

P=31415.

P_q =(5+10(1+10(4+10(1+10·3)))) mod 13=31415 mod 13=7

Необхідно шукати підрядки тексту T, які дорівнюють 7 mod 13.

Обчислимо значення T_S для всього тексту:



Відповідь: Зразок P входить в текст T зі зсувом S=6, позиція *i*=7.

Алгоритм Кнута-Морриса Пратта (КМП).

Алгоритм КМП складається з двох етапів: підготовчого (побудова префікс-функції) і основного (пошуку).

Задано: текст T= *abababaababababcbabababababcabb* (length(T)=27).
підрядок, який треба знайти P=*abababababca* (length(P)=10).

Підготовчий етап (*побудова префікс-функції*).

Для слова P розглянемо всі його префікси, що одночасно є суфіксами і серед них оберемо найдовший (не враховуючи самого P). Його будемо позначати $n(P)$ і називати найбільшим префікс-суфіксом.

Знайти повну префіксну функцію f для зразка P=*abababababca.*

$n(a)=L$, $f(1)=0$
$n(ab)=L$, $f(2)=0$
$n(aba)=a$, $f(3)=1$
$n(abab)=ab$, $f(4)=2$
$n(ababa)=aba$, $f(5)=3$
$n(abababab)=ababab$, $f(6)=4$
$n(abababaababca)=ababa$, $f(7)=5$
$n(abababababc)=ababab$, $f(8)=6$
$n(ababababc)=L$, $f(9)=0$
$n(abababababca)=a$, $f(10)=1$

або

i	1	2	3	4	5	6	7	8	9	10
$P[i]$	a	b	a	b	a	b	a	b	c	a
$f[i]$	0	0	1	2	3	4	5	6	0	1

Основний етап (*пошук*).

$f[i]$=[0,0,1,2,3,4,5,6,0,1].

1) Зсув S_0=0, q=5, $f[q]=f[5]$=3,
$S_1=S_0+(q-f(q))$=0+(5-3)=2, i=3.
2) Зсув S_1=2, q=3, $f[q]=f[3]$=1,
$S_2=S_1+(q-f(q))$=2+(3-1)=4, i=5.
3) Зсув S_2=4, q=1, $f[q]=f[1]$=0,
$S_3=S_2+(q-f(q))$=4+(1-0)=5, i=6.
4) Зсув S_3=5, q=9, $f[q]=f[9]$=0,
$S_4=S_3+(q-f(q))$=5+(9-0)=14, i=15.
5) Зсув S_4=14, q=0,
$S_5=S_4+1$=15, i=16.
6) Зсув S_5=15, q=10, $f[q]=f[10]$=1,
$S_6=S_5+(q-f(q))$=15+(10-1)=24, i=25. (Пошук зупиняємо, так як закінчився текст T).

T	*abababaababababcbabababababcabb*
P	*ababa*babca *aba*bababca *a*bababbca *abababab*ca abababbca *abababab*ca

Відповідь: Зразок P входить в текст T зі зсувом S=15, позиція i=16.

Завдання.

1. Задано: текст *T*, підрядок *P*, який треба знайти, просте число *q* (табл. 1), алфавіт {0,1,2,3,4,5,6,7,8,9}, *d*=10, $h = d^{m-1} = 10^9 = 1000000000$.
2. Знайти всі входження зразка P в текст T, використовуючи алгоритм РК.
3. Знайти всі входження зразка P в текст T, використовуючи алгоритм КМП.
4. Скласти алгоритм та написати програму реалізації алгоритмів РК та КМП за варіантами (парні варіанти за списком - РК, непарні - КМП).
5. Провести порівняльний аналіз алгоритмів пошуку в рядках.

Таблиця 1

Варіант	Текст T	Зразок P	Просте число *q*
1.	T = afdfabcabcabcfdca	P = abcabcabcf	13
2.	T = cbafcbcbacbcbcdcak	P = cbcbacbcbc	11
3.	T = hdacbaccbacabadhac	P = baccbacaba	17
4.	T = fdfacbacbacbadafc	P = cbacbacba	19
5.	T = ftfabababadbadfdkt	P = bababadbad	23
6.	T = bnbfdbdadbdaddnfdn	P = dbdadbdadd	13
7.	T = batfkatktkatkkfabk	P = katktkatkk	11
8.	T = baxazzazzxzzazxaxb	P = zzazzxzzaz	17
9.	T = npxamnmpmnmpmnxapn	P = mnmpmnmpmn	19
10.	T = nbnamanamnmanabanb	P = manamnmana	23
11.	T = batfxyxzxyxzxzfabki	P = xyxzxyxzxz	13
12.	T = tpotpptpopptpoapkp	P = pptpopptpo	11
13.	T = ankakbbykbnnakapkp	P = kbbykbbybk	17
14.	T = kerpedecedeceerkep	P = edecedecee	19
15.	T = kacvcmppmcmppcvcvk	P = cmppmcmppc	23
16.	T = ytrtqqptqqptqtytyq	P = qqptqqptqt	13
17.	T = ssfaffcfcffcfqasfs	P = ffcfcffcfq	11
18.	T = scbabkfakfakakfbass	P = kfakfakakf	17
19.	T = nbfbegfgegfefenbnf	P = egfgegfefe	19
20.	T = bgagcddcddcecdegba	P = cddcddcecd	23
21.	T = cvcxmcmcpmcmcpcxcz	P = mcmcpmcmcp	13
22.	T = bgnbxzyxyxzyxzbncz	P = xzyxyxzyxz	11
23.	T = nsabaappkaappkbnas	P = aappkaappk	17

24.	T = fdghmamapmamaahfgf	P = mamapmamaa	19
25.	T = fcbaslmlnlmlnlmgbac	P = lmlnlmlnlm	23
26.	T = nvbvmpkpmpkpmknbvc	P = mpkpmpkpmk	13
27.	T = tbtdcffcffcfpctbtd	P = cffcffcfpc	11
28.	T = scaskfcckfccfkbasb	P = kfcckfccfk	17
29.	T = bnfssrrtrsrrtrnbdf	P = srrtrsrrtr	19
30.	T = baacfkpfkpfkkfbdac	P = fkpfkpfkkf	23

Контрольні запитання.

1. Дайте визначення: рядок, підрядок, префікс слова, суфікс слова.
2. Алгоритм послідовного пошуку: переваги та недоліки.
3. Ідея алгоритму Рабіна-Карпа.
4. Ідея алгоритму КМП.
5. Префікс-функція. Принцип побудови префікс-функції.
6. Етапи алгоритму КМП. Охарактеризувати кожний етап.

```
Procedure RK (T,P,d,q)
Var …
    Begin  n:=length (T);      //довжина тексту
           m:= length (P);     //довжина зразка
           P:=0;               //числовий покажчик зразка P
           Ts :=0;         //змінний числовий покажчик підрядка тексту T
           h:=1:
    for i:=1 to m-1 do h:=d*h mod q;      //h=d^(m-1) mod q
    for i:=1 to m do
           begin
                  P:=(d*P+ord(P[i])) mod q;
                  Ts:=(d*Ts +ord(T[i])) mod q;
           end;
    for s:=0 to n-m do
      begin
                  if P=Ts then         //якщо співпадання за модулем
           begin
                  k:=1;                //перевірка тексту зі зразком
                  while (k≤m) and (P[k]=T[s+k]) do inc(k);
                  if k>m then          //якщо співпали P і Ts, то виводимо зсув
                  writeln (s);
           end;
           //обчислюємо наступне Ts
           Ts:=(d*(Ts– ord T[s+1]*h)+ ord T[s+m+1]) mod q;
      end
           End.
```