

The Rabin-Karp String Matching Algorithm

- Basic idea is to use **hash function and compute hash values for pattern and each shift in text.**
- The Rabin-Karp string matching algorithm calculates a hash value for the pattern, as well as for each M-character substring of text to be compared.
- If the hash values are unequal, the algorithm will determine the hash value for next M-character substring sequence.
- If the hash values are equal, the algorithm will analyse the pattern and the M-character sequence (i.e) compare characters in pattern and M-character substring of text.
- Therefore, **character matching is only required when the hash values match.**
- **Valid match (Valid Shift)** – hash values of pattern and M-character substring (ts) of text are equal and $T[s+1.....s+m] = P[1.....m]$.
- **Spurious hit** – hash values of pattern and M-character substring of text are equal but $T[s+1.....s+m] \neq P[1.....m]$.
- **No. of substrings/shifts (ts) = $n-m+1$ (i.e.) from $s = 0$ to $n-m$ (i.e.) t_0 to t_{n-m}**
- To compute hash value of pattern and M-character substring of text, **we use hash function.**
- **Hash function takes numerical value as input and gives numerical value as output. Output from hash function is considered to be the hash value.**
- **So, characters in text and pattern need to be represented as numbers.** Either we use ASCII values (Example: $a = 97$) or simple assignment like $a=0$, $b=1$, $c=2$ and so on.
- **Rolling hash function is used.** A rolling hash is a hash function that is used to **efficiently** compute a hash value for a sliding window of data.

- The idea behind a rolling hash is to compute the hash value for a fixed-size window of data, and then “roll” the window one position at a time and re-compute the hash value.
- **Sliding window size = length of pattern**
- By using a sliding window, the rolling hash can avoid re-computing the hash value from scratch for each position, which can be time-consuming for large data sets.
- Instead, the rolling hash only needs to update the hash value for the new data.

Examples to understand rolling hash function and sliding window:

Example:1

eg-1

Text : ^{1 2 3 4 5 6} a a a a a b

Pattern : ^{1 2 3} a a b

$n=6$ $m=3$ no. of shifts/comparisons = $6-3+1=4$

Pattern : a a b

$0+0+1=1 \Rightarrow$ hash value. Substring in Text.

Text : a a a a a b

(Shift 0) to : $0+0+0 \rightarrow 0$

(Shift 1) t1 : $0+0+0 \rightarrow 0$

(Shift 2) t2 : $0+0+0 \rightarrow 0$

(Shift 3) t3 : $0+0+1 \rightarrow 1$ Possibility of match. so compare

$T[s+1 \dots s+m] = P[1 \dots m]$

Example:2

No. of shifts/comparisons = $n-m+1 = 8-3+1 = 6$
(i.e) t_0 to t_5

eg-2

Text : a b c d a b c e n=8
m=3

pattern : b c e

$1+2+4 = 7 \Rightarrow$ Hash value of pattern

Text : a b c d a b c e hash value

(Shift 0) $t_0 : \rightarrow 0+1+2 \dots \rightarrow 3$

(Shift 1) $t_1 : \rightarrow 1+2+3 \dots \rightarrow 6$

(Shift 2) $t_2 : \rightarrow 2+3+0 \dots \rightarrow 5$

(Shift 3) $t_3 : \rightarrow 3+0+1 \dots \rightarrow 4$

(Shift 4) $t_4 : \rightarrow 0+1+2 \dots \rightarrow 3$

(Shift 5) $t_5 : \rightarrow 1+2+4 \rightarrow 7$

(Shift 5) $t_5 : \rightarrow 1+2+4 \rightarrow 7$

possibility of match.
So compare characters.
 $T[s+1 \dots s+m] = P[1 \dots m]$

Example to understand the terms spurious hit and valid match:

Text : c c a c c a a e d b a $n=11$ $n-m+1=11-3+1=9$
 pattern : d b a $m=3$ to t_0 to t_8
 $3+1+0 = 4 \Rightarrow \text{hash value.}$

Text :	c	c	a	c	c	a	a	e	d	b	a			
t_0		2	+	2	+	0						$\rightarrow 4$		
t_1			2	+	0	+	2					$\rightarrow 4$		
t_2				0	+	2	+	2				$\rightarrow 4$		
t_3					2	+	2	+	0			$\rightarrow 4$		
t_4						2	+	0	+	0		$\rightarrow 2$		
t_5							0	+	0	+	4	$\rightarrow 4 \rightarrow \text{spurious hit}$		
t_6								0	+	4	+	3		
t_7									4	+	3	+	1	
t_8										3	+	1	+	0

$\rightarrow 4 \rightarrow \text{Valid match (or) valid shift. (S=8)}$

- Spurious hit is because of simple hash function.
- Strong hash function is needed to minimize spurious hits.
- Rabin Karp algorithm suggest strong hash function.

String matching using Rabin Karp algorithm

Example:1

Let us consider same text and pattern that has been considered above to understand spurious hit and valid match.

Pattern : $\begin{matrix} 1 & 2 & 3 \\ d & b & a \end{matrix}$
 $3 \times 10^2 + 1 \times 10^1 + 0 \times 10^0 \rightarrow$ base value. Here base = 10. So only 10 alphabets can be considered. (i.e) a to j preferably prime no.

$$\therefore [P[1] \times 10^{m-1} + P[2] \times 10^{m-2} + P[3] \times 10^{m-3}] \bmod q$$

If 26 alphabets are considered then base will be 26.
 If all lower case, upper case and special characters are considered, then base = 127.

Pattern(P) : $\begin{matrix} 1 & 2 & 3 \\ d & b & a \end{matrix}$
 $3 \times 10^2 + 1 \times 10^1 + 0 \times 10^0 = 310 \bmod 13 \rightarrow$ preferably prime no.
 $= 11 \Rightarrow$ high value of pattern.

Likewise, hash value of each m-character substring is computed.

Text : $\begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ c & c & a & c & c & a & a & e & d & b & a \\ 2 & 2 & 0 & 2 & 2 & 0 & 0 & 4 & 3 & 1 & 0 \end{matrix}$

$t_0 : c c a$
 $2 \times 10^2 + 2 \times 10^1 + 0 \times 10^0 = 220 \bmod 13 = 12 \times$

$t_1 : [2 \times 10^2 + 2 \times 10^1 + 0 \times 10^0] - 2 \times 10^2 \times 10 + 2 \times 10^0$
 $[cac] : 2 \times 10^2 + 0 \times 10^1 + 2 \times 10^0 = 202 \bmod 13 = 7 \times$

$t_2 : [2 \times 10^2 + 0 \times 10^1 + 2 \times 10^0] - 2 \times 10^2 \times 10 + 2 \times 10^0$
 $[acc] : 0 \times 10^2 + 2 \times 10^1 + 2 \times 10^0 = 22 \bmod 13 = 9 \times$

$t_3 = cca$
 $t_4 = caa$
 $t_5 = aae$
 $t_6 = aed$
 $t_7 = edb$
 $t_8 = dba$

$$t_3 : \left[\left[\cancel{0 \times 10^2} + 2 \times 10^1 + 2 \times 10^0 \right] - \cancel{0 \times 10^2} \right] \times 10 + 0 \times 10^0$$

cca $2 \times 10^2 + 2 \times 10^1 + 0 \times 10^0 = 220 \bmod 13 = \boxed{12} \times$

$$t_4 : \left[\left[\cancel{2 \times 10^2} + 2 \times 10^1 + 0 \times 10^0 \right] - \cancel{2 \times 10^2} \right] \times 10 + 0 \times 10^0$$

caa $2 \times 10^2 + 0 \times 10^1 + 0 \times 10^0 = 200 \bmod 13 = \boxed{5} \times$

$$t_5 : \left[\left[\cancel{2 \times 10^2} + 0 \times 10^1 + 0 \times 10^0 \right] - \cancel{2 \times 10^2} \right] \times 10 + 4 \times 10^0$$

aae $0 \times 10^2 + 0 \times 10^1 + 4 \times 10^0 = 4 \bmod 13 = \boxed{4} \times$

$$t_6 : \left[\left[\cancel{0 \times 10^2} + 0 \times 10^1 + 4 \times 10^0 \right] - \cancel{0 \times 10^2} \right] \times 10 + 3 \times 10^0$$

aed $0 \times 10^2 + 4 \times 10^1 + 3 \times 10^0 = 43 \bmod 13 = \boxed{4} \times$

$$t_7 : \left[\left[\cancel{0 \times 10^2} + 4 \times 10^1 + 3 \times 10^0 \right] - \cancel{0 \times 10^2} \right] \times 10 + 1 \times 10^0$$

edb $4 \times 10^2 + 3 \times 10^1 + 1 \times 10^0 = 431 \bmod 13 = \boxed{2} \times$

$$t_8 : \left[\left[\cancel{4 \times 10^2} + 3 \times 10^1 + 1 \times 10^0 \right] - \cancel{4 \times 10^2} \right] \times 10 + 0 \times 10^0$$

dba $3 \times 10^2 + 1 \times 10^1 + 0 \times 10^0 = 310 \bmod 13 = \boxed{11} \checkmark$

Here, comparison of $P[1..m]$ with

$T[st1 \dots stm]$ takes place.

It is found to have exact match.

$\therefore$ Valid shift $S = 8$ No spurious hits with strong hash fn.

Algorithm:

RABIN-KARP-MATCHER (T, P, d, q)

1. $n \leftarrow \text{length}[T]$
2. $m \leftarrow \text{length}[P]$
3. $h \leftarrow d^{m-1} \bmod q$ //q preferably prime number, d - base
4. $p \leftarrow 0$
5. $t_0 \leftarrow 0$
6. for $i \leftarrow 1$ to m //Pre-processing
7. do $p \leftarrow (dp + P[i]) \bmod q$
8. $t_0 \leftarrow (dt_0 + T[i]) \bmod q$
9. for $s \leftarrow 0$ to $n-m$ //Matching
10. do if $p = t_s$
11. then if $P[1 \dots m] = T[s+1 \dots s+m]$
12. then "Pattern occurs with shift" s
13. If $s < n-m$
14. then $t_{s+1} \leftarrow (d(t_s - T[s+1]h) + T[s+m+1]) \bmod q$ //formula to compute hash value of next sliding window.

- Given a pattern $P[1 \dots m]$, we let p denote its hash value of pattern.
- In a similar manner, given a text $T[1 \dots n]$, we let t_s denote the hash value of the length- m substring $T[s+1 \dots s+m]$, for $s = 0, 1, \dots, n-m$.
- If $t_s = p$ and if $T[s+1 \dots s+m] = P[1 \dots m]$; thus, s is a valid shift.
- If $t_s = p$ and if $T[s+1 \dots s+m] \neq P[1 \dots m]$; thus, s is a spurious hit.

Time Complexity:

- We can compute p in time $\theta(m)$ using **Horner's rule**
$$p = P[m] + 10(P[m-1] + 10(P[m-2] + \dots + 10(P[2] + 10P[1]) \dots))$$

- The value t_0 can be similarly computed from $T[1 \dots m]$ in time $\theta(m)$.
- If we could compute p in time $\theta(m)$ and all the t_s values in a total of $\theta(n - m + 1)$ time, then we could determine all valid shifts s in time $\theta(m) + \theta(n - m + 1) = \theta(n)$
- Therefore, Total running time of string-matching algorithm = pre-processing time + matching time

Example: 2 – With valid shift and spurious hit

Text T: 2 3 5 9 0 2 3 1 4 1 5 2 6 7 3 9 9 2 1
 Pattern P: 3 1 4 1 5

$n = 19$ $m = 5$ prime no. $q = 13$ base = 10

No. of substrings (or) shifts = $n - m + 1 = 19 - 5 + 1 = 15$
 (i.e) from t_0 to t_{14}

Hash value of pattern:

P: 3 1 4 1 5
 $3 \times 10^4 + 1 \times 10^3 + 4 \times 10^2 + 1 \times 10^1 + 5 \times 10^0$
 $= 31415 \text{ mod } 13 = \boxed{7}$

Hash value of substrings from t_0 to t_{14}

t_0 : $2 \times 10^4 + 3 \times 10^3 + 5 \times 10^2 + 9 \times 10^1 + 0 \times 10^0$
 $\boxed{23590} = 23590 \text{ mod } 13 = \boxed{8} \times$

t_1 : $\left[\left[\cancel{2 \times 10^4} + 3 \times 10^3 + 5 \times 10^2 + 9 \times 10^1 + 0 \times 10^0 \right] - \cancel{2 \times 10^4} \right] \times 10 + 2 \times 10^0$
 $\boxed{35902} = 35902 \text{ mod } 13 = \boxed{9} \times$

t_2 : $\left[\left[\left[\cancel{3 \times 10^4} + 5 \times 10^3 + 9 \times 10^2 + 0 \times 10^1 + 2 \times 10^0 \right] - \cancel{3 \times 10^4} \right] \times 10 + 3 \times 10^0 \right]$
 $\boxed{59023} = 59023 \text{ mod } 13 = \boxed{3} \times$

$$t_3: \boxed{90231} = \left[\left[\left[5 \times 10^4 + 9 \times 10^3 + 0 \times 10^2 + 2 \times 10^1 + 3 \times 10^0 \right] - 5 \times 10^4 \right] \times 10 \right] + 1 \times 10^0 \\ = 90231 \bmod 13 = \boxed{11} X$$

$$t_4: \boxed{02314} = \left[\left[\left[9 \times 10^4 + 0 \times 10^3 + 2 \times 10^2 + 3 \times 10^1 + 1 \times 10^0 \right] - 9 \times 10^4 \right] \times 10 \right] + 4 \times 10^0 \\ = 02314 \bmod 13 = \boxed{0} X$$

$$t_5: \boxed{23141} = \left[\left[\left[0 \times 10^4 + 2 \times 10^3 + 3 \times 10^2 + 1 \times 10^1 + 4 \times 10^0 \right] - 0 \times 10^4 \right] \times 10 \right] + 1 \times 10^0 \\ = 23141 \bmod 13 = \boxed{1} X$$

$$t_6: \boxed{31415} = \left[\left[\left[2 \times 10^4 + 3 \times 10^3 + 1 \times 10^2 + 4 \times 10^1 + 1 \times 10^0 \right] - 2 \times 10^4 \right] \times 10 \right] + 5 \times 10^0 \\ = 31415 \bmod 13 = \boxed{7} X$$

Here $s=6$ Hence, compare $T[s+1 \dots s+m]$ with $P[1 \dots m]$
 $\Rightarrow$ exact match.

$$t_7: \boxed{14152} = \left[\left[\left[3 \times 10^4 + 1 \times 10^3 + 4 \times 10^2 + 1 \times 10^1 + 5 \times 10^0 \right] - 3 \times 10^4 \right] \times 10 \right] + 2 \times 10^0 \\ = 14152 \bmod 13 = \boxed{8} X$$

$$t_8: \boxed{41526} = \left[\left[\left[1 \times 10^4 + 4 \times 10^3 + 1 \times 10^2 + 5 \times 10^1 + 2 \times 10^0 \right] - 1 \times 10^4 \right] \times 10 \right] + 6 \times 10^0 \\ = 41526 \bmod 13 = \boxed{4} X$$

$$t_9: \boxed{15267} \left[\left[\cancel{4 \times 10^4} + 1 \times 10^3 + 5 \times 10^2 + 2 \times 10^1 + 6 \times 10^0 \right] - \cancel{4 \times 10^4} \right] \times 10 + 7 \times 10^0 \\ = 15267 \bmod 13 = \boxed{5} \times$$

$$t_{10}: \boxed{52673} \left[\left[\cancel{1 \times 10^4} + 5 \times 10^3 + 2 \times 10^2 + 6 \times 10^1 + 7 \times 10^0 \right] - \cancel{1 \times 10^4} \right] \times 10 + 3 \times 10^0 \\ = 52673 \bmod 13 = \boxed{10} \times$$

$$t_{11}: \boxed{26739} \left[\left[\cancel{5 \times 10^4} + 2 \times 10^3 + 6 \times 10^2 + 7 \times 10^1 + 3 \times 10^0 \right] - \cancel{5 \times 10^4} \right] \times 10 + 9 \times 10^0 \\ = 26739 \bmod 13 = \boxed{11} \times$$

$$t_{12}: \boxed{67399} \left[\left[\cancel{2 \times 10^4} + 6 \times 10^3 + 7 \times 10^2 + 3 \times 10^1 + 9 \times 10^0 \right] - \cancel{2 \times 10^4} \right] \times 10 + 9 \times 10^0 \\ = 67399 \bmod 13 = \boxed{7} \checkmark$$

Here $s=12$. Compare $T[s+1 \dots s+m]$ with $P[1 \dots m]$

$$T[s+1 \dots s+m] \neq P[1 \dots m]$$

$\therefore$ Spurious Hit

$t_{13}:$

$$\boxed{73992} \left[\left[\cancel{6 \times 10^4} + 7 \times 10^3 + 3 \times 10^2 + 9 \times 10^1 + 9 \times 10^0 \right] - \cancel{6 \times 10^4} \right] \times 10 + 2 \times 10^0 \\ = 73992 \bmod 13 = \boxed{9} \times$$

$$t_{14}: \boxed{39921} \left[\left[\cancel{7 \times 10^4} + 3 \times 10^3 + 9 \times 10^2 + 9 \times 10^1 + 2 \times 10^0 \right] - \cancel{7 \times 10^4} \right] \times 10 + 1 \times 10^0 \\ = 39921 \bmod 13 = \boxed{11} \times$$

$\therefore$ Valid shift $s=6$.

Spurious hit $s=12$.

$\therefore$ pattern is found at shift $\boxed{s=6}$.