

KMP algorithm

문자열 매칭 알고리즘

어떤 문자열 S에서, 어떤 패턴 P를 찾아내는 알고리즘이다.
오토마타 알고리즘 라빈-카프 알고리즘 KMP알고리즘 등이 있다.

KMP 알고리즘이란?

KMP알고리즘은 만든 사람 이름이 KNUTH, MORRIS, PRETT이기 때문에 앞글자를 하나씩 따서 KMP알고리즘이라 이름이 붙었다.

KMP알고리즘의 시간 복잡도는 $O(N+M)$ 으로 하나 하나 비교하는 방법 $O(NM)$ 보다 매우 빠르다.

단순한 방식

인덱스	0	1	2	3	4	5	6	7	8	9	10	11
텍스트	A	B	C	D	A	B	C	D	A	B	E	E
패턴	A	B	C	D	A	B	E					



인덱스	0	1	2	3	4	5	6	7	8	9	10	11
텍스트	A	B	C	D	A	B	C	D	A	B	E	E
패턴		A	B	C	D	A	B	E				

인덱스	0	1	2	3	4	5	6	7	8	9	10	11
텍스트	A	B	C	D	A	B	C	D	A	B	E	E
패턴	A	B	C	D	A	B	E					

여기까지 맞다는 정보를 무시하고 다음 인덱스로 진행

이 정보를 활용한다면?

인덱스	0	1	2	3	4	5	6	7	8	9	10	11
텍스트	A	B	C	D	A	B	C	D	A	B	E	E
패턴		A	B	C	D	A	B	E				

인덱스	0	1	2	3	4	5	6	7	8	9	10	11
텍스트	A	B	C	D	A	B	C	D	A	B	E	E
패턴			A	B	C	D	A	B	E			

인덱스	0	1	2	3	4	5	6	7	8	9	10	11
텍스트	A	B	C	D	A	B	C	D	A	B	E	E
패턴				A	B	C	D	A	B	E		

인덱스	0	1	2	3	4	5	6	7	8	9	10	11
텍스트	A	B	C	D	A(i)	B	C	D	A	B	E	E
패턴					A(j)	B	C	D	A	B	E	

접두사와 접미사(공통 머리 공통 꼬리)를 잘 보면 공통 되는 부분부터 순회를 할 수 있다는 것을 알 수 있다.

인덱스	0	1	2	3	4	5	6	7	8	9	10	11
텍스트	A	B	C	D	A	B	C	D	A	B	E	E
패턴	A	B	C	D	A	B	E					

틀린 부분 바로 전 인덱스까지의 패턴부분
공통 접두/접미사를 안다면 시간 단축을 할 수 있을 것!

미리 패턴의 모든 인덱스마다 해당 공통 접두/접미사 일치 정보를 저장한다면 더 빠른 검색이 가능

해당 정보가 담긴 배열을 pi 배열 이라고함

pi 배열 구하는 알고리즘

Pi	인덱스	0	1	2	3	4	5	6	7									
	값	0	0															
패턴 P	인덱스	0	1 (i)	2	3	4	5	6	7									
	값	A	B	A	B	A	B	A	C									
패턴 P	인덱스		0 (j)	1	2	3	4	5	6	7								
	값		A	B	A	B	A	B	A	C								

Pi	인덱스	0	1	2	3	4	5	6	7									
	값	0	0	1														
패턴 P	인덱스	0	1	2 (i)	3	4	5	6	7									
	값	A	B	A	B	A	B	A	C									
패턴 P	인덱스			0 (j)	1	2	3	4	5	6	7							
	값			A	B	A	B	A	B	A	C							

pi 배열 구하는 알고리즘

Pi	인덱스	0	1	2	3	4	5	6	7								
	값	0	0	1	2												
패턴 P	인덱스	0	1	2	3 (i)	4	5	6	7								
	값	A	B	A	B	A	B	A	C								
패턴 P	인덱스			0	1 (j)	2	3	4	5	6	7						
	값			A	B	A	B	A	B	A	C						

pi 배열 구하는 알고리즘

Pi	인덱스	0	1	2	3	4	5	6	7									
	값	0	0	1	2	3												
패턴 P	인덱스	0	1	2	3	4 (i)	5	6	7									
	값	A	B	A	B	A	B	A	C									
패턴 P	인덱스			0	1	2 (j)	3	4	5	6	7							
	값			A	B	A	B	A	B	A	C							

Pi	인덱스	0	1	2	3	4	5	6	7									
	값	0	0	1	2	3	4											
패턴 P	인덱스	0	1	2	3	4	5 (i)	6	7									
	값	A	B	A	B	A	B	A	C									
패턴 P	인덱스			0	1	2	3 (j)	4	5	6	7							
	값			A	B	A	B	A	B	A	C							

pi 배열 구하는 알고리즘

Pi	인덱스	0	1	2	3	4	5	6	7									
	값	0	0	1	2	3	4	5										
패턴 S	인덱스	0	1	2	3	4	5	6 (i)	7									
	값	A	B	A	B	A	B	A	C									
패턴 P	인덱스			0	1	2	3	4 (j)	5	6	7							
	값			A	B	A	B	A	B	A	C							

Pi	인덱스	0	1	2	3	4	5	6	7									
	값	0	0	1	2	3	4	5										
패턴 P	인덱스	0	1	2	3	4	5	6	7 (i)									
	값	A	B	A	B	A	B	A	C									
패턴	인덱스			0	1	2	3	4	5 (j)	6	7							
	값			A	B	A	B	A	B	A	C							

pi 배열 구하는 알고리즘

Pi	인덱스	0	1	2	3	4	5	6	7									
	값	0	0	1	2	3	4	5										
패턴 P	인덱스	0	1	2	3	4	5	6	7 (i)									
	값	A	B	A	B	A	B	A	C									
패턴 P	인덱스					0	1	2	3 (j)	4	5	6	7					
	값					A	B	A	B	A	B	A	C					

Pi	인덱스	0	1	2	3	4	5	6	7									
	값	0	0	1	2	3	4	5										
패턴 S	인덱스	0	1	2	3	4	5	6	7 (i)									
	값	A	B	A	B	A	B	A	C									
패턴 P	인덱스							0	1 (j)	2	3	4	5	6	7			
	값							A	B	A	B	A	B	A	C			

pi 배열 구하는 알고리즘

Pi	인덱스	0	1	2	3	4	5	6	7									
	값	0	0	1	2	3	4	5	0									
패턴 P	인덱스	0	1	2	3	4	5	6	7 (i)									
	값	A	B	A	B	A	B	A	C									
패턴 P	인덱스								0 (j)	1	2	3	4	5	6	7		
	값								A	B	A	B	A	B	A	C		

pi 배열 구하는 알고리즘

Pi	인덱스	0	1	2	3	4	5	6	7									
	값	0	0	1	2	3	4	5	0									
패턴 P	인덱스	0	1	2	3	4	5	6	7	8 (i)								
	값	A	B	A	B	A	B	A	C									
패턴 P	인덱스									0 (j)	1	2	3	4	5	6	7	
	값									A	B	A	B	A	B	A	C	

pi 배열은 구했고 그 다음은?

실제 텍스트랑 패턴이랑 비교하면서 틀렸을 경우
해당 인덱스 파이 배열 정보만큼 건너 뛰기

파이배열 구하는 로직이랑 동일

pi 배열은 구했고 그 다음은?

인덱스	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
텍스트	A	B	A	B	A	B	A	B	B (i)	A	B	A	B	A	B	A	B	C
패턴	A	B	A	B	A	B	A	B	C (j)									

인덱스	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
텍스트	A	B	A	B	A	B	A	B	B (i)	A	B	A	B	A	B	A	B	C
패턴			A	B	A	B	A	B	A (j)	B	C							

pi 배열은 구했고 그 다음은?

인덱스	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
텍스트	A	B	A	B	A	B	A	B	B (i)	A	B	A	B	A	B	A	B	C
패턴					A	B	A	B	A (j)	B	A	B	C					

인덱스	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
텍스트	A	B	A	B	A	B	A	B	B (i)	A	B	A	B	A	B	A	B	C
패턴							A	B	A (j)	B	A	B	A	B	C			

pi 배열은 구했고 그 다음은?

인덱스	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
텍스트	A	B	A	B	A	B	A	B	B	A (i)	B	A	B	A	B	A	B	C
패턴										A (j)	B	A	B	A	B	A	B	C

실제 코드로 돌려보기!

END