

## *k*-mer slide run 코딩

**[문제]**  $k$ -mer는 일반적으로 문자열에서 길이가  $k$ 인 모든 부분문자열이다. 예를 들어 문자열이 agatcgagt에서 3-mers는 aga gat atc tcg cga gag agt가 된다. 3-mer slide 부분문자열은 중복부분을 제외한 aga tcg agt를 의미한다.  $k$ 는 부분문자열의 길이를 뜻하고 이 예에서는 3이다. 문서의 압축을 위해 많은 방법이 있지만 아주 간단한 run length 코딩을 응용하여 주어진 문자열을 압축하고자 한다. 이 방법은 문자열에서 같은  $k$ -mer가 연속해서 나타나는 것을 그 부분문자열의 개수와 반복되는  $k$ -mer값으로 표현하여 더 짧은 문자열로 줄여서 표현하는 알고리즘이다.

간단한 예로 "aaggattt"의 경우 1-mer slide run 코딩은 run length 코딩과 같은 기능을 하게 된다. "2a2ga3t"(문자가 반복되지 않아 한번만 나타난 경우 1은 생략함)와 같이 표현할 수 있는데, 이러한 방식은 반복되는 문자가 적은 경우 압축률이 낮다는 단점이 있다. "agttagtacac"와 같은 문자열은 전혀 압축되지 않는다. 이런 경우에는 2-mer slide run 코딩을 사용하면 "agttagt2ac"로 9바이트로 표현되고 3-mer slide run 코딩을 적용하면 "2agtacac"로 8바이트로 표현할 수 있다. 반복되는 부분문자열이 없는 경우에는 그대로 출력한다. 또 다른 예를 보면, 16바이트인 "atatcgatcgatcgatcgatcg"의 경우 문자를 1-mer slide run 코딩을 하면 압축되지 않지만, 2-mer slide run 코딩을 하면 "2at2cg2at2cg"로 12바이트로 표현할 수 있다. 다른 방법으로 8-mer slide run 코딩을 하면 "2atatcgat"로 9바이트로 표현할 수 있으며, 이때가 가장 짧게 표현할 수 있다.

다른 예로, "agttagtcg"와 같은 경우, 2-mer slide run 코딩은 "agttagt2cg"가 되지만, 3-mer slide run 코딩은 "2abcdede"가 되어 이 경우가 가장 짧은 코딩이 된다. 이때 3개 단위로 자르고 마지막에 남는 문자열은 그대로 붙여준다.

[입출력] 입력은 첫줄에 처리해야 할 문자열 개수  $n$ 이 주어지고 2번째 줄부터  $n$ 개의 문자열  $s$  ( $1 \leq s \leq 1000$ )가 주어진다. 출력은  $s$ 가 주어질 때, 각각의 문자열에 대해  $k$ -mer slide run 코딩할 때 가장 짧은 길이로 표현되는 코딩길이를 출력한다. 단 문자열은 대소문자를 구분한다.

### [예제]

입력 stdin	출력 stdout
5	7
aabbaccc	9
ababcdcdababcdcd	8

abcabcdede abcabcabcabcdededededede xababcdcdababcdcd	14 17
---	----------

[제한조건] 프로그램의 이름은 pa04\_runcode.{py,c,cpp,java}이다. 제출 횟수는 최대 15번이며 허용 시간은 데이터 당 제한 시간은 1초, 허용가능 코드의 최대 크기는 3,000 bytes이다. 문제 풀이 마감시간은 2022년 5월5일 24:00이다. 제출한 프로그램에 대한 풀이(방법과 코드설명)를 작성하여 2022년 5월6일 24:00까지 NESPA “설명게시판”에 제출해야 한다. 제출한 프로그램 풀이과정은 마감이 지나면 공개된다.