String

- A string is a sequence of elements drawn from a finite set S

eg Consider the finite set $S = \{0, 1\}$, then examples of the strings over the set S are 01, 001, etc

- A string of length k is called k-string

eg There are 8 strings of length 3 over the set $S = \{0, 1\}$.

000, 001, 010, 011, 100, 101, 110, 111.

- A substring s' of a string s is an ordered sequence of consecutive elements of s

eg 01101001

010 ✓ (substring of length 3, and begins at position 4)

111 × (not a substring)

Total no of substrings

Consider a string of length n, then

no. of substrings of length 1 : n

—— " —— 2 : $n-1$

—— " —— 3 : $n-2$

⋮

no. of substrings of length k : $n-(k-1)$

⋮

no of substring of length n : $n-(n-1) = 1$

Total no. of substrings of all lengths (1 to n) $= n+(n-1)+(n-2)+\cdots 2+1$

$$= \frac{n(n+1)}{2}$$

★ These are the total number of non-empty substrings of a string with 'n' characters. But, it includes the string itself.

subsequence

- A subsequence of a given sequence is just the given sequence with zero or more elements left out.

- Formally, Given a sequence $X = \langle x_1, x_2, \cdots x_m \rangle$, another sequence $Z = \langle z_1, z_2 \cdots z_k \rangle$ is called a subsequence of X if there exists a strictly increasing sequence $\langle i_1, i_2, \cdots i_k \rangle$ of indices of X such that

$\forall\ j = 1, 2, \cdots k$ ($\Rightarrow$ for each element of Z)

$$\boxed{z_j = x_{i_j}}$$

eg $X = \langle A, B, C, B, D, A, B \rangle$

then $Z = \langle B, C, D, B \rangle$ is a subsequence of X

with corresponding index sequence $\langle 2, 3, 5, 7 \rangle$

total no. of subsequences

Consider a sequence X of length n, then total no. of subsequences will be 2^n.

because each element of X may or may be chosen in a subsequence, therefore there are 2 choices for each element.

$$\underline{2} \times \underline{2} \ \underline{\quad} \ \ldots\ldots \times \underline{2}$$

$$\Rightarrow 2^n$$

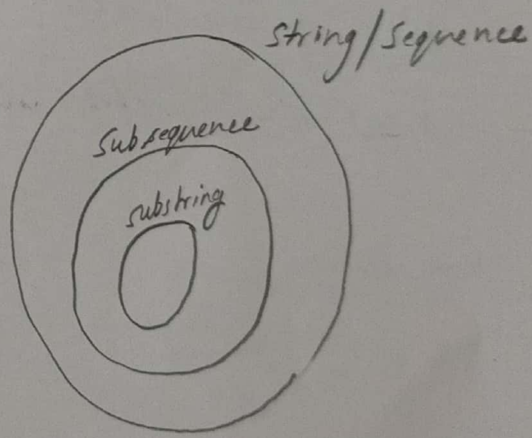
* This number includes subsequence with zero length, as well as the sequence itself.

subsequence Vs substring

eg Let $X = \langle NITKKR \rangle$

then $\langle IKR \rangle$ is a subsequence but not a substring.

Further, any substring is also a subsequence, therefore we can say that concept of "subsequence" is a generalization of the concept of "substring".



Prefix

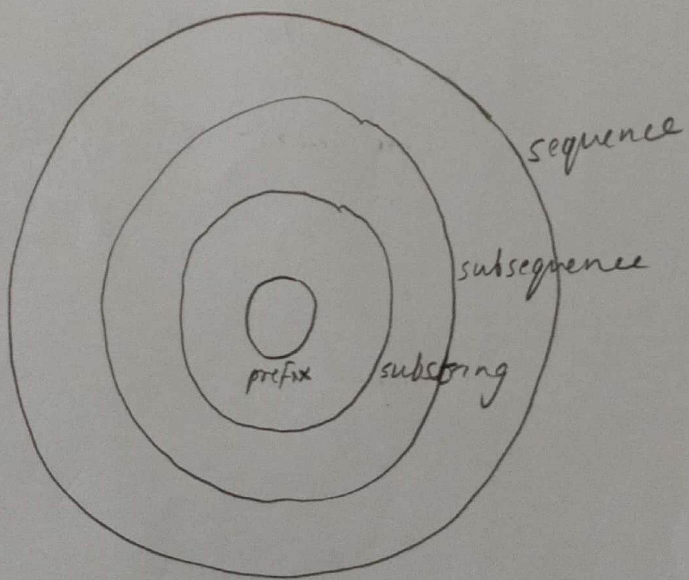
- Given a sequence $X = \langle x_1, x_2, \dots x_m \rangle$ then we define i^{th} prefix of X as follows:

$$X_i = \langle x_1, x_2 \dots x_i \rangle \; ; \text{for } i = 0, 1, 2 \dots m.$$

- eg if $X = \langle N, I, T, K, K, R \rangle$ then

$$X_3 = \langle N, I, T \rangle$$

- it should be noted that X_0 is the empty sequence and X_m is the sequence itself.
- Total no of prefixes of a sequence of length n will be $(n+1)$



Common Subsequence

Given two sequences X and Y, then a sequence Z is called a "common subsequence" of X and Y if Z is a subsequence of both X & Y.

eg

$X = \langle A, B, C, B, D, A, B \rangle$

$Y = \langle B, D, C, A, B, A \rangle$ then the sequence $\langle B, C, A \rangle$ is a common subsequence of both X & Y. but it is not a longest common subsequence (LCS) of X & Y.

the sequence $\langle B, C, B, A \rangle$ is also common subsequence of both X & Y and has length 4.

Since, there is no common subsequence of length 5 or more, therefore the sequence $\langle B, C, B, A \rangle$ is an LCS of X & Y.

☆ LCS is not unique.

LCS Problem

We are given two sequences

$$X = \langle x_1, x_2, \cdots x_m \rangle \quad \text{and}$$

$$Y = \langle y_1, y_2 \cdots y_n \rangle$$

we have to find a maximum length common subsequence of X and Y.

Theorem Let $X = \langle x_1, x_2 \cdots x_m \rangle$ and $Y = \langle y_1, y_2 \cdots y_n \rangle$ be sequences, and let $Z = \langle z_1, z_2 \cdots z_k \rangle$ be any LCS of X and Y.

(1) if $x_m = y_n$ then $z_k = x_m = y_n$ and Z_{k-1} is an LCS of X_{m-1} and Y_{n-1}

(2) if $x_m \neq y_n$, then $z_k \neq x_m$ implies that Z is an LCS of X_{m-1} and Y.

(3) if $x_m \neq y_n$, then $z_k \neq y_n$ implies that Z is an LCS of X and Y_{n-1}.