SOEN 6461 Fall 2015 Design Assignment 1

Design a program to gather statistics of all strings P of length k in a string S of length n.

Greg Butler

Computer Science and Software Engineering
Concordia University, Montreal, Canada

Email: gregb@cs.concordia.ca

Design Assignment 1 — Count Substrings

Small Example

Alphabet $= \{$ a, c, g, t $\}$ String $S=$ ataaaa size $n = 6$

Substrings of size $k = 1$

Substrings

a

t

Counts

a : 5

t : 1

Total $= n$

Positions

a : 5 : 0,2,3,4,5

t : 1 : 1

Substrings of size $k = 2$

Substrings

aa

at

ta

Counts

aa : 3

at : 1

ta : 1

Total $= n - 1$

Positions

aa : 3 : 2,3,4

at : 1 : 0

ta : 1 : 1

Compute substrings, counts, and positions!

Naive Solution: Design Assignment 1 — Count Substrings

```
//Construct collection C of
// triples <pattern, cnt, position>

for each pattern p in alphabet^k do
   C[ p ] := < p, 0, empty_list >;
end for

for i := 0 to n-k do
   ss = S.substring(i, i+k-1));
   C[ ss ].count++;
   C[ ss ].list.append(i);
end for
```

Count operations
& data movement

$|alphabet|^k$ iterations

$n - k$ iterations
data movement?
indexing cost?
indexing cost?

Design Assignment 1 — Count Substrings

Issues

Correctness

system must compute the right answers!

Efficiency = Resource Usage

Computation time, memory, disk, elapsed time

Formulas in terms of n, k, size of alphabet

Scaleability

size n of string S, size k of substring P

string S may be of size $k = 10^{10}$ or more

size k of substring P is often 17 to 37

potential number of different substrings P is $|alphabet|^k$

$4^{37} = 2^{78} = 10^{23}$ approximately

Design Issues for Design Assignment 1 — Count Substrings

Data representations
- character in the string
- the string
- the substring
- the collection of statistics

Algorithms
- for enumerating each substring
- for updating statistics of a substring in the collection
- indexing and searching the collection of statistics

Interfaces
- String: how to iterate over string
- Collection: how to update the statistics for each substring

packed representation of characters in the string, or not?

pass-by-value versus pass-by-reference