Joshua Johnston

COT 4400

Project 17

Summary

The problem is to read two text files, and find the longest sequence of text that appears in both of the files. My first solution to this problem was to read each file and build a string for each of the files. Get all the sub-strings of each string, and store them into separate lists. Take the intersection of the two lists, and then take the longest string from that list. The time complexity of this solution is O(n^3). However, this solution runs out of memory when the strings length is over 2000.

My second solution is the same as above in theory, but with a much-needed improvement. I'm not storing every sub-string into a list. When I get a sub-string from string1 I see if string2 contains that sub-string or not then I add that sub-string into the list if and only if string2 contains it, and it's not a duplicate. Then find the longest string that is in the list. The time complexity is the same as above O(n^3).

I compared both solutions with strings length ranging from 15 to 2500. I can see a difference from the two. When I ran a test on the first solution with string lengths of 500, and 2500. It ran for four hours before running out of memory. I did the same test again on the second solution, and the time was only 282 milliseconds. I tested the second solution with strings length up to 10,000. Here is a table for the rest of my tests.

Name of file	Number of char	Simple: time in milliseconds	Improved: time in milliseconds
test'1'	15	3	1
test'2'	15		
test'1'	15	169153	2
test'3'	500		
test'3'	500	953218	107
test'4'	500		
test'4'	500	989688	102
test'5'	500		
test'8'	700	4331437	174
test'11'	700		
test'7'	1000	5927255	256
test'5'	500		
test'4'	500	4 hours then ran out of memory	282
test'6'	2500		
test'9'	3000	N/A	12056
test'10'	3000		
test5000'1'	5000	N/A	82290
test5000'2'	5000		
test7000'1'	7000	N/A	488926
test7000'2'	7000		
test10000'1'	10,000	N/A	2538038
test10000'2'	10,000		