平成 23 (2011) 年度 夏入試

東京大学情報理工学系研究科創造情報学専攻

プログラミング

注意事項

1. 試験開始の合図まで，この問題冊子を開いてはいけない．
2. この表紙の下部にある受験番号欄に受験番号を記入しなさい．
3. 解答用紙および下書き用紙が 1 枚ずつ配られる．それぞれに受験番号を記入しなさい．
4. 受験者に配られた USB メモリに ASCII コードで書かれたテキストファイル `c1.txt`, `c2.txt`, `s1.txt`, `s2.txt`, `s3.txt` が含まれている．
   試験開始前に，USB メモリから上記のファイルを自分の PC にコピーしなさい．ファイルの中身を確認し，PC から手を離しなさい．ファイルにアクセスできないなどの場合は試験監督に申し出なさい．USB メモリの中身は全受験者に共通である．
5. プログラミング言語は何を使ってもよい．
6. プログラミング言語のマニュアルは 1 冊に限り試験中に参照してもよい．**ネットワーク接続をしてはいけないが，各自の PC に入っているライブラリやプログラム断片を使用・流用することは自由である．**
7. 試験終了時までに，自分の PC 上に受験番号名のディレクトリ/フォルダを作成し，作成したプログラムおよび関連ファイルをその下にコピーしなさい．作成したディレクトリ/フォルダを各受験者に渡された USB メモリにコピーしなさい．
8. 試験終了時に，USB メモリ，解答用紙，下書き用紙を回収する．
9. 回収後，試験監督が巡回し，各受験者のプログラムの結果を簡単に確認するので，そのまま座席で待機しなさい．全員の確認が終わるまで部屋を出てはいけない．
10. 午後のプログラミングの口頭試問中にプログラムの動作をより精密に確認する．各自の PC 上でプログラムがすぐに実行できるようにしておきなさい．
11. 全員の確認が終了した後，**各自の PC とこの問題冊子を残し**，部屋から退出しなさい．

受験番号＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿

このページは空白.

このページは空白.

本問題では，辞書式の可逆圧縮を扱う．同じ文字列が繰り返し出現することが多いという性質を使う辞書式圧縮では，与えられた文字列に対し，その文字列の一部分（置換対象文字列）を別のより短い文字列（置換指定文字列）に置換して圧縮する．圧縮された文字列を元に戻すことを展開という．圧縮する前の文字列を元文字列，圧縮して得られた文字列を圧縮文字列と呼ぶ．

本問題では，n 文字（$n \leq 1000$）の元文字列 S において，a から b までの部分文字列を $S[a;b]$ と表記する（$0 \leq a \leq b \leq n-1$）．元文字列の長さは問 5 までは 1000 文字以下であり，問 6 で一般の場合に拡張する．元文字列はアルファベット小文字（a〜z），空白，ピリオドとコンマを要素とする．圧縮文字列は，元文字列を構成する文字に加え，数字（0〜9）を用い 10 進数表記の 3 桁，000 から 993 までの置換指定文字列を要素とする．なお，10 進数の表記では先頭の 0 は省略しない．置換対象文字列は固定長であり，長さ 6 であるとする．

i，j（$0 \leq i < j \leq n-6$）について部分文字列 $S[i;i+5]$ と部分文字列 $S[j;j+5]$ が等しい時，$S[j;j+5]$ が置換対象文字列となり，i の 10 進 3 桁表記である置換指定文字列と置き換える（図 1）．なお，元文字列における置き換えた $S[j;j+5]$ 部分は，さらなる置換の候補とはしない．置換により 6 文字が 3 文字に圧縮される．

置換指定文字列で利用される i を一意に特定するため，探索に次の制限を加える．置換対象文字列 $S[j;j+5]$ の探索は，$j=1$ から j を 1 ずつ増加する方向で行う．このとき，部分文字列 $S[i;i+5]$ に該当するものが複数ある場合には，最小の i を置換指定文字列とする．

$i+5 \geq j$ の場合，置換対象文字列と，置換元の部分文字列が重複する．展開において重複がある場合には，図 2 のように，重複していない部分から順次展開することにより，元文字列を得ることができる．

例	元文字列	圧縮文字列
	vwabcdefxyabcdefst	vwabcdefxy002st
	abababababababab	ab000000ab

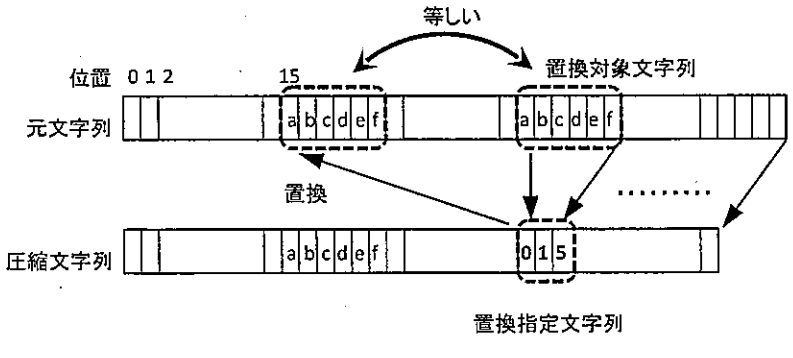

図 1　文字列の置換

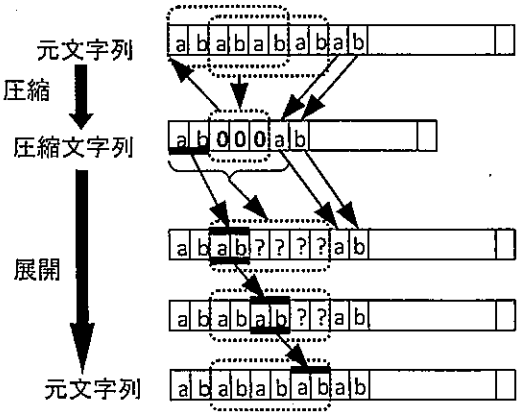

図 2　重複する置換対象文字列の例

問 1　プログラムを作らずに次の問いに答えよ．

1-1　圧縮文字列 aabbba000c001008a に対応する元文字列を答えよ．

1-2　元文字列 aabbccddaabbccddbbccddaa に対応する圧縮文字列を答えよ．

問 2　c1.txt，c2.txt に格納されている圧縮文字列に，10 進数 3 文字の置換指定文字列がそれぞれ何個含まれているか，プログラムを作成して答えよ．なお，ASCII コードでは，'a'=0x61（10 進表記 97），' '=0x20（10 進表記 32），','＝0x2C（10 進表記 44），'.'＝0x2E（10 進表記 46），'0'＝0x30（10 進表記 48）である．

問 3　元文字列を圧縮するための辞書を構築するプログラムを作成せよ．辞書は 6 文字の部分文字列に対し，その出現する位置の最小値を対応させるデータ構造である．図 3 に圧縮するための辞書構成の一例を示す．

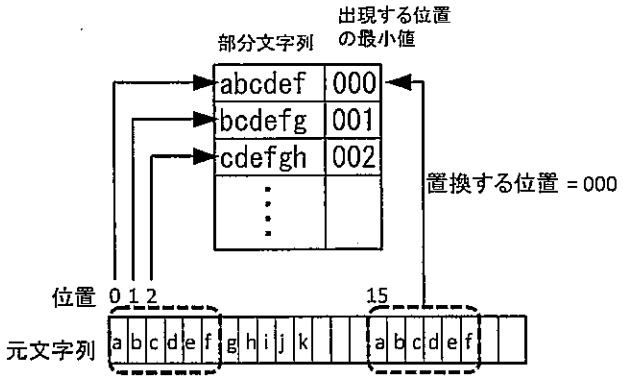

図 3　圧縮辞書の構成の一例

長さ n，$n \leq 1000$ の文字列 S から長さ 6 の部分文字列を検索するための辞書を作成せよ．具体的には，$i = 0$ から i を 1 ずつ増加させながら，順次 $S[i; i + 5]$ を切り出し，辞書を検索する．検索が成功する場合には，当該部分文字列が出現した位置の最小値を返す．検索が不成功である場合には，当該部分文字列と部分文字列の位置 i を辞書に追加する．

作成したプログラムを利用して s1.txt に格納されている文字列（89 文字）から辞書を作成せよ．終了後の辞書中に登録された部分文字列の出現位置の数を答えよ．

問 4　問 3 で作成したプログラムを利用して，元文字列を圧縮するプログラムを作成せよ．作成したプログラムを利用して，s1.txt，s2.txt それぞれについて，格納されている元文字列から得られる圧縮文字列の，長さと末尾の 10 文字を答えよ．

問 5　圧縮文字列を展開し，出力するプログラムを作成せよ．作成したプログラムを利用して，c1.txt，c2.txt にそれぞれ格納されている圧縮文字列について，展開して得られる元文字列の，長さと末尾の 10 文字を答えよ．

問 6　元文字列が 1000 文字より長い場合への拡張として，元文字列を先頭から 1000 文字毎のブロックに区切り，ブロック単位で処理を行うように問 4 のプログラムを拡張せよ．同様に，問 5 のプログラムも 1000 文字より長い元文字列に対応するように拡張せよ．

s3.txt に格納されている元文字列を用いて，拡張した圧縮プログラムを用いて元文字列を圧縮し，得られた圧縮文字列を拡張した展開プログラムで展開することにより，元文字列が復元することを確認せよ．

このページは空白.

このページは空白.