Laboratorio di Programmazione a.a. 2015-16

Progetto "Codifica di Huffman" – Parte III

16 Maggio 2016

1. Codici di Huffman basati su statistiche relative alla frequenza delle lettere nei testi

Supponendo di voler comprimere testi "letterari" in cui compaiono prevalentemente lettere (minuscole e qualche volta maiuscole) e spazi, accompagnati da altri simboli meno frequenti (interpunzione, apice, doppio apice, parentesi, cifre...), è possibile costruire un albero di Huffman basato su dati statistici condivisi relativamente alle frequenze di lettere e parole nei testi. In tal caso non è necessario codificare l'albero nell'intestazione del documento compresso perché la sua struttura è convenuta una volta per tutte e resa disponibile a chi deve comprimere o decomprimere documenti.

Scrivi un programma per costruire un "buon" albero di Huffman interpretando opportunamente i dati statistici riportati nella pagina seguente, relativi alla lingua inglese utilizzata negli articoli di tipo giornalistico. Tieni conto che le codifiche devono essere estese in modo ragionevole ad altri caratteri che si possono ritrovare in un testo giornalistico.

Per per poter riutilizzare i programmi già sviluppati, il "peso" assegnato a un carattere può essere definito dal numero *atteso* (in senso statistico) di occorrenze in un ipotetico documento di 100000 caratteri.

2. Compressione e decompressione

Modifica, quindi, il programma sviluppato a lezione (e disponibile attraverso le pagine del corso) in modo tale da utilizzare l'albero realizzato nel punto precedente sia per la compressione che per la decompressione, senza codificarlo nel file compresso, indipendentemente dal contenuto del documento specifico che si vuole comprimere.

L'albero di Huffman può essere gestito convenientemente applicando la classe `HuffmanTree` realizzata nella II parte di questo progetto.

3. Sperimentazione

Infine, confronta sperimentalmente i risultati del programma realizzato e di quello sviluppato a lezione in termini di fattore di compressione. In particolare, puoi utilizzare il campione di testo associato a questo esercizio di laboratorio (un breve collage di brevi articoli giornalistici).

Tipiche statistiche relative al linguaggio inglese giornalistico

Lunghezza media di una parola: 5.1 lettere
Lunghezza media di una frase: 24.5 parole

Frequenza percentuale delle lettere dell'alfabeto inglese

a	8.167
b	1.492
c	2.782
d	4.253
e	13.00
f	2.228
g	2.015
h	6.094
i	6.966
j	0.153
k	0.772
l	4.025
m	2.406
n	6.749
o	7.507
p	1.929
q	0.095
r	5.987
s	6.327
t	9.056
u	2.758
v	0.978
w	2.360
x	0.150
y	1.974
z	0.074