Laboratorio di Programmazione a.a. 2018-19

Progetto "Codifica di Huffman" – Parte III

24 Maggio 2019

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. Risulta invece ancora opportuno codificare il numero complessivo di caratteri presenti nel documento originale.

Scrivi un programma per costruire un "buon" albero di Huffman interpretando opportunamente i dati statistici riportati nella pagina seguente, relativi alla lingua inglese utilizzata in articoli di tipo 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. Tieni inoltre conto che le codifiche devono essere estese in modo ragionevole a tutti i caratteri corrispondenti ai 128 codici ASCII ammessi, per cui si rende necessario attribuire preliminarmente una frequenza a ciascuno di essi. In particolare:

- *Lettere maiuscole*: ogni *frase* inizia con una lettera maiuscola, ma saltuariamente se ne possono trovare delle altre in corrispondenza alle occorrenze di nomi propri, acronimi, ecc.; in prima approssimazione si può assumere che la distribuzione relativa (%) delle lettere maiuscole sia analoga a quella delle lettere minuscole.
- *Cifre*: saltuariamente possono essere riportati dei dati numerici; in prima approssimazione, si può assumere che le cifre siano distribuite uniformemente (cioè siano equiprobabili).
- *Simboli di interpunzione*: ogni frase termina generalmente con un punto, con rare eccezioni; fra gli altri simboli di interpunzione la virgola è il più frequente, probabilmente almeno altrettanto frequente del punto.
- *Spazi bianchi*: dopo ogni *parola*, con l'eventuale eccezione dell'ultima, c'è uno spazio bianco.
- *Capolinea*: ogni paragrafo si conclude con un capolinea; in prima approssimazione si può assumere che un paragrafo si componga di poche *frasi* (per esempio due o tre).
- *Virgolette*: nei testi che riportano dialoghi e discorsi indiretti le virgolette possono essere relativamente frequenti (anche una coppia per frase), ma occorre mediare con la probabilità che un testo sia di questo tipo.
- *Altri caratteri*: ai caratteri meno frequenti si associa il minimo numero di occorrenze attese (una su 100000).

2. Compressione e decompressione

Modifica 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. Poiché il peso di ciascun carattere nell'albero è convenzionale, per determinare la lunghezza del documento originale occorre contare i caratteri nel corso della lettura.

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 i campioni di testo associati a questo esercizio di laboratorio (un collage di brevi articoli giornalistici, un articolo di divulgazione scientifica e un testo letterario in inglese).

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	12.702
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	2.361
x	0.150
y	1.974
z	0.074