Laboratorio di Programmazione a.a. 2023-24

Progetto “Codifica di Huffman” – Parte II

29 Maggio 2024

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

Supponendo di voler comprimere testi “narrativi” 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 (per evitare problemi tecnici nel ripristino dell’ultimo carattere del testo).

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 ogni frase contenga esattamente una lettera maiuscola e che la distribuzione relativa (%) delle lettere maiuscole sia analoga a quella delle lettere minuscole.
- *Simboli di interpunzione*: ogni frase termina generalmente con un punto, con rare eccezioni; in prima approssimazione si può assumere che ci siano tanti punti quante frasi e che in ogni frase ci sia, in media, uno dei seguenti simboli di interpunzione diversi dal punto (tra parentesi è riportata la distribuzione indicata in percentuale relativa all’insieme dei simboli di interpunzione diversi dal punto): la virgola (45%), l’apostrofo (40%), i due punti (5%), il punto e virgola (5%), e i doppi apici (5%).
- *Spazi bianchi*: dopo ogni *parola*, con l’eventuale eccezione dell’ultima, c’è uno spazio bianco. Per semplicità si può assumere che ci siano tanti spazi bianchi quante parole.
- *Altri caratteri e cifre*: a tutti gli altri caratteri (incluso le cifre ed il capolinea) si associa il minimo numero di occorrenze attese (una su 100000).

Si può poi raffinare l’analisi statistica considerando in maniera più mirata altri caratteri frequenti, come, ad esempio, le cifre ed il carattere di capolinea (carattere per andare a capo). Ad esempio, si può assumere che ci sia un carattere capolinea per ogni paragrafo e che un paragrafo si componga di poche frasi (ad esempio 2 o 3 frasi).

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