Laboratorio di Programmazione a.a. 2019-20

Progetto "Codifica di Huffman" – Parte I

19 Maggio 2020

1. Codici di Huffman dei caratteri

Scrivi un metodo statico in Java che, dati i nomi (`String`) di due file di testo, uno di input e l'altro di output, analizza il file di input per determinarne l'*istogramma* delle frequenze dei caratteri contenuti (concretamente, un array i cui indici corrispondono a caratteri e i cui valori indicano il numero complessivo di occorrenze di ciascun carattere nel testo); quindi riporta nel file di output una tabella conforme alle seguenti specifiche: i caratteri ammessi sono quelli con codici ASCII nell'intervallo 0–127; il file generato deve prevedere una riga di testo per ciascun carattere effettivamente utilizzato, riga nella quale saranno riportati il codice ASCII, il simbolo del carattere corrispondente, il numero di occorrenze nel file di input, il codice di Huffman e la rispettiva lunghezza (numero di bit). I caratteri speciali *nuova-linea*, *capo-linea* e *tabulazione* possono essere rappresentati in Java dai letterali `'\n'`, `'\r'`, `'\t'` (e lo sono in modo analogo nelle stringhe).

Per generare l'istogramma, l'albero e la tabella dei codici di Huffman puoi avvalerti dei metodi statici sviluppati a lezione. Per l'accesso e le operazioni sui file di testo utilizza le classi del package `huffman_toolkit` disponibile (con una sintetica documentazione) attraverso le pagine del corso.

2. Analisi di file di testo

Scrivi un metodo statico in Java che, dato il nome (`String`), genera un file di testo "random" composto da caratteri i cui codici ASCII sono prodotti in modo casuale, con distribuzione uniforme, nell'intervallo 0–127. (Per l'accesso e le operazioni sui file di testo utilizza le classi del package `huffman_toolkit` disponibile attraverso le pagine del corso.)

Per la scelta aleatoria dei caratteri puoi basarti sulla funzione predefinita della classe Java `Math`

```
public static double random()
```

che ad ogni invocazione restituisce un numero (`double`) casuale compreso nell'intervallo [0.0, 1.0[.

Analizza la tabella di cui al punto **1** per un sorgente Java e per un file di testo "random" di dimensioni analoghe, in termini di numero complessivo di caratteri. Confronta, quindi, le caratteristiche generali dei codici di Huffman associati ai caratteri nelle due situazioni, in particolare riguardo alla variabilità delle relative lunghezze.

3. Stima e verifica delle dimensioni dei file "compressi"

In base ai dati riportati nella tabella, in particolare lunghezza del codice di Huffman e numero di occorrenze di ciascun carattere, calcola la dimensione in *byte* del testo compresso. Tieni conto che, in relazione al modello sviluppato a lezione, un byte si compone di 7 bit utili sugli 8 disponibili. Calcola, inoltre, il numero di caratteri che compongono l'intestazione e quindi la dimensione complessiva del file "compresso". Esegui questi calcoli, che possono essere portati a termine agevolmente tramite uno spreadsheet, per entrambi i file di testo analizzati nel punto precedente.

Applica, infine, il programma di compressione sviluppato a lezione e verifica le stime effettuate confrontandole con le dimensioni effettive dei file "compressi" prodotti.