Laboratorio di Algoritmi

Progetto "Digitalizzazione" (giugno 2024)

Nota: La scadenza del progetto è fissata per lunedì 10 giugno **compreso**.

Nota: Si consiglia di consultare sulla pagina web il documento che riporta le avvertenze utili per lo svolgimento del progetto. Si consiglia anche di verificare di tanto in tanto gli aggiornamenti a questo documento, che potranno riportare risposte ai dubbi degli studenti e correzioni di eventuali errori.

Il problema La pubblica amministrazione di un paese vuole digitalizzare tutti i documenti conservati negli archivi distribuiti sull'intero territorio. Istituisce quindi un centro specializzato e lo dota delle risorse sufficienti a eseguire la scansione di quantità considerevoli di documenti. A questo centro i singoli archivi dovranno spedire parte dei documenti in loro possesso, organizzando la spedizione per lotti giornalieri. Ogni documento è caratterizzato da un numero di pagine, che determina il tempo necessario a compierne la scansione. In generale, di ogni documento possono esistere più copie, conservate in archivi diversi e di qualità non necessariamente uniforme. Dati il numero dei documenti e il tempo limitato a disposizione, digitalizzare tutti i documenti esistenti è impossibile, per cui si eviterà per prima cosa di digitalizzare più copie dello stesso documento. Poiché comunque non è garantito che il tempo stabilito sia sufficiente a digitalizzare una copia di ciascun documento esistente, si pone il problema di scegliere quali considerare.

La scelta è complicata dal fatto che, prima di poter cominciare a inviare documenti, gli archivi devono essere visitati da tecnici specializzati, ed esiste un solo gruppo di tecnici. Ogni visita richiede approssimativamente una giornata di lavoro. L'archivio visitato può cominciare a inviare documenti il giorno dopo la visita. La trasmissione è abbastanza rapida da assumere che i documenti arrivino al centro prima che comincino le operazioni di scansione della giornata[1]. Per ridurre il rischio di errori, ogni documento deve essere digitalizzato interamente nel giorno in cui viene trasmesso.

Dal punto di vista gestionale, si vuole compiere un'analisi preliminare dei dati, che ne estragga informazioni utili. In particolare, si vuole completare l'anagrafe dei documenti determinando per ogni documento il numero di copie e la loro qualità, nonché la distribuzione empirica del numero di copie. Analogamente, si vuole completare l'anagrafe degli archivi determinando informazioni aggregate sulla consistenza quantitativa, la qualità e la rarità dei documenti che conservano.

Si intende poi calendarizzare la spedizione dei documenti da ciascun archivio al centro di digitalizzazione con lo scopo di massimizzare la quantità e qualità totale dei documenti processati. Il problema è computazionalmente difficile, e verrà quindi affrontato decomponendolo in problemi più facilmente affrontabili. L'idea fondamentale sarà di determinare un ordine ragionevole di visita degli archivi in base al numero, alla qualità e alla rarità dei documenti che contengono, privilegiando quelli più ricchi di copie migliori di documenti più rari, in modo che abbiano il massimo numero di giorni a disposizione per inviarli al centro di digitalizzazione. Più in dettaglio, dividendo la qualità di una copia per il numero totale di copie dello stesso documento conservate nei vari archivi (ovvero moltiplicando la qualità della copia e la rarità del documento) si ottiene una misura del pregio della copia stessa. Sommando tale misura per tutti i documenti conservati in un archivio si ottiene un

[1] In pratica, basta che alcuni documenti arrivino prima e gli altri via via, a un ritmo non inferiore a quello di digitalizzazione, in modo che non ci si debba mai fermare ad attenderli.

indice di pregio per l'archivio stesso, in base al quale determinare l'ordine di visita, cioè per pregio non crescente (in caso di parità, si consideri semplicemente l'indice dell'archivio stesso). È possibile che alcuni archivi non siano visitabili nel tempo disponibile, e quindi i loro documenti risultino del tutto indisponibili. Ciò fatto, sarà determinato giorno per giorno l'insieme di archivi che possono spedire documenti, e quindi l'insieme delle copie disponibili. Si potrà allora scegliere quale sottoinsieme di copie spedire in quel giorno dai vari archivi abilitati. L'obiettivo è massimizzare la qualità totale delle copie, rispettando i seguenti vincoli: i) il numero di pagine da digitalizzare non deve superare la capacità del centro; ii) ogni copia va scandita interamente, non in parte; iii) non devono essere digitalizzate copie diverse dello stesso documento; iv) solo copie conservate in archivi già abilitati possono essere digitalizzate. Per semplicità, non aggiorneremo l'indice di rarità dei documenti digitalizzati via via, e quindi gli indici di pregio e l'ordine di visita degli archivi rimanenti, anche se avrebbe più senso farlo.

Infine, si vuole studiare il grado di sovrapposizione fra archivi diversi e la possibilità di chiuderne alcuni in quanto ridondanti, ovvero convertirli in archivi secondari dedicati a conservare copie di sicurezza. In questo caso, la capacità non pone problemi, dato che non c'è un orizzonte temporale limitato. Il problema che si pone è quindi di determinare un sottoinsieme di archivi (al limite, tutti) che nel complesso contenga almeno una copia di ognuno dei documenti complessivamente disponibili (esclusi quindi quelli che nessun archivio possiede). Siccome questo problema è computazionalmente difficile, verrà affrontato in maniera euristica con il seguente algoritmo *greedy adattivo*, cioè un algoritmo *greedy* il cui criterio di scelta non è costante durante l'esecuzione, ma viene aggiornato via via. L'algoritmo parte da un insieme di archivi selezionati vuoto, e quindi un insieme di documenti selezionati (cioè disponibili negli archivi selezionati) anch'esso vuoto. Ad ogni passo, l'algoritmo determina l'archivio non ancora selezionato che contiene il massimo numero di documenti non ancora selezionati. Quindi, i documenti via via selezionati vanno esclusi dal conteggio, aggiornando il criterio di scelta. In caso di parità, l'algoritmo sceglie l'archivio i cui documenti non ancora selezionati hanno la qualità totale massima. In caso di ulteriore parità, sceglie l'archivio di indice minimo. L'algoritmo termina quando tutti i documenti disponibili compaiono in almeno uno degli archivi selezionati, cioè gli eventuali archivi non selezionati contengono solo documenti già selezionati.

Il progetto Il progetto richiede la stesura di un programma che legga da un file di testo le informazioni relative ai documenti disponibili in ogni archivio. Il file si apre con una riga che contiene il numero di giorni disponibili per l'intero procedimento, il numero di pagine digitalizzabili ogni giorno dal centro, il numero di documenti e il numero di archivi. Per esempio:

```
5 100 10 3
```

indica che l'orizzonte temporale è di 5 giorni, che ogni giorno il centro può scandire 100 pagine e che ci sono 10 documenti e 2 archivi.

Ciascuna riga seguente si riferisce a un singolo documento e ne riporta il nome e il numero di pagine. Il nome è descritto da una sola parola, di al più 20 caratteri[2].

```
doc6 10
doc1 24
atto3 9
mappa2 3
```

[2]Questa indicazione serve a dimensionare le stringhe. Non va intesa come un invito a considerare $O(1)$ il numero di possibili documenti nelle analisi di complessità.

```
atto9 58
mappa4 33
atto5 28
mappa3 83
doc4 14
doc3 52
```

indica che il documento `doc6` è di 10 pagine, `doc1` è di 24 pagine, e così via.

Seguono blocchi di righe dedicate ai singoli archivi. Ciascun blocco comincia con una riga che riporta il numero di documenti disponibili nell'archivio. Ogni riga seguente descrive un documento, riportandone il nome e la qualità, espressa da un numero intero positivo. È possibile che ci siano documenti non disponibili in alcuno degli archivi elencati. Al contrario, tutti i documenti disponibili negli archivi compaiono nell'elenco generale dei documenti. Inoltre, ogni archivio contiene una sola copia dei propri documenti, dato che è stata preliminarmente scelta la migliore. Infine, vi assume che i documenti siano integri, e quindi tutte le copie dello stesso documento conservate nei diversi archivi abbiano lo stesso numero di pagine, indicato più sopra. Per esempio:

```
7
doc1 65
atto9 12
doc6 95
mappa2 43
mappa4 29
mappa3 15
doc3 72
6
doc6 47
mappa2 87
doc1 43
atto3 28
mappa3 87
doc4 14
5
mappa2 3
atto9 58
mappa4 33
doc4 74
doc3 32
```

indica che il primo archivio contiene 7 documenti, cioè `doc1` con qualità 65, `atto9` con qualità 12, `doc6` con qualità 95, ecc. . . ; il secondo archivio contiene 6 documenti, il terzo 5 documenti.

L'analisi preliminare richiede di determinare per ogni documento disponibile (escludendo quindi quelli che nessun archivio conserva) quanti archivi lo possiedono e quali siano la qualità massima, media e minima disponibile. I documenti vanno stampati a video: la prima riga fornirà il loro numero seguito dalla parola chiave `documenti`, le successive (una per documento) il nome, la qualità massima, media (arrotondata alla prima cifra decimale) e minima, separati da spazi singoli. Queste informazioni saranno separate da singoli spazi. I documenti vanno ordinati per qualità massima non crescente. In caso di parità, vanno ordinati per qualità media non crescente, poi per qualità minima non crescente e infine in ordine alfabetico. Nell'esempio:

```
9 documenti
doc6 95 71.0 47
mappa3 87 51.0 15
mappa2 87 44.3 3
doc4 74 44.0 14
doc3 72 52.0 32
doc1 65 54.0 43
atto9 58 35.0 12
mappa4 33 31.0 29
atto3 28 28.0 28
```

cioè solo 9 dei 10 documenti in anagrafe sono effettivamente disponibili. Quello con la massima qualità è `doc6 95 71.0 47`, che ha qualità massima 95, minima 47 e media 71 (essendo presente in due copie, per cui $95 + 47/2 = 71$). I due documenti successivi sono `mappa3` e `mappa2`, entrambi con qualità massima pari a 87, ma `mappa3` ha qualità media $(87+15)/2 = 51$, superiore a quella di `mappa2`, che è solo $(43 + 87 + 3)/3 = 44.\overline{3}$.

Bisognerà poi determinare il profilo di distribuzione del numero di copie disponibili per ciascun documento, vale a dire il numero dei documenti non disponibili, quelli dei documenti per i quali esiste una sola copia, due copie, ecc... Questa informazione andrà stampata a video, preceduta da una riga che contiene solo la parola chiave `profilo`. Ogni riga seguente indica il numero di copie e il numero di documenti corrispondente, per valori crescenti da zero al massimo numero di copie, senza lacune (cioè può succedere che nessun documento compaia in uno specifico numero di copie inferiore al massimo). Nell'esempio:

```
profilo
0 1
1 1
2 7
3 1
```

cioè un solo documento è del tutto assente (`atto5`), uno è presente in un solo archivio (`atto3`), 7 documenti sono presenti in due copie e uno in tre copie (`mappa2`).

Si vuole poi determinare per ogni archivio il numero totale di pagine disponibili, la loro qualità totale (stimata attraverso la somma delle qualità dei singoli documenti) e la loro rarità. Quest'ultima è definita per ogni documento dall'inverso del numero di copie totale conservato nei vari archivi (ovviamente, escludendo i documenti non conservati in alcun luogo). L'indice di rarità dell'intero archivio è semplicemente la somma di quelli dei documenti che conserva. Si stamperà a video una riga contenente il numero di archivi seguito da uno spazio e dalla parola chiave `archivi`. Ogni riga seguente riporterà le informazioni relative a un archivio, vale a dire l'indice numerico (cioè il numero d'ordine con cui appare nel file dei dati), il numero totale di pagine, il valore totale della qualità e quello della rarità (arrotondata alla seconda cifra decimale), separati da singoli spazi. Gli archivi vanno ordinati per numero totale di pagine non crescente. In caso di parità, vanno ordinati per indice crescente. Nell'esempio:

```
3 archivi
1 263 331 3.33
3 160 200 2.33
2 143 306 3.33
```

cioè l'archivio con il massimo numero di pagine è il primo, che contiene 263 pagine, con qualità totale pari a $65 + 12 + 95 + 43 + 29 + 15 + 72 = 331$ e rarità totale pari a $0.5 + 0.5 + 0.5 + 0.\overline{3} + 0.5 + 0.5 + 0.5 = 3.\overline{3}$. Segue il terzo archivio, con 160

pagine, qualità totale pari a 200 e rarità totale pari a $2.\bar{3}$. Infine, il secondo archivio contiene 143 pagine, con qualità totale pari a 306 e rarità totale pari a $3.\bar{3}$.

Calendarizzate le visite e le spedizioni come indicato nella sezione precedente, si stamperà la soluzione come segue. La prima riga deve riportare la parola chiave `calendario`. Ciascuna riga seguente riporta le operazioni relative a un giorno, precisamente:

- l'indice dell'archivio visitato, sostituito da un trattino nei giorni in cui non vi sono più visite;
- la qualità totale e il numero totale di pagine delle copie di documenti digitalizzati;
- l'elenco delle copie digitalizzate (descritte dal nome del documento e dall'indice dell'archivio di provienienza), ordinato alfabeticamente per nome.

Tutte queste informazioni sono separate da spazi bianchi. Nell'esempio:

```
calendario
1 0 0
2 275 89 doc1 1 doc3 1 doc6 1 mappa2 1
3 115 92 atto3 2 mappa3 2
- 132 72 atto9 3 doc4 3
- 33 33 mappa4 3
```

cioè il primo giorno si visita il primo archivio (perché ha pregio massimo, pari a $158.\bar{3}$) e non si digitalizza alcun documento. Il secondo giorno si visita il secondo archivio (perché ha pregio 152.5) e si digitalizzano i documenti `doc1`, `doc3`, `doc6` e `mappa2`, tutti conservati nel primo archivio (ovviamente, dato che è l'unico abilitato), per un totale di $24 + 52 + 10 + 3 = 89$ pagine e una qualità pari a $65 + 72 + 95 + 43 = 275$. Il terzo giorno si visita il terzo archivio (perché ha pregio 99.5) e si digitalizzano i documenti `atto3` e `mappa3`, conservati nel secondo archivio, per un totale di $9 + 83 = 92$ pagine e una qualità pari a $28 + 87 = 115$. Si noti che sarebbe disponibile anche una copia di `atto3` conservata nel primo archivio, ma con una qualità inferiore, oltre ad altri documenti non ancora digitalizzati in entrambi gli archivi abilitati. Il quarto giorno non si visita alcun archivio e si digitalizzano i documenti `atto9` e `doc4` conservati nel terzo archivio, mentre l'ultimo giorno si digitalizza il documento `mappa4`, conservato nel terzo archivio.

Infine, si dovrà stampare l'elenco degli archivi selezionati. La prima riga riporterà il numero di archivi selezionati, seguito dalle parole chiave `archivi selezionati`. La riga seguente riporterà l'elenco degli indici degli archivi selezionati, separati da spazi, nell'ordine con cui sono stati selezionati dall'algoritmo. Nell'esempio:

```
2 archivi selezionati
1 2
```

cioè i primi due archivi contengono complessivamente almeno una copia di tutti i documenti disponibili. Al primo passo, l'algoritmo seleziona il primo archivio, che ha 7 documenti nuovi. Al secondo passo, l'archivio 2 ha 2 documenti aggiuntivi (`atto3` e `doc4`), mentre l'archivio 3 ne ha uno solo (`doc4`). Di conseguenza, l'algoritmo sceglie l'archivio 2. A questo punto, l'archivio 3 non ha nuovi documenti da aggiungere, e quindi l'algoritmo termina.

Chiarimenti

In questa sezione saranno riportate le risposte a domande e dubbi.

È possibile che vi siano documenti con un numero di pagine superiore alla capacità giornaliera del centro di digitalizzazione?

Sì, è possibile. Ovviamente, sono documenti che risultano disponibili (e anche "digitalizzabili"), ma non verranno mai digitalizzati a causa del vincolo di non distribuire la digitalizzazione di uno stesso in più giorni.

La lunghezza dei nomi dei documenti è assunta limitata ai fini dell'allocazione di memoria per contenerli, ma non limitata ai fini della valutazione del numero di possibili documenti. Come va considerata ai fini dell'analisi di complessità temporale e spaziale?

A rigore dovrebbe essere considerata non limitata. Considererò corrette, però, entrambe le ipotesi.