

Cognome e Nome _____

Matricola _____

Big data – Graph Mining A.A. 2024/2025

Appello del 13 febbraio 2025

Esercizio 1 [8 punti]

Siano date delle quadruple (**id_studente**, **id_insegnamento**, **voto**, **data**). Ciascuna quadrupla rappresenta il verbale di un esame svolto in data **data** dallo studente con matricola **id_studente** per l'insegnamento **id_insegnamento**, con voto **voto**.

Si progettino (eventualmente utilizzando più step di computazione tra loro concatenati) le funzioni *Map* e *Reduce* per restituire in output **uno degli studenti che ha conseguito la media aritmetica più alta dei voti degli esami sostenuti**, utilizzando spazio locale $O(n^{1/3})$, dove n è il numero di studenti. Si assuma che le matricole siano numeri interi da 0 a $n-1$, e che ogni studente ha svolto un numero costante di esami.

Suggerimento: si utilizzi un primo step di computazione per calcolare la media aritmetica dei voti di ciascuno studente; successivamente, si utilizzino gli step di computazione necessari a trovare uno studente con la media più alta.

Esempio di input:

- (0, 100, 20, 10/04/2023)
- (1, 100, 30, 07/09/2024)
- (0, 210, 28, 05/07/2023)
- (1, 210, 28, 10/07/2024)
- (0, 300, 24, 01/07/2024)

In questo input abbiamo 5 esami.

Il suggerimento consiglia di calcolare prima le coppie (**id_studente**, **media**)

che nell'esempio risulteranno essere:

- (0, 24)
- (1, 29)

Per poi calcolare uno degli studenti che ha conseguito la media più alta (1 nell'esempio).

Esercizio alternativo [valutazione massima 4 punti]: assumere di avere in input le coppie (**id_studente**, **media**) (con **id_studente** che varia da 0 a $n-1$) e dare in output uno degli studenti con la media più alta, utilizzando spazio locale $O(n^{1/3})$.

Esercizio 2 [9 punti]

Si consideri la **variante multi-hash con due tabelle hash** dell'algoritmo *Park Chen Yu*, nell'ambito del problema degli *itemset frequenti*, e se ne descriva in dettaglio il funzionamento.

Si dica inoltre, in modo **sintetico e puntuale**, **giustificando la risposta**:

- a. cosa deve accadere perché una coppia venga considerata candidata ad essere frequente;
- b. come avviene la compressione delle tabelle hash;
- c. quale vantaggio offre rispetto all'algoritmo PCY con singola tabella hash.

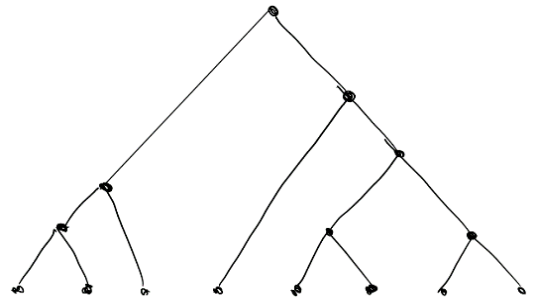
Esercizio 3 [9 punti]

Si descriva come operano gli algoritmi di clustering gerarchici o agglomerativi che partono da tutti cluster singoletti.

Si dica in modo **sintetico e puntuale**, giustificando le risposte:

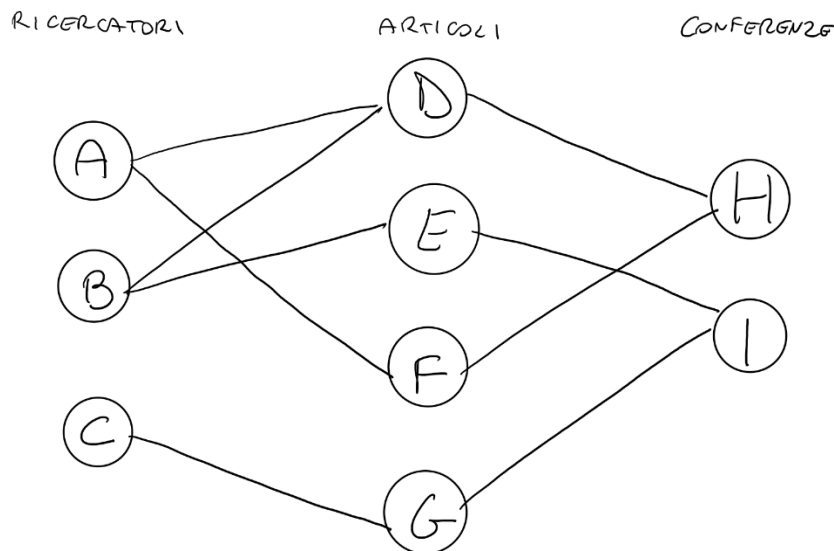
- in che modo possono essere scelti i cluster da unire;
- quali possono essere le condizioni di terminazione del clustering.

Infine, si faccia un esempio di clustering sul piano euclideo (partendo da un insieme di 8 punti) applicando un criterio di unione di cluster a scelta (si dica quale criterio si sceglie) e andando avanti fino ad ottenere un unico cluster, in modo che il processo di clusterizzazione effettuato corrisponda al dendrogramma in figura (nel quale i nodi vanno etichettati con i cluster corrispondenti).



Esercizio 4 [8 punti]

Si consideri la rete sociale in figura, che rappresenta con un grafo 3-partito le relazioni tra ricercatori, articoli scientifici e conferenze: l'insieme di nodi più a sinistra rappresenta i ricercatori, quello centrale gli articoli e quello di destra le conferenze; un arco tra un ricercatore e un articolo indica il fatto che il ricercatore è autore dell'articolo, mentre un arco tra un articolo e una conferenza indica il fatto che l'articolo è stato presentato alla conferenza.



Si impostino le equazioni per il calcolo del **SimRank** rispetto all'autore A con parametro di teleport β generico (si noti che il grafo è non diretto, e quindi ogni arco va considerato come doppiamente diretto).

Regole per lo svolgimento della prova scritta:

- Per svolgere il compito si hanno a disposizione **90 minuti**
- Scrivere **subito** nome, cognome, matricola su OGNI FOGLIO (**compreso questo**).
- Durante la prova scritta **non** è possibile abbandonare l'aula.
- Non è ammesso **per nessun motivo** comunicare in qualsiasi modo con altre persone
- Non è possibile consultare appunti, libri e dispense.
- Qualsiasi strumento elettronico di calcolo o comunicazione (telefoni cellulari, calcolatrici, palmari, computer, etc...) deve essere **completamente disattivato** e **depositato in vista sulla cattedra**
- Mettere in vista sul banco un valido documento di identità.