Cognome e Nome _____

Matricola _____

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

Esonero del 29 novembre 2024

Esercizio 1 (8 punti)

Siano date in input:

- Delle coppie **(id_autore, id_articolo)** che rappresentano i dati sull'essere autore di un articolo scientifico: l'autore **id_autore** (che varia da 0 a m-1) ha scritto l'articolo **id_articolo** (che varia da 0 a n-1); si noti che un articolo può avere più di un autore.
- Delle triple **(id_articolo, "cita", id_articolo_citato)** che contengono le informazioni circa le citazioni tra articoli scientifici: nell'articolo **id_articolo** è presente una citazione all'articolo **id_articolo_citato**.

Si progettino (eventualmente utilizzando più step di computazione tra loro concatenati) le funzioni *Map e Reduce* per dare in output un articolo scientifico che ha ricevuto il più alto numero di citazioni, con la lista di tutti i suoi autori, utilizzando spazio locale $O(n^{1/3})$.

L'output deve essere nel formato **(id_articolo, (id_autore_1, …, id_autore_k))**, dove k è il numero di autori dell'articolo **id_articolo**.

Si discuta il *reducer size* di tutti gli step individuati, assumendo che ogni articolo riceve un numero di citazioni costante, non dipendente quindi da *n*, e che ugualmente costante sia il numeri di autori di ogni articolo.

Esercizio 2 (9 punti)

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

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

a. In che modo vengono compresse le tabelle hash, e quale è il fattore di compressione.
b. Quando in questa variante, si può escludere che una coppia sia candidata ad essere frequente.
c. Perché viene impiegato il metodo delle triple per memorizzare i contatori delle coppie di itemset candidate ad essere frequenti.

Regole per lo svolgimento della prova scritta:

- Per svolgere il compito si hanno a disposizione **80** 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.