Cognome e Nome _____

Matricola _____

Big data – Graph Mining
A.A. 2019/2020

Esonero del 16 dicembre 2019

Esercizio 1 (8 punti)

Si progettino (eventualmente utilizzando più step di computazione tra loro concatenati) le funzioni *Map* e *Reduce* per generare tutti i path di lunghezza 3 archi (quindi con 4 nodi) in un **grafo diretto non pesato** G=(V,E). Si assuma che un arco del grafo di input, che va dal nodo i al nodo j, sia codificato come la coppia (i, j).

Suggerimento: prima trovare i path di 2 archi con la tecnica vista a lezione: il Map genera per ogni arco (i, j) 2 coppie (j, ("primo", i)) e (i, ("secondo",j)) e il reduce analizza le coppie che hanno in comune la stessa chiave k e…

Esercizio 2 (10 punti + 4 punti + 2 punti)

Si calcoli, evidenziando tutti i passaggi fatti riga per riga, la matrice delle firme minhash dei seguenti insiemi S_1, S_2, S_3, S_4 e S_5 secondo le permutazioni indotte dalle 3 funzioni hash evidenziate in tabella.

Riga	S_1	S_2	S_3	S_4	S_5	(x+2) mod 7	(2x+1) mod 7	(3x) mod 7
0	**0**	**0**	**1**	**1**	**0**			
1	**0**	**1**	**1**	**1**	**0**			
2	**1**	**0**	**0**	**0**	**0**			
3	**1**	**1**	**0**	**0**	**1**			
4	**1**	**0**	**0**	**0**	**1**			
5	**0**	**1**	**0**	**1**	**1**			
6	**1**	**0**	**0**	**0**	**1**			

Qual è la proprietà fondamentale che mette in relazione le funzioni minhash con la Jaccard Similarity degli insiemi? Discutere se nell'esercizio svolto tale proprietà risulta empiricamente rispettata.

[Opzionale: si fornisca la dimostrazione formale della proprietà enunciata]

Esercizio 3 (10 punti)

Si consideri l'algoritmo di *Bradley, Fayyad e Reina* per il clustering di punti in uno spazio euclideo con d dimensioni. Si dica in modo **sintetico e puntuale, giustificando la risposta**:

a. Se tale algoritmo è di tipo gerarchico o di tipo con assegnamento di punti
b. L'assunzione di fondo che questo algoritmo fa sui punti di ogni cluster
c. Come viene memorizzato in memoria un cluster, quanta memoria occupa tale memorizzazione e perché si memorizzano tali informazioni
d. Cosa sono i *discard set* e perché vengono chiamati così.
e. Come è definita la distanza di Mahalanobis e che ruolo riveste nell'algoritmo.

Si fornisca infine un esempio pratico di dominio applicativo in cui ha senso applicare l'algoritmo BFR, giustificando la risposta.

Regole per lo svolgimento della prova scritta:

- Per svolgere il compito si hanno a disposizione **100** 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.