Organización de Datos (75.06) Primer Cuatrimestre de 2014. Examen parcial, primera oportunidad. [2014_1c_Parcial_1]

Importante: Antes de empezar complete nombre y padrón en el recuadro. Lea bien todo el enunciado antes de empezar. Para aprobar se requiere un mínimo de 60 puntos (60 puntos = 4) con al menos 15 puntos entre los ejercicios 1 y 2. Este enunciado debe ser entregado junto con el parcial si quiere una copia del mismo puede bajarla del grupo de la materia. En el ejercicio 3 elija 2 de los 4 ejercicios y resuelva única y exclusivamente 2 ejercicios. **Si tiene dudas o cosultas levante la mano**, está prohíbido hablar desde el lugar, fumar o cualquier actividad que pueda molestar a los demás. El criterio de corrección de este examen estará disponible en forma pública en el grupo de la materia.

"The question of whether a computer can think is no more interesting than the question of whether a submarine can swim." (E. Dijkstra)

#	1	2	3.1	3.2	4	5	6	7	Entrega Hojas:
Corr									Total:
Puntos	/15	/15	/10	/10	/15	/15	/10	/10	/100

Nombre:
Padrón:
Corregido por:

1) Se tiene un archivo con información sobre visitas a páginas web de la forma: (URL, visitas, fecha). Existe solo un registro por día para cada URL. Se quiere generar un archivo que, por cada URL, indique cuál fue la fecha en la que tuvo mas visitas y la cantidad de visitas.

Programar lo pedido en Map Reduce usando agregación para minimizar la cantidad de datos que deben transferirse en la red.

Atención: La resolución es muy simple, trivial, asi que es fundamental resolver la agregación para el puntaje completo. (**) (15 pts)

2) Se tiene un archivo que representa una red social con billones de links entre usuarios de la forma (id1,id2). Se quieren encontrar todos los triángulos de tipo (id1,id2,id3) tales que existen (id1,id2),(id2,id3),(id1,id3). Explicar, sin programar, que algoritmo usaría para listar todos los triángulos que existen en la red social usando Map Reduce. En concreto se pide para el siguiente fragmento del archivo:

(A,B) (B,C) (A,C) (A,D) (C,D) (D,E) (D,F) (E,F) (A,E)

- Qué recibe el metodo Map, qué hace y qué emite.

- Qué recibe el metodo Reduce, qué hace y qué emite.

Si lo hace en dos o mas iteraciones entonces explique los puntos anteriores por cada iteración. (****) (15pts)

(opcional x puntos extras) ¿De qué forma podría usarse el contar la cantidad de triángulos para saber si un determinado grafo es o no una red social? Justifique. (*****) (5 pts)

3) Resolver 2 (dos) y solo 2 de los siguientes ejercicios a elección (si resuelve mas de 2 el ejercicio vale 0 puntos, sin excepciones). En cada caso indicar V o F justificando adecuadamente sus respuestas. Si no justifica vale 0 puntos sin excepciones.

a) En una colección de 1300 documentos el b optimo para un término que aparece en 900 de ellos es 1. (**) (10 pts)

b) Si para una cierta consulta q d1 es mas relevante que d2 entonces luego de aplicar LSI d1 seguirá siendo mas relevante que d2. (***) (10 pts)

c) Si la probabilidad de las distancias 8 a 15 es similar a 1/256 entonces es buena idea representarlas usando código Delta. (**) (10 pts)

d) En los casos en los que hay mas términos que documentos el código unario es óptimo. (*) (10 pts)

4) Comprimir con PPMC de orden máximo 2 (dos) el siguiente archivo: ABABAA

Trabajar con el set de caracteres A,B

Indicar en cada paso el estado de los modelos y el archivo final comprimido en binario. Importante: Si deja expresado el archivo final como las probabilidades vale 0 puntos (sin excepciones) (***) (15 pts)

5) Se quiere aplicar LSH a un conjunto de documentos para encontrar los pares de documentos mas similares. Queremos que si J(D1,D2)>=0.7 entonces la probabilidad de que D1 y D2 sean candidatos sea >= 0.9 y queremos que si J(D1,D2) <= 0.5 entonces la probabilidad de que sean candidatos sea <= a 0.3. Indique cuántas funciones minhash usaría y que combinación de AND y OR usaría para lograr lo pedido. (****) (15 pts)

6) Si tenemos mas de un dispositivo físico disponible (discos) para hacer el sort externo de un archivo de varios gigabytes de longitud. ¿Qué algoritmo usaría para realizar el sort? Justifique. (*) (10 pts)

7) Encontrar el m minimo y los parámetros a y b de forma tal que la función de hashing ax + b mod m sea perfecta para las siguientes claves: 1,3,5,12 (***) (10 pts)