Organización de Datos (75.06) Primer Cuatrimestre de 2015. Examen parcial, primera oportunidad. [2015_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 consultas levante la mano**, está prohíbido hablar desde el lugar, fumar, escuchar a Arjona 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.

"Anagrams never lie, reveals a renaming." ***(Donald Holmes)***

#	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) Dada una colección de documentos queremos encontrar frases de 1 , 2 o 3 palabras que sean anagramas de otras. Por ejemplo: ("Postmaster", "Stamp store") o ("A telescope" , "To see Place") o ("The cockroach", "cook catch her"). Esta tarea implica una combinatoria muy difícil por lo que se decide usar Map-Reduce para paralelizarla. Usando Map-Reduce programar la solución a este problema listando todos los pares de anagramas entre frases de 1, 2 o 3 palabras. Como puede verse en los ejemplos se ignoran mayúsculas y minúsculas y los espacios en blanco, puntuación, etc. Suponer que existe la función word_tokenizer que recibe un texto y devuelve un vector de palabras ya convertidas a minúsculas y sin puntuación. (****) (15 pts)

2) El archivo Amigos.txt contiene informacion sobre las amistades entre usuarios en una red social. Cada relacion de amistad esta representada por dos registros. Por ejemplo si los usuarios A y B son amigos, el archivo contiene los registros (A,B) y (B,A). Se pide realizar un programa en Pig que, utilizando el archivo Amigos.txt, obtenga cual es el usuario con mayor cantidad de amigos. (**) (10 puntos).

Utilizando el archivo Usuarios.txt que contiene (user_id, user_name) incluir en el resultado el nombre del usuario con mas amistades. (**) (5pts)

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) Utilizando una funcion de hashing con resolución de colisiones Hopscotch de distancia màxima 4 me aseguro insertar como minimo 4 sinonimos. (*) (10 pts)

b) Dado que una funcion de hashing criptografica tiene pocas colisiones es ideal para asegurarnos buscar claves rapidamente y su uso es recomendable en la mayoria de los casos. (*) (10 pts)

c) Un compresor estadistico estático comprime siempre mejor que un compresor estadístico dinámico. (***) (10 pts)

d) LZW siempre comprime mejor que LZ78. (**) (10 pts)

4) Tomando bloques de 5 caracteres comprimir el siguiente archivo usando block-sorting, MTF y Huffman dinámico. Considere que solo son posibles los caracteres A,B,C y D. Se pide la salida en BINARIO completa del archivo comprimido. (***) (15 pts)

"ABABACCCCC"

5) Se tiene un set de datos sobre autos en donde cada columna mide en una escala de 0 a 1 diferentes datos de performance. En total tenemos 52 mediciones por cada vehículo. Se usa la SVD para descomponer la matriz de datos.

a) Explique de que forma podemos estimar a cuántas dimensiones podríamos reducir el set de datos para mantener un 99.5% de la información en el mismo. (5 pts) (*)
b) Suponiendo que el punto "a" nos indique que podemos usar 5 dimensiones. ¿De qué forma reducimos el set de datos a 5 dimensiones? (5 pts) (*)
c) Explique que representan las primeras "k" columnas de la matriz "V" (5 pts) (***)

6) Usamos LSH para encontrar documentos parecidos a un documento consulta. Desafortunadamente nuestra función LSH no está encontrando varios documentos que son similares a los buscados y esto es importante para nuestro problema.Indique diferentes formas de solucionar este problema. (**) (10 pts)

7) Es facil construir casos en los cuáles el código unario sea mejor que Gamma o Delta y es fácil construir casos en los cuales Delta sea mejor que Gamma y Unario. Le pedimos que explique que características debe tener una colección de documentos para que Gamma sea mejor que Delta y Unario.No alcanza con decir que debe haber mayoría de distancias en donde Gamma sea mas corto que Delta o Unario, eso ya lo sabemos. (****) (10 pts)

¿8099 es primo?