Organización de Datos 75.06. Primer Cuatrimestre de 2018. Examen parcial, segunda oportunidad:

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 20 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 consulta levante la mano, está prohibido hablar desde el lugar, fumar o cualquier actividad que pueda molestar a los demás. El criterio de corrección de este examen está disponible en forma pública en el grupo de la materia.

"Peace has cost you your strength! Victory has defeated you!" Bane, The Dark Knight Rises

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

Nombre:
Padrón:
Corregido por:

El GCPD (Gotham City Police Dept) recolecta la información de casos policiales que acontecen en Ciudad Gótica. Esta información se encuentra guardada en un archivo con el siguiente formato: **(fecha, id_caso, descripción, estado_caso, categoría, latitud, longitud).**

Los posibles estados que puede tener un caso son 1: caso abierto, 2: caso resuelto, 3: cerrado sin resolución. Las fechas se encuentran en el formato YYYY-MM-DD.

Por otro lado el comisionado Gordon guarda un registro detallado sobre en cuáles casos fue activada la batiseñal para pedir ayuda del vigilante, Batman. Esta información se encuentra en un archivo con el siguiente formato **(id_caso, respuesta)**, siendo campo respuesta si la señal tuvo una respuesta positiva (1) o negativa (0) de parte de él.

El sector encargado de las estadísticas oficiales del GCPD quiere analizar las siguientes situaciones:

a) Las categorías que hayan incrementado su tasa de resolución en al menos un 10% en el último trimestre, con respecto al trimestre anterior.

b) Tasa de participación de Batman por categoría, para los delitos contra la propiedad (que enmarcan las categorías **incendio intencional**, **robo**, **hurto**, y **robo de vehículos**)

1) Resolver ambas consultas, utilizando **Apache Spark** (***) (15 pts)

2) Resolver ambas consultas utilizando **Pandas**. (****) (15 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) Si hasheamos strings de longitud fija 6 caracteres en donde cada caracter puede ser cualquiera de las 26 letras del alfabeto de forma uniforme necesitamos que una función de hashing genere un número de al menos 32 bits para que la probabilidad de que dos strings colisionen sea menor a 0.1. (****) (10 pts)

b) Si f es una función de hashing genérica y g es una función de hashing criptográfica entonces la composición de funciones f o g es una función criptográfica. (**) (10 pts)

c) Si tenemos una tabla de hash de 512 posiciones a la cual accedemos usando una función de hashing que devuelve un número de 64 bits y queremos aumentar el tamaño de la tabla a 1024 posiciones será necesario re-hashear todos y cada uno de los elementos presentes en la tabla original. (****) (10 pts)

d) Si queremos hashear textos de mucha longitud, por ejemplo libros del proyecto Gutemberg minimizando la probabilidad de que dos textos diferentes generen el mismo hash es necesario usar una función de hashing criptográfica como por ejemplo SHA-256. (**) (10 pts)

4) Sean los siguientes vectores en 5 dimensiones: v1 = [4 4 -5 -2 3] ; v2= [-3 -2 -4 5 0]; v3=[3 2 -1 -2 1].

Y sean los siguientes 6 hiperplanos aleatorios: r1=[1 1 1 1 -1]; r2=[-1 1 1 -1 -1]; r3=[1 -1 -1 -1 -1]; r4=[1 -1 -1 -1 1]; r5=[1 -1 -1 -1 1]; r6=[-1 1 1 1 1].

Se pide estimar usando LSH cuál es el par de vectores más similares entre sí, mostrando el análisis realizado. (*) 10ptos

5) Dados los siguientes documentos:

D1: JACINTO JAZMIN JOCOSO

D2: JARRON JAZMIN

D3: JARRO JAPONES

D4: JEQUE JOCOSO JACINTO JEQUE

D5: JARRO JARRON JAZMIN

Se pide: a) Construir un índice invertido utilizando front coding parcial (n=3), utilizando códigos gamma para la codificación de punteros.. Indicar paso a paso cómo se construye el índice, y como quedan las estructuras resultantes.
b) Resolver la consulta "JEQUE" explicando cómo se utiliza para esto las estructuras antes creadas. (***) (15 pts)

6) La siguiente es la salida de un compresor LZ78:

A B (256) (258) (257) (260) (260)

Recuperar el archivo original, detallar cada paso. Cuanto ocupa el archivo original y cuanto el comprimido? (***) (10pts)

7) En base a los datos utilizados para los puntos 1 y 2, se pide realizar una única visualización que muestre la tasa de participación de Batman por categoría a lo largo del tiempo. La visualización tiene que comunicar efectivamente la **influencia de Batman en cada categoría** en la resolución de casos, y su **aporte en la disminución del crimen** en Ciudad Gótica. (****) (15 pts)