Organización de Datos 75.06. Segundo Cuatrimestre de 2018. Examen parcial, primera oportunidad: Criterio

1- Mas alla de que la resolución haga lo pedido, se evalua la calidad de la solución presentada. Map innecesarios descuento de 3 puntos.
Join sin claves compatibles descuento de 5 puntos. No filtrar lo antes posible descuento de al menos 3 puntos.

2- a) El item vale 7 puntos. Luego de filtrar por fecha, es necesario realizar un inner join entre eventos y locación por id_evento. Luego realizar un group by por sede y realizar la agregación de cantidad de espectadores. Si muestran los datos de la locación y no la sede se aplica un descuento de 5 puntos. Si hacen doble group by también aplica descuento, ya que esto no es necesario.
b) El item vale 8 puntos. Esencialmente es necesario realizar un inner join, para luego poder evaluar filtrando si la cantidad de espectadores es mayor a la que tienen la locación.
Para ello es necesario adicionar una columna en eventos que calcule la capacidad máxima que puede tener una determinada locación para poder evaluar. Descuento de 5 puntos si solamente usan capacidad_extendida.
Sobre los valores filtrados, es necesario realizar luego otro inner join con categorias_deportivas.csv.
De no retornarse el formato final pedido (nombre, categoria_deportiva), descuento de 2 puntos.

3- V/F: Respuestas sin justificación valen cero.

a) Sea un archivo que contiene cinco millones de dígitos de Pi, no se puede saber si es random porque K(X) es intractable.
Respuesta: Falso. Este archivo no es random porque si bien K(X) es incomputable aún así sabemos que hay un programa que ocupa menos de lo que ocupan cinco millones de dígitos (bytes) y en consecuencia K(X) tiene que ser menor que |X| (con lo cual no sería random)
b) Si un archivo es random entonces la entropía de Shannon es máxima.
Respuesta: Verdadero, aunque la inversa no vale. Es verdadero porque si la entropía no fuera máxima entonces hay símbolos que se pueden codificar en una menor longitud.
c) Las tablas de frecuencias de los compresores dinámicos de orden 3 o superior pueden ocupar tanto espacio que la compresión termina siendo ineficiente.
Respuesta: Falso. Los compresores dinámicos no almacenan sus tablas de frecuencias sino que las calculan en el momento.
d) Solo con compresores aritméticos podemos alcanzar la longitud ideal indicada por la entropía ya que permiten codificar un mensaje en cantidades no enteras de bits
Respuesta: Falso. Comparar el nivel de compresión con la entropía solo tiene sentido en el universo de los compresores estadísticos, pero hay otros compresores que no se basen en la probabilidad de ocurrencia de cada carácter que podrían mejorar un aritmético en nivel de compresión

4- Un ejemplo de solución es, como el espacio de direcciones de la función de hashing es 0..3. Armar dos tablas de la siguiente forma (usando h1 y h2 para la tabla 1, y usando h3 y h4 para la tabla 2).

Tabla 1

0	{14}
1	{12}
2	{22,10}
3	{22,12}

Tabla 2

0	{22,10,12}
1	{14,12}
2	{10}
3	{22,14}

Cálculos:
h1(22)=3, h2(22)=2
h1(14)=0, h2(14)=0
h1(10)=2, h2(10)=2
h1(12)=1, h2(12)=3

h3(22)=3, h4(22)=0
h3(14)=1, h4(14)=3
h3(10)=0, h4(10)=2
h3(12)=0, h4(12)=1

Existen distintas alternativas de solución que se considerarán correctas. Se evalua el entendimiento del tema, y que la estructura planteada en el punto a sea utilizada correctamente luego para resolver el punto b.
b) Tomando como resolución del punto a la anterior…. Si nuestro query es el 16, entonces primero buscamos la intersección entre h1(16)=2 y h2(16)=0. Con eso recuperamos de la primer tabla el {22,10} y el {14}, cuya intersección es nula. Por otro lado, en la 2da tabla buscamos la intersección con h3(16)=1 y h4(16)=1 entonces obtenemos los documentos {14,12}. Haciendo el OR entre { } y {14,12} se obtiene {14,12} con lo cual debemos comparar contra esos elementos. Si plantearon otra estructura de tablas se evaluará en función de como se haya resuelto el punto a.
c) Simplemente aumentando la cantidad de tablas a utilizar reducimos la cantidad de falsos negativos. Para reducir cantidad de falsos positivos aumentamos la cantidad de funciones de Hashing.

5- Cálculo incorrecto de IDF descuento de 3 puntos.
Cálculo incorrecto de TF.IDF descuento de 7 puntos.
Dar a D3 como último documento recuperado descuento de 3 puntos.
Descuento de un punto y medio por cada evaluador mal calculado.

6- a. Dos formas naturales de hacerlo es por el método del "codo", o bien utilizando la energía acumulada y si supera cierto umbral entonces podemos probar con ese valor de k.
b. Sí, claro. Tomando el valor de K igual al rango de la matriz tenemos la SVD compacta y en ese caso podríamos ver si las matrices U, Sigma y V ocupan menos que la matriz original.
c. Tres opciones: Usando las k columnas de U (sin escalarlas por Sigma), usando las k columnas de U escaladas por sigma o bien multiplicando V^t por la matriz A^t (la segunda y tercer opción arrojan el mismo resultado)
d. Al tener todas las imágenes del set en K dimensiones, lo que hacemos es proyectar la nueva imágen a K dimensiones y como estan vectorizadas podemos calcular la distancia euclídea contra las imágenes y ver cuál es la más cercana (hay que tener el cuidado aquí que pueden haber muchas imágenes para calcular la distancia).

7- Más allá de que se pueden proponer diversos enfoques para el punto, es necesario plantear algún tipo de visualización que se pueda mostrar una serie temporal para los días del evento.

El enfoque más simple partiendo de los datos sería mostrar el crecimiento que puede haber tenido una categoria deportiva en relación al público durante el certamen, a partir de line plot de y axes, para aquellas categorías nuevas. Podría aplicar también un stacked área plot. Se evaluará la

creatividad de la solución propuesta por los alumnos en este punto.
Descuento de puntaje si la visualización no tiene título, ni ejes rotulados. Tambien aplican descuentos si la propuesta carece de originalidad, explicaciones o justificación sobre las decisiones de diseño tomadas.