**Organización de Datos 75.06**. Segundo Cuatrimestre de 2018. Examen parcial, primera 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.

*"You're strong. But I could snap my fingers, and you'd all cease to exist.".- Thanos, Avengers: Infinity War*

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

**Nombre**:
Padrón:
Corregido por:

1) (***) Tenemos información sobre recetas en 3 RDD de Spark.

**Recetas**: (id_receta, nombre, tiempo_preparación, dificultad)

**Ingredientes**: (id_ingrediente, nombre)

**Ingredientes por Receta**: (id_receta, id_ingrediente, cantidad)

Se pide:

a) Obtener el nombre de todas las recetas que tengan Cordero. (7 puntos)

b) Calcular la cantidad total de cada ingrediente si queremos hacer todas las recetas con Cordero que sean fáciles. (8 puntos)

2) (***) Dada la exitosa convocatoria de los Juegos Olímpicos de la Juventud por parte del público, sus organizadores realizan distintos análisis para planificar las jornadas finales del certamen. Es por ello que cuentan con información en los siguientes archivos csv:

**eventos.csv** (id_evento, fecha, id_locacion, nombre_evento, id_categoria_deportiva, cantidad_espectadores)

**locacion.csv** (id_locacion, nombre, capacidad, capacidad_extendida, sede, latitud, longitud)

**categorias_deportivas.csv** (id_categoria_deportiva, nombre, año_de_adopcion)

El primer archivo cuenta con información de los eventos, indicando la fecha (en formato "YYYY-mm-dd hh:mm:ss"), el lugar donde ocurrió (id_locacion) y la cantidad de espectadores que asistieron. Además se aporta información sobre la categoría deportiva a la cual pertenece el evento.

Por otro lado se tiene información sobre las distintas locaciones en la sedes del certamen en las que ocurrieron los eventos. Contamos con información de su capacidad total de espectadores así como de su capacidad extendida (cuantos asientos extras se pueden brindar sobre la capacidad de la locación).

Se desea obtener:

a) Nombre de la sede que acumuló la mayor cantidad de espectadores en eventos durante el certamen del 14 al 15 de octubre inclusive. Esto es de vital importancia para distribuir el merchandising oficial del evento, para las fechas finales. (7 pts)

b) Nombre del evento y nombre de la categoría deportiva de aquellos eventos cuya cantidad de espectadores superó **la capacidad de la locación, más allá de la capacidad extendida**. Esto es de vital importancia para detectar problemas de seguridad o si es necesario realizar algún cambio de locación. (8 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) Sea un archivo que contiene cinco millones de dígitos de Pi, no se puede saber si es random porque K(X) es intractable. (*) (10 pts)

b) Para poder suponer que un archivo es random debe verificarse que la entropía de Shannon sea máxima (*) (10 pts)

c) Las tablas de frecuencias de los compresores dinámicos de orden 3 o superior pueden ocupar tanto espacio que la compresión termina siendo ineficiente. (*) (10 pts)

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 (*) (10 pts)

4) Desafortunadamente, tenemos un set de datos con muchos puntos y necesitamos utilizar LSH para buscar los puntos más cercanos. Contamos con el siguiente set: {22,14,10,12} y las siguientes 4 funciones de hashing: h1(x)=(3x mod 7) mod 4, h2(x) = (2x mod 7) mod 4, h3(x)=((2x+1) mod 7) mod 4 y h4(x) = ((x+3) mod 7) mod 4. Se pide:
a. Usando b=2 y r=2, indique cómo quedan las tablas
b. ¿Cuáles puntos deberíamos comparar si nuestro query es el {16}? Explique
c. ¿Qué podríamos hacer para reducir la cantidad de falsos negativos? ¿Y si quiséramos reducir la cantidad de falsos positivos? (***) (15pts)

5) (**) Se tienen los siguientes documentos:

D1: CORDERO SAL PIMIENTA ROMERO

D2: CERDO CORDERO SAL CORDERO

D3: SAL CERDO LIMON

D4: CORDERO ENTRAÑA

D5: PIMIENTA PAPA CORDERO PIMIENTA

D6: CORDERO CORDERO CORDERO CORDERO

Dada la consulta "CORDERO PIMIENTA" dar el resultado de la consulta rankeada utilizando TF.IDF. (10 pts)

Considerando como relevante los documentos que no tengan otra carne que no sea CORDERO, calcular la Precisión, Recall y F1 Score. (5 pts)

6) Se tiene una matriz muy grande donde cada fila representa una imágen de una cara. Se quiere aplicar algún algoritmo de reconocimiento facial utilizando la SVD:
a. ¿Cómo podemos determinar el valor de k (cant. de dimensiones a utilizar)?. Justifique
b. ¿Se podría reducir el espacio que ocupa la matriz sin perder información?
c. Una vez obtenido *k*, ¿Cómo podemos reducir los puntos originales a *k* dimensiones?
d. Si ahora quisiera reconocer una imágen, ¿Cómo podría usar la SVD para ello? (***) (10pts)

7) El COI desea evaluar la aceptación de las nuevas categorías deportivas que se sumaron en el año 2018 a los Juegos Olímpicos de la Juventud. Para ello es necesario que nuestra área de análisis de datos prepare una visualización que muestre a lo largo del tiempo de duración del certamen como fue evolucionando la cantidad de público que han tenido estas nuevas categorías

Para desarrollar el punto debe partir como base de la información que cuenta en el punto 2, ampliando con otras posibles fuentes de datos, el contenido de la misma. (***) (10 pts)