Organización de Datos 75.06. Primer Cuatrimestre de 2020. Examen parcial, primera oportunidad:

Importante: Lea bien todo el enunciado antes de empezar. Para aprobar se requiere un mínimo de 60 puntos (60 puntos = 4) con al menos 30 puntos entre los ejercicios 1 y 2. Si tiene dudas o consulta estaremos disponibles vía meet, pero tengan en cuenta que solo se contestarán dudas de enunciado, y no deben compartir por esa vía nada relacionado con la resolución. Está prohibido realizar 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.

Throughout human history, we have been dependent on machines to survive. Fate, it seems, is not without a sense of irony.
Morpheus - The Matrix

#	1	2	3	4	Entrega Hojas:
Corrección					Total:
Puntos	/25	/25	/30	/20	/100

Nombre:
Padrón:
Corregido por:

1) Dado los acontecimientos en USA, deseamos obtener datos que nos den mayor información sobre las muertes de personas de raza negra por parte de oficiales de policía.

Para ello, tenemos un csv con información sobre las muertes por parte de oficiales de policía en USA desde 2015 hasta 2017:

`(name, date, race, city, state)`

Y otro csv con información sobre el porcentaje de cada raza en las ciudades de USA:

`(state, city, share_white, share_black, share_native_american, share_asian, share_hispanic)`

Se pide:

a) Obtener el estado con mayor porcentaje de muertes de personas de raza negra teniendo en cuenta la cantidad total de muertes por parte de oficiales en ese estado. (10 pts)

b) Obtener los 10 estados con mayor diferencia entre el porcentaje de muertes y el porcentaje de gente de raza negra en ese estado. Para ello, considerar el porcentaje de raza de un estado como el promedio de los valores de sus ciudades. Por ejemplo si en Texas el porcentaje de muertes de personas de raza negra por parte de la policía es del 36% y el promedio de share_black para Texas es 24% la diferencia es 0.12. (15 pts)

Resolver ambos puntos usando la API de RDDs de PySpark. `

2) Un importante servicio de monitoreo de aplicaciones cloud, sufrió un importante incidente en las últimas semanas por lo cual está conduciendo una investigación. Nuestro equipo participa en el análisis de forma externa para lo cual nos provee los siguientes dataframes en csv:

- `metrics.csv` describiendo información de métricas obtenidas. El mismo tiene el formato `('client_id', 'integration_id' ,'metric', 'timestamp', 'value')` donde la columna `'metric'` indica el nombre de una métrica que se registra en un cierto momento (`'timestamp'`) cuyo valor se guarda en la columna `'value'`.

- `clients.csv` con información sobre el cliente, con el formato `('client_id, 'account_number', 'name')` donde `'account_number'` indica el número de cuenta de cliente y 'name' el nombre del mismo.

Como miembro de nuestro equipo es necesario que obtengas las siguientes métricas para aportar insights a nuestro análisis:

a) Calcular el valor promedio de las metricas `'aws.vpc.network_out', 'aws.vpc.network_in', 'aws.vpc.network_rate'` para los clientes de `'account_number'` mayor al `'25679247'` devolviendo el resultado en el siguiente formato `(account_number, 'aws.vpc.network_out_mean', 'aws.vpc.network_in_mean', 'aws.vpc.network_rate_mean')`(15pts)

b) Calcular la diferencia entre el valor promedio obtenido en cada métrica por cada cliente, y el valor promedio general de esa métrica (sin utilizar group by) (10pts)

(****)

3) LSH, algo de teoría, algo de práctica.

a) "Algo de Teoría" (20 puntos)

Tenemos una colección de millones de archivos binarios y queremos encontrar rápidamente archivos parecidos con el propósito de detectar malware. A fin de resolver este problema se nos plantea la idea de usar LSH para estos archivos. A modo de guía le pedimos que piense y resuelva los siguientes problemas:

a.1) ¿Cuál sería la función de minhash a utilizar? compruebe que cumple las propiedades necesarias para ser un minhash.
a.2) En base a la función de minhash planteada indique de qué forma se realizaría la amplificación de la misma usando b=2 y r=16
a.3) En base a los dos puntos anteriores indique qué pre-procesamiento debería realizar sobre los archivos para poder encontrar los similares a un archivo dado.
a.4) Considerando el punto anterior una vez realizado el pre-procesamiento indique de qué forma encontraría los archivos candidatos a ser similares a un nuevo archivo.

b) "Algo de Práctica" (10 puntos)

Los siguientes vectores representan calificaciones de canciones del 1 al 5, donde 0 significa que ese usuario no escuchó la canción:
v1 = [1 5 3 2] v2 = [0 0 1 2] v3 = [4 4 5 0] v4 = [5 1 0 1]
Utilizando la técnica de los hiperplanos, con b = 1, r = 4, hallar hiperplanos correspondientes para que el más similar a v1 sea v3 (y ningún otro).

4) A partir de la siguiente estructura de un índice Invertido:

=	≠	Char	Doc
0	5	0	0
4	6	5	10
2	8	11	33
0	9	19	38
4	2	28	55
2	7	30	69
0	10	37	76

`madreigueragistradomaleficiotanantialmanifiesto`

`10100100110110101001001010101111101011111010101101010100100100100010001110100101011010100100`

`(las posiciones en los documentos están numeradas desde la posición 1, y todo está codificado en gamma)`

Armado con la finalidad de poder resolver consultas por proximidad, se pide extraer la información de los documentos y resolver la consulta "**Madriguera Maleficio** " utilizando **TF-IDF**. Detallar cada paso realizado (****) (20pts)