Organización de Datos 75.06. Segundo Cuatrimestre de 2017. 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.

"Your time will come. You will face the same evil, and you will defeat it.", (The Lord of the Rings. The Fellowship of the Ring)

#	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) A partir de la plataforma online (e-shop) de los países en los que opera, Nintendo tiene información **de ventas de videojuegos diarias digitales por país** en el siguiente RDD: (id_videojuego, codigo_pais, fecha, visitas_diarias, total_ventas_diarias).

Por otro lado se tienen otro RDD que tiene información de todos los videojuegos que se venden en su plataforma con el siguiente formato (id_videojuego, titulo, rating_pegi, rating_esbr).
Tener en cuenta que un mismo videojuego se puede vender en distintos países y esos nos permitirá obtener métricas a nivel global.

Con esta información escribir un programa en pySpark que permita:
a) Obtener el videojuego con más ventas digitales globales (es decir en todos los países) en un RDD con el siguiente formato: (id_videojuego, titulo, total), siendo total la cantidad total de ventas digitales globales
b) Para el videojuego con mas ventas, obtener cual es el país para el cual ser registra una mayor tasa de conversión (es decir, mayor total_ventas_diarias / visitas_diarias) (**) (15 pts)

2) La Agencia Nacional de Estadísticas de Buenos Aires recolecta información de nacimientos cuando los padres registran a sus hijos en el registro civil a partir de una encuesta. Esa información se encuentra disponible para su análisis en un csv con el siguiente formato (dia_nacimiento, mes_nacimiento, anio_nacimiento, peso_al_nacer, longitud_al_nacer, id_hospital, tipo_parto), donde el tipo de parto 1 es natural y 2 es cesárea.
Por otro lado la agencia cuenta con información histórica de los hospitales en otro csv con siguiente formato (id_hospital, dirección, promedio_nacimientos_mensual).

Se pide usar Pandas para:
a) Calcular la cantidad de nacimientos para cada uno de los hospitales para el mes de Octubre de 2017 e indicar aquellos hospitales que superan el promedio de nacimientos mensuales.
b) Comparando el mes de Octubre de 2017 indicar programáticamente si se incremento el % de cesáreas con respecto a ese mes del año 2016. (****) (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) Usamos KNN para predecir la cantidad de dinero que los clientes van a gastar en un sitio de e-commerce. El algoritmo predice valores muy similares para todos los clientes y eso no tiene sentido. Podemos afirmar entonces que el valor de "k" que estamos usando es muy alto (**) (10 pts)

b) Si optimizando k no obtenemos buenos resultados podemos afirmar que KNN no es un buen algoritmo para nuestro problema (**) (10 pts)

c) Si tenemos muchas clases y usamos KNN como algoritmo de clasificación, necesitaremos muchos datos para que el algoritmo funcione bien. (**) (10 pts)

D) Para puntos en dos dimensiones podemos lograr que el orden de complejidad de KNN para clasificar un punto sea aproximadamente O(log N). (**) (10 pts)

4) Dados los siguientes documentos:
D1: google gmail hangout
D2: yahoo gmail hotmail outlook
D3: google yahoo excite lycos altavista
D4: hotmail outlook google
D5: android mobile gmail google
a) Construir el índice invertido para esa colección. Detallar como se construye el índice y las estructuras resultantes (10pts)
b) Resolver la consulta "google AND gmail" explicando paso a paso su resolución. (**) (5pts)

5) El archivo (en bits) 00100010 10000000 es el resultado de utilizar **PPMC de Orden 1**. Asumiendo que los únicos caracteres posibles son A, B y C y que el compresor asigna los intervalos ordenando los caracteres alfabéticamente (comenzando por la A en el intervalo inferior) se pide recuperar el archivo original (***) 15ptos.

6) Supongamos que asignamos a cada letra del alfabeto un numero de la forma A=1, B=2, C=3, etc... Proponemos como función de hashing sumar el valor correspondiente a cada carácter del string y luego tomar el modulo con un cierto número primo p. Analizar la función propuesta indicando: a) Cantidad de colisiones b) Facilidad de encontrar sinónimos c) Eficiencia d) Efecto avalancha (***) (10 pts)

7) Usando LSH indicar si una construcción de 2 ANDs, 3 ORs, 2 ANDs y 3 ORs es equivalente a una construcción de 3 ORs, 2 ANDs, 3 ORs y 2 ANDs. (***) (10 pts)