Organización de Datos 75.06. Segundo Cuatrimestre de 2015. 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 cosultas levante la mano, esta prohíbido hablar desde el lugar, fumar o cualquier actividad que pueda molestar a los demás. El criterio de corrección de este examen esta disponible en forma pública en el grupo de la materia.

"Toto, I've a feeling we're not in Kansas anymore." (The Wizard of Oz)

#	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:

1) Se tiene un RDD con las coordenadas de rectángulos de la forma (x1,x2,y1,y2). Se pide programar en PySpark un programa que encuentre el rectángulo de superficie mínima que contiene al punto (w,z) (*?) (15 pts)

2) Dados los archivos:

Songs (SongID, SongTitle, Genre);
WhoSangIt (SongID, SingerID)

Realizar un programa en PIG que liste para cada canción que haya sido interpretada por tres o mas cantantes distintos, el nombre de la canción, y su genero. (***) (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 queremos usar KNN para un cierto k fijo podemos eliminar del set de entrenamiento todos los puntos que tienen a sus k vecinos pertenecientes a la misma clase. (**) (10 pts)

b) Tenemos una matriz con las calificaciones de películas x usuario. La matriz es de 50000 películas por 150000 usuarios. La dimensionalidad intrínseca de este set de datos seguramente es menor a 150000. (***) (10 pts)

c) En un boxplot podemos identificar si los datos presentan el fenómeno conocido como "skewing". (**) (10 pts)

d) Un plot de tipo "coordenadas paralelas" perimite visualizar datos en múltiples dimensiones mientras que un plot de burbujas solo permite visualizar dos dimensiones. (*) (10 pts)

4) Dados los siguientes puntos en 4 dimensiones: p1:(3,2,0,3) p2: (0,1,5,4) p3: (2,0,6,3) p4:(2,2,1,2). Aplicando la SVD obtenermos la siguiente descomposición (por filas):

U= [-0.2942, 0.7870, -0.0268, -0.5417; -0.6193,-0.2611,0.7362,-0.0794; -0.6715, -0.3012, -0.6759, -0.0395; -0.2811, 0.4709,0.0207, 0.8359]

S = [10.0500, 0, 0, 0; 0, 4.6033 , 0,0;0, 0 1.8316,0; 0,0, 0,0.6727]

V=[-0.2774, 0.5866,-0.7594,-0.0477;-0.1761, 0.4898,0.3953,0.7569;-0.7370,-0.5738, -0.1929, 0.3006; -0.5907, 0.2944 0.4795, -0.5784]

a) Graficar los puntos en dos dimensiones. (5 pts) (**)

b) Interpretar el resultado de la SVD en base a los datos. (5 pts) (***)

5) Tenemos la siguiente tabla representando el valor de 6 minhashes para tres documentos:

	D1	D2	D3
MH1	1	0	1
MH2	3	1	3
MH3	1	2	1
MH4	2	2	3
MH5	0	0	2
MH6	0	0	2

Se usa b=2 y r=3. Se decide usar 7 buckets para cada banda. Encontrar una <u>única</u> función de hashing perteneciente a una flia universal LSH(r1,r2,r3) de forma tal que en la primer banda solo D1 y D3 sean candidatos a ser similares pero en la segunda banda los tres documentos sean candidatos a ser similares. (****) (15 pts)

6) Dadas las siguientes frases:
"this is not fine"
"this is right"
Calcular la probabilidad de la frase "this is not right" en un modelo de bigramas usando Stupid Backoff como método de smoothing. (***) (10 pts)

7) Se tiene un archivo con 10 caracteres en total formado por tres caracteres distintos (ej: ABC). De todos los archivos posibles con estas características mostrar el archivo de máxima entropía que se pueda comprimir mejor usando LZ77. No es necesario comprimir el archivo. (***) (15 pts)