Organización de Datos 75.06. Primr Cuatrimestre de 2016. 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 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.

"If you want to be succesful perserverance is the key quality" (George Lucas)

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

Nombre:
Padrón:
Corregido por:

1) Una red social almacena el contenido de los chats entre sus usuarios en un RDD que tiene registros con el siguiente formato: (chat_id, user_id, nickname, text). Queremos saber cuál es el usuario (user_id) que mas preguntas hace en sus chats, contabilizamos una pregunta por cada caracter "?" que aparezca en el campo text. Programar en Spark un programa que identifique al usuario preguntón. (*) (15 pts)	2) Dados los archivos: ratings (movie_id, title, avg_rating) genres (movie_id, genre) Realizar un programa en PIG que liste los 5 géneros con mayor promedio de rating en sus películas. (**) (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 H es una familia de funciones de hashing universal definida por h(x,a) en donde "a" es un parámetro entonces la familia H2 definida h2(x,a)=h(x,a) mod p también es universal.(****) (10 pts)	b) Si sabemos que la dimensionalidad intrínseca de los datos es 1 entonces podemos aplicar la SVD y quedarnos con la primer columna de U para representar a los mismos. (******) (10 pts)	c) Si PPMC no comprime bien un archivo entonces es muy poco probable que Block Sorting lo comprima bien. (****) (10 pts)	d) El teorema de Johnson Lindenstraus nos da una forma óptima de proyectar un set de datos a una cantidad de dimensiones "k" (***) (10 pts)

4) Dar un archivo/string que, comprimido con un compresor artimético dinámico de orden 0 ocupe exactamente 23 bits. (15 pts) (****)	5) Usamos LSH para la distancia de Jaccard para comparar frases breves usando 4-shingles con 6 funciones de hashing que agrupamos en 3 construcciones OR de 2 construcciones AND cada una. Queremos obtener los strings que sean al menos 80% semejantes a "use the force". Describa detalladamente todos los pasos necesarios para encontrar las frases que cumplan con lo pedido. (****) (15 pts)

6) Dada la siguiente colección y considerando a cada línea como un documento. Se pide aplicar "the hashing trick" sabiendo que los vectores tienen 6 dimensiones. Luego calcular la semejanza todos contra todos para los 4 documentos de la colección usando el método del coseno. (***) (15 pts) alfa beta alfa alfa gamma alfa beta beta beta delta delta beta gamma gamma gamma gamma alfa gamma delta h(alfa) = 1 h(beta) = 5 h(gamma)=0 h(delta)=1	7) En 2014 American Airlines canceló el 1.5% de sus vuelos, Delta canceló el 1.8% y United el 2.2%. En 2015 American canceló el 1.7%, Delta el 2.2% y United el 3.4% de sus vuelos. Realice una visualización coherente y agradable para estos datos y explique la misma. (***) (10 pts)