Organización de Datos 75.06. Segundo Cuatrimestre de 2017. 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, esta 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.

"For he that strikes the first blow, if he strikes it hard enough, may need to strike no more.", (The Lord of the Rings. The Two Towers)

#	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) Se cuenta con un RDD con información sobre patentamientos de autos con la siguiente información (patente, marca, modelo, versión, tipo_vehiculo, provincia, fecha), donde tipo_vehiculo indica si la unidad patentada es auto, pickup, camión o moto.

Se pide generar un programa en pySpark que indique la marca y modelo del auto más patentado por tipo de vehículo en la provincia de Buenos Aires en el mes de Abril de 2017. (***) (15 pts)

2) Tenemos un dataframe con la información de distintas playlists armadas por usuarios con el formato (playlist, song_id, description). A su vez, contamos con un dataframe de canciones que contiene (song_id, singer, year, lenght, genres).

Se pide generar un programa en Pandas que indique para cada playlist cual es el cantante predominante (con mas canciones incluidas dentro de esa lista). (**) (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) Todo archivo con complejidad de kolmogorov baja tendrá una entropía de Shannon baja. (**) (10 pts)

b) Es imposible que un archivo comprimido mediante huffman estático de orden 1 iguale la máxima compresión dada por la entropía del mismo. (***) (10 pts)

c) Podemos determinar la longitud final del archivo comprimido utilizando huffman estático de orden 1, calculando la entropía y multiplicándola por la cantidad de caracteres del archivo (**) (10 pts)

D) Tenemos 2 archivos, uno con longitud pequeña y el otro muy grande que se comprimen utilizando huffman estático de orden 5. Si observamos que tienen la mismas tablas de frecuencias podemos afirmar que el cociente entre el tamaño del archivo sin comprimir y el tamaño del archivo comprimido será similar. (***) (10 pts)

4 Se tiene la siguiente colección de documentos:

1: cucharita chacarita cucaracha chacarita
2: cucharita cucurucho
3: cartucho cucharita cartucho cartucho
4: cucaracha mamaracha cucaracha cucaracha
5: chacarita cucharita cucharita cucharita cucaracha

Resolver la consulta "chacarita cucurucho " utilizando BM25 (K=2) y sin normalizar la longitud de los documentos (modificar el parámetro necesario en la fórmula para ajustarse a esto). (**) (15pts)

5) Descomprimir el siguiente archivo comprimido con LZ78/LZW:

B C 256 258 B A 260 262 260

Indicar cuanto se logro comprimir el archivo en este caso.
(***) (15 pts)

6) Si la probabilidad de que dos vectores colisionen usando un único hiperplano es mayor a 0.95.
a) ¿Cuál es el ángulo máximo entre los vectores?
b) De un ejemplo de un hiperplano para el cual dos vectores que están a la distancia indicada en el punto anterior no colisionen.
(10 pts ***)

7) Tenemos un total de 10.000 claves de 32 bytes c/u. Si usamos el esquema FKS y la primer tabla tiene 1000 posiciones. ¿Cuánto espacio necesitamos en total para almacenar las 10.000 claves?
(10 pts *)