Organización de Datos 75.06. Primer Cuatrimestre de 2019. 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.

"Ask me again in 10 years." - Tyrion Lannister, "The Imp", "The Hangman", "The Hand", Game of Thrones.

#	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) ACARA posee información histórica sobre la venta de autos 0km en la Argentina. Posee un RDD con información de cada modelo `(marca, modelo, motor, transmisión, origen)` y otro con la información de ventas `(marca, modelo, fecha, concesionario)`.

Se desea conocer, para los modelos de origen nacional, cuales son los modelos que ya se discontinuaron (un modelo discontinuado es aquel que no tuvo ventas en los últimos 12 meses), obteniendo un nuevo RDD con `(marca, modelo, total_vendido, fecha_inicio_venta, fecha_discontinuación)`, donde la fecha de discontinuación es la fecha cuando se vendió el último auto de ese modelo, ordenado ascendentemente por esta fecha. (***) (15pts)

Aclaración: Se puede asumir que el primer RDD tiene un único registro para cada Marca y Modelo.

2) Se tiene información diaria de la cotización de acciones en el NYSE en el archivo `nyse_daily.csv` en el siguiente formato `(symbol, date, open, measure_midday, measure_afternoon, close, volume)`. Por cada acción cuyo nombre está indicado en el campo `symbol`, tendremos una entrada por fecha con los valores `open`, `measure_midday`, `measure_afternoon`, y `close` indicando respectivamente a qué valor abrio la acción, cuál fue el valor que tuvo al mediodía, cual fue su valor que tuvo por la tarde y cual fue su valor al cierre del mercado. Asimismo en `volume` se indica el volumen operado ese día para esa acción.

Por otro lado se cuenta con el archivo `s&p500.csv` de formato `(symbol, company_name)` que indica aquellas acciones que deben ser consideradas para calcular el índice Standard & Poor's 500 (S&P 500).

Se pide calcular el valor diario del índice S&P 500, teniendo en cuenta que el mismo se calcula como el promedio del valor promedio de las mediciones que tuvo cada acción ese día (`open, measure_midday, measure_afternoon, close`), para las 500 acciones que se encuentran en el archivo `s&p500.csv`.

El resultado debe estar en un dataframe de la forma `(date, index_name, value)`. Por ejemplo, una entrada del mismo sería `('2019-03-24', 'SP500', '35.46')`.

Nota: A los efectos prácticos del ejercicio consideraremos como estadísticamente significativo calcular el promedio con esas pocas mediciones.
(***) (15pts)

3) Resolver 2 (dos) y solo 2 de los siguientes ejercicios a elección (si resuelve más 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) Utilizar SVD o PCA para reducir dimensiones de un set de datos es exactamente lo mismo, incluso a nivel de utilización de recursos. En consecuencia, no hay ningún motivo por el cual preferir usar un método u otro. (**)

b) Las proyecciones vistas que cumplen el lema JL son independientes de la distribución de los datos, mientras que SVD encuentra la proyección según ellos. (**)

c) Se puede utilizar reducción de dimensiones como mecanismo de feature engineering. (***)

d) Si tenemos un algoritmo que sabemos que performa en orden cuadrático para una gran dimensión y volumen de datos, no tenemos herramientas para poder utilizarlo en ese tipo de sets de datos. (**)

4) La empresa UBER tiene registrados los siguientes viajes:
Viaje1: A -> B -> D -> E
Viaje2: A -> C -> D -> E
Viaje3: A -> C -> D -> G
Se pide:

a) Explicar qué hace UBER para detectar viajes similares utilizando LSH Jaccard (3pts).
b) Teniendo las siguientes funciones de hashing h1(x) =(2x+1) mod 6, h2(x)=(3x+2) mod 6, h3(x)= (5x+1) mod 6, h4(x)= (x+1) mod 6, realizar las permutaciones correspondientes y encontrar los candidatos de viajes a ser similares utilizando r=2 ***y luego*** b=2 para el siguiente viaje (12 pts):
ViajeQ: C->D-> E

(***) (15 pts)

5) Comprimir el siguiente archivo utilizando LZHuf: **ABACBACACBABABAB.** Tomar 6 caracteres para la ventana y considerar un mínimo de 2 caracteres para las repeticiones (a los efectos del ejercicio está bien considerar ABC como únicos caracteres posibles) Pueden considerar también que la longitud máxima de match es 6. Indicar paso a paso cómo se procesa el archivo y como queda el archivo comprimido. (***) (15pts)

6) Un sitio de empleos desea utilizar la metodología Learning to Rank para rankear sus avisos de búsqueda de empleos.

Proponga una solución para el mismo, explicando qué factores tendría en cuenta y cómo funcionaria el mismo, con el mayor detalle posible.(****) (10 pts)

7) Con la información provista en el punto 1, realizar una visualización en la que pueda comparar **de forma diaria** las cotizaciones de las acciones de símbolo AMZN, GOOG, FB. Además indicar en la visualización para cada acción en cada dia, si cerro a la alza o a la baja (si el valor para ese dia con el que abrió la acción es mayor al de cierre se dice que cerró a la alza. En caso contrario, a la baja).
¿Cómo podría agregar el índice calculado en el punto 1 a la visualización para comparar el desempeño de esas acciones contra el índice S&P 500? (***) (10 ptos)