Organización de Datos 75.06. Primer Cuatrimestre de 2019. Examen parcial, segunda oportunidad: Resolución

1. Posible solución:
   Filter autos por origen nacional.
   Map con key (marca, modelo) en ambos RDD.
   Inner Join entre ambos.
   ReduceByKey para obtener la primer fecha donde se vendio y la última.
   Filter para chequear si la última fecha esta alejada en 12 meses a la actual.
   Map para quedarnos con la info pedida.
   Sort por la fecha.
   Map innecesarios, descuento de 3 puntos. Filter tardíos, descuento de 3 puntos. Joins con RDD sin clave descuento de al menos 5 puntos.

2. El ejercicio consiste en realizar un join entre los dos dataframes con how='inner' para solamente quedarnos con información de las 500 acciones del indice S&P 500. De esta forma contaremos en el dataframe resultante solamente con información de las acciones que nos interesan para el cálculo.
   Luego por cada fila tienen que realizar un promedio por 'fila' por las columnas open, measure_midday, measure_afternoon, close.
   Esto lo pueden realizando un groupby por dia y symbol y luego haciendo un apply para promediar esos valores generando una única fila que tiene el valor promedio diario para el symbol de la acción por dia.
   Es de vital importancia en el ejercicio entender cómo realizan el apply, dado que es importante también que mantengan valores para las operaciones siguientes de forma correcta y puede invalidar completamente el ejercicio.
   Luego se tendrá una estructura del tipo (date, symbol, daily_mean) sobre la cual hay que realizar un group by por fecha y realizar el promedio sobre los promedios diarios obtenidos.
   Para llegar al final pedido resultado final es necesario que fijen en el dataframe resultante una columna con el valor `'SP500'`.
   Se puede evaluar el tipo de join, mas alla de que se puede utilizar un inner siendo esto lo más correcto o un left o right teniendo a la izquierda o derecha respectivamente el dataframe de s&p500.csv. Otro tipo de joins sin considerar esto, descuento de 5 puntos.
   Mal manejo de índices luego de groupby descuento de 2 puntos.
   Si no llegan al formato final pedido agregando la columna al dataframe, descuento de 1 puntos.
   Errores en el apply serán a evaluar por el corrector hasta un máximo de 7 puntos.

3. a. Falso.
   b. Verdadero.
   c. Verdadero,
   d. Falso,

4. Para poder realizar la detección de viajes similares, Uber considera "shingles" de la forma A->B, B->D, D->E, E->H y así sucesivamente. Luego, con cada shingle, utilizan simplemente la distancia de Jaccard. Entonces es cuestión de armar la matriz con los shingles y luego aplicar LSH para Jaccard. Notar que el ejercicio pide primero utilizar r=2 y luego b=2, que no es lo mismo que utilizar primero b=2 y luego r=2.
   La matriz resultante debería quedar de la siguiente forma:

	V1	V2	V3
A->B	1	0	0
B->D	1	0	0
D->E	1	1	0
A->C	0	1	1
C->D	0	1	1
D->G	0	0	1

5. En cuanto al seguimiento, tienen un buffer de longitud 6 y minimo de repeticiones de 2 chars, por lo que tienen que considerar posiciones de 0 a 5 (podrian considerar que en la 0 no van a encontrar ningun match y contar de 1 a 5) y longitudes de L2 a L6. Con esto en mente, tienen que armar un arbol de huffman con todos los caracteres (pueden tomar ABC a los efectos del ejercicio) y todas las posibles longitudes: A, B, C, L2, L3, L4, L5, L6, inicialmente todos con frecuencia 1. Para las posiciones pueden proponer cualquier codigo prefijo.

   Y comenzar a leer el achivo: ABACBACACBABABAB.

   A
   ventana A leo B -> B
   ventana: AB leo AC -> A
   ventana ABA leo C -> C
   ventana ABAC leo BAC -> L3P2
   ventana BACBAC leo ACBA -> L4P4
   ventana ACACBA leo BABAB -> L5P1

   A medida que leen el archivo y emiten un Char o un par ordenado Longitud - Posicion, y actualizan el arbol de huffman. Deben incluir el desarrollo del seguimiento. Tiene que ser claro que el proceso implica una sola pasada al archivo, si de alguna forma lo procesan como dos pasos independientes (2 metodos en secuencia) esta mal. Es una unica pasada y durante esa unica pasada por cada caracter o par ordenado se genera un codigo que es lo que va conformando el archivo comprimido. Si actualizan mal los arboles -5, si no encuentran bien algun match pero muestran conocimiento del metodo -5. Deben mostrar exactamente bit a bit como queda el archivo comprimido (ya que es la salida del árbol) y como manejan el padding. El archivo se debe poder descomprimir, sino el ejercicio vale 0.

6. Hay que evaluar la calidad de la solución propuesta. Deberian tener en cuenta features del aviso en sí, como ser título, descripción, requisitos, area, etc, Otros en relación a la consulta, por ej si la consulta aparece en el título del aviso, descripción, area, etc. Tambien se podria personalizar el ranking por usuario, teniendo en cuenta la información del usuario en su CV, e intentando rankean mejor los avisos que matcheen con su perfil. Luego para el feedback para poder hacer la parte de aprendizaje se podría utilizar la información de clicks en los avisos y además las postulaciones a los mismos, y de esa forma ir ajustando los pesos que se le da a cada feature para mejorar el rankeo en el tiempo.
   Si no explican en base a qué se realiza el aprendizaje descuento de al menos 5 puntos.
   Si no tienen en cuenta que hay features del aviso en sí y otros del aviso con relación a la query, descuento de 5 puntos.

7. Esencialmente se espera que grafique una serie temporal con un line plot por cada símbolo a analizar, de tal forma que podamos comparar el valor de la acción dia por dia, indicando la referencia a cada símbolo de la acción.
   Para resolver mostrar en la visualización si ese dia la acción cerró a la alza o a la baja, se puede indicar en el punto del plot un color específico que indique si la misma cerró a la alza o baja (por ejemplo se puede utilizar verde para alza, rojo para baja). También se puede en el punto del plot específico en vez de usar un punto usar un triangulo hacia arriba o hacia abajo como alternativa (a la que se suma en conjunto con la de color).
   Si no tiene titulo descuento de 2 puntos.
   Si no tiene referencias, tanto de los códigos de las acciones para interpretar la misma o los colores utilizados para indicar cómo cerro, descuento de 2 puntos. El resto del puntaje queda a criterio del corrector evaluando la solución propuesta.