Organización de Datos 75.06. Segundo Cuatrimestre de 2018. Examen parcial, tercera 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.

"No resurrections this time". - Thanos, Avengers Infinity War

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

Nombre:
Padrón:
Corregido por:

1) La empresa Instagram almacena información de usuarios en un RDD de la forma (id_usuario, nickname, país, fecha_alta) y en otro RDD almacena información de las publicaciones como (id_publicacion, id_usuario, id_foto, comentario, fecha). En Instagram, un hashtag es una palabra que comienza con #. Se pide realizar lo siguiente utilizando el api de RDD en pyspark:

a- Encontrar el top 10 de hashtags que más aparecen y que fueron publicados **únicamente** por usuarios del país Argentina. (Por ej: Si un hashtag fue publicado por alguien que no es de Argentina entonces no hay que considerar dicho hashtag para el top 10, aunque sea el que mayor cantidad de ocurrencias haya acumulado). (9 ptos)

b- Construir y cachear un RDD con los pares de usuarios (id_usuario1, id_usuario2) que tienen hashtags publicados en común. (***) (6 ptos)

2) El equipo comercial en Argentina de instagram cuenta con información sumarizada de usuarios del país en el siguiente Data Frame (id_usuario, nickname, país, fecha_alta, cantidad_seguidores, cantidad_seguidos, cantidad_total_publicaciones).

Uno de sus analistas quiere obtener los potenciales 10 usuarios "influencers" que van a liderar su primera campaña de vía pública en el país y para ello necesita de nuestra ayuda. El criterio para detectar un influencer es que el mismo debe:

- Tener una cantidad de seguidores mayor al 80% de la máxima cantidad de seguidores de todos los usuarios dentro de su país. (Por ejemplo, si un usuario es Argentino y el máximo de seguidores en Argentina es 1000 entonces requerirá tener al menos 800 seguidores).

- El número de usuarios que esa persona sigue no debe superar el 15% de la cantidad de seguidores que él tiene.

El criterio final para seleccionar a los candidatos será a partir de la mayor cantidad de seguidores que tengan.

Escribir un programa en python Pandas que pueda obtener esos usuarios (***)(15 pts)

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) En t-SNE, si dos puntos A y B a distancia 1 se consideran como cercanos entonces si C y D (distintos a A y B) también están a distancia 1 también se consideran cercanos. (**)(10 pts)

b) Para matrices grandes la SVD no es una opción para reducir dimensiones por un tema de eficiencia. (**) (10 pts)

c) En PCA, se obtendrá una mayor varianza en los datos si los proyectamos contra la segunda componente principal en vez de la primera. (**) (10 pts)

d) Dado que t-SNE escala cuadráticamente dada la cantidad de inputs que tenga no podemos utilizarlo para grandes volúmenes de datos. (**) (10 pts)

4) Sean los siguientes vectores en 5 dimensiones: v1 = [4 4 -5 -2 3] ; v2= [-3 -2 -4 5 0]; v3=[3 2 -1 -2 1]. Y sean los siguientes 6 hiperplanos aleatorios: r1=[1 1 1 1 -1]; r2=[1 -1 -1 -1 1]; r3=[1 -1 -1 -1 -1];r4=[-1 1 1 -1 -1]; r5=[1 -1 -1 -1 1]; r6=[-1 1 1 -1 1]. Utilizando b=2, r=3 y tablas hash de 13 buckets encontrar una función de hashing universal h(a,b,c) tal que en la primer banda solo haya colisión entre v1 y v2 y en la segunda banda colisionen todos.

(***) (10pts)

5) (15ptos) (***) Dados los siguientes documentos:

Abanico Azul Aro

Arco Aro Avion

Arco Arpa Azul Avioneta

Avion Avioneta

Sabiendo que se construyó sobre ellos un índice invertido utilizando front coding parcial (n=3) y códigos gamma para codificar los punteros, **indicar cuántos accesos son necesarios** para resolver la consulta "Arco Azul" detallando paso a paso la forma en que se resuelve la misma.

6) Se implementa un algoritmo de Block Sorting utilizando en su última etapa un compresor de Huffman. Teniendo el siguiente archivo con la metadata correspondiente, se pide descomprimir el mismo. Suponer que para el armado del árbol se respeta el orden en que aparecen los símbolos en la tabla. (***) (15pts)

```
<tabla_frecuencias>Símbolo0=7, Símbolo2:1, Símbolo3:1, Símbolo4:3</>
<Índice de BW>L=1</>
<Vector Move to Front>ABCDE</>
1111010000101001100
```

7) A partir de toda la información disponible sobre publicaciones de usuarios de instagram y sus comentarios, utilizada en el primer punto para calcular el top10, realizar una visualización en la que se puedan distinguir la relación que existe entre los usuarios (similitudes y diferencias) en base a los hashtags utilizados en sus publicaciones (***) (10 pts)