**Organización de Datos 75.06**. Primr Cuatrimestre de 2016. 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 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.

"What you are will show in what you do." (Thomas Edison)

| # | 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) UBER almacena en un cluster todos los datos sobre el movimiento y viajes de todos sus vehículos. Existe un proceso que nos devuelve un RDD llamado trip_summary con los siguientes campos: (driver_id, car_id, trip_id, customer_id, date (YYYYMMDD), distance_traveled), Programar usando Py Spark un programa que nos indique cual fue el conductor con mayor promedio de distancia recorrida por viaje para Abril de 2016. (***) (15 pts)

2) Dados los archivos:

- Notas (Padron, Codigo de Materia, Codigo de Curso, Nota)
- Cursos (Codigo de Materia, Nombre de Materia, Codigo de Curso, - Profesor a Cargo)

Hacer un programa en PIG que liste para cada curso de cada materia, el promedio de notas de los alumnos que aprobaron la misma. El listado debe contener Codigo de Materia, Codigo de Curso, Profesor a cargo y Promedio de Notas, y debe estar ordenado por materia. Tener en cuenta solo los cursos que tengan al menos 100 alumnos aprobados.
aero. (***) (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 las claves a hashear son numéricas y aleatorias entonces no es necesario que el número m sea primo en la función de hashing . a*x mod m.(*) (10 pts)

b) Si tenemos una matriz de mxn y queremos reducirla a k dimensiones entonces podemos aplicar la SVD y luego calcular X=U*S*V(t) y quedarnos las primeras k columnas de X. (*) (10 pts)

c) Aumentar la cantidad de funciones de hashing o la cantidad de registros por bucket en el método del cuckoo sirve para reducir la cantidad de accesos promedio que necesitamos para recuperar una clave. (*) (10 pts)

d) La maldición de la dimensionalidad para KNN está representada porque en muchas dimensiones todos los puntos estarán mas o menos a igual distancia uno del otro (****) (10 pts)

4) Tenemos un compresor aritmético dinámico de orden 0 que trabaja procesando bit por bit. Si comprimimos un archivo que está formado por una serie de 1000 bits en 1 y luego dos bits en 0. ¿cuántos bits ocupará el archivo comprimido? (15 pts) (****)

5) Dados los soguientes textos considerando a cada línea como un documento:

Luna Apolo programa NASA

Marte mision Pathfinder NASA

Luna mision NASA Marte mision NASA NASA mision Marte

Luna Marte Luna Marte Luna Marte Luna Marte

a) Usando TF-IDF calcular la relevancia de cada documento para la consulta "mision Apolo,Luna". Indicar detalladamente todos los cálculos realizados. (*) (10 pts)

b) Explique que ventaja tiene BM25 sobre el esquema usado en el punto "a" (esto es en general, no solo para la consulta) (**) (5 pts)

6) Usando la distancia de Jaccard y 36 minhashes se quiere comparar el efecto de usar 6 construcciones OR y luego 6 AND contra usar primero 6 construcciones AND y luego 6 OR. ¿En qué casos tendremos mas falsos positivos y en que casos mas falsos negativos? Si fijamos d1=0.2 y d2=0.5 ¿cuál es la probabilidad de que dos documentos sean candidatos en cada caso? (****) (10 pts)

7) Un sensor registra la intensidad de los terremotos para una determinada zona, los valores que registró en el último año son: 2.83-4.5-1.52-2.5-6.45-1.8-3.3-5.3-1.75-2.6-1.2-4.1-2.35. Se pide realizar un histograma para visualizar estos datos usando 7 bins o intervalos. Luego indicar que conclusiones pueden obtenerse a partir de la visualización realizada. (**) (10 pts)