Métodos Analíticos **Examen parcial**
Primavera 2021

Nombre: _____

Instrucciones (124 puntos) Encierra en un círculo la respuesta o respuestas correctas. No hay preguntas capciosas.

(8^{pts}) **1.** Supongamos que aplicamos el algoritmo a-priori para contar conjuntos frecuentes en datos de transacciones. Selecciona todas las que apliquen:

(a) Es posible que no encontremos algunos pares que son frecuentes.

(b) El algoritmo funcionará bien si el número de artículos por canasta es grande (cientos de miles).

(c) Es necesario contar la frecuencia de cada artículo que ocurre en las canastas.

(d) Es necesario contar la frecuencia de cada par de artículos que ocurren en las canastas.

(8^{pts}) **2.** Supongamos que sabemos que $\{A, B, C\}$ y $\{A, C, D, E\}$ son conjuntos frecuentes en un análisis de canastas. ¿Cuáles de los siguientes tienen que ser conjuntos frecuentes?

(a) $\{A, C\}$ (b) $\{B, D\}$ (c) $\{A, D\}$ (d) $\{A, B, C, D\}$

(5^{pts}) **3.** Supón que la regla Fresas ->Crema tiene confianza 0.3 y lift 1.5. ¿Cuál es el soporte de Crema?

(a) 0.015 (b) 0.15 (c) 6.0 (d) 0.67 (e) 0.2

(8^{pts}) **4.** Considera los documentos DIARIA y SANDIA. Si utilizamos tejas de tamaño 2, ¿cuáles de las siguientes son verdaderas?

(a) El conjunto de tejas de DIARIA es de tamaño 4

(b) El conjunto de tejas de SANDIA es de tamaño 5

(c) El conjunto de tejas de SANDIA es de tamaño 4

(d) Los dos documentos tienen 4 tejas en común

(8^{pts}) **5.** Considera la siguiente matriz tejas-documentos:

	c1	c2	c3	c4
1	0	1	1	0
2	1	0	1	1
3	0	1	0	1
4	0	0	1	0
5	1	0	1	0
6	0	1	0	0

Cuáles de las siguientes son verdaderas, si usamos similitud de jaccard?

(a) c1 y c2 tienen similitud 1/5

(b) c3 y c4 tienen similitud 1/5

(c) c3 y c4 tienen similitud 1/6

(d) c1 y c2 tienen similitud 1/6

(8^{pts}) **6.** Considera la matriz de tejas-documentos del ejercicio anterior. Tomamos una permutación que resulta en el orden de renglones 4,6,1,3,5,2. ¿Cuáles de las siguientes son verdaderas? (Nota: cuenta el minhash comenzando en 1).

(a) El minhash del documento c4 es 5.

(b) El minhash del documento c3 es 2.

(c) El minhash del documento c2 es 2.

(d) El minhash del documento c1 es 5.

(6 pts) **7.** En la práctica, cuando hacemos minhashing usamos funciones hash apropiadas de los renglones en lugar de simular permutaciones de los renglones de la matriz tejas-documentos. La razón es que

(a) Las funciones hash dan una mejor aproximación de la similitud jaccard que las permutaciones.

(b) Calcular permutaciones de un gran número de renglones es más costoso que calcular hashes.

(c) Las funciones hash garantizan que no hay distintos reglones o tejas.

(8 pts) **8.** ¿Cuáles de las siguientes son verdaderas?

(a) La proporción de coincidencias entre dos firmas minhash de dos documentos da la similitud de jaccard.

(b) La probabilidad de que el mismo minhash de dos documentos sea igual es aproximadamente la similitud de jaccard de los dos documentos.

(c) Las firmas minhash de dos documentos idénticos son iguales

(d) Si dos firmas minhash son iguales, entonces los documentos son idénticos.

(5 pts) **9.** Supón que calculamos las firmas con 5 minhashes de 100 documentos. Si dos documentos tienen similitud de jaccard 0.2, ¿cuál es la probabilidad aproximada de que los 5 minhashes coincidan?

(a) $(1-0.2)^5$ (b) $(1-0.2)^{100}$ (c) $(0.2)^5$ (d) $(0.2)^{100}$

(5 pts) **10.** Supón que tienes una colección de 100 millones de objetos. Si en un segundo puedes calcular 1000 millones de similitudes, ¿cuánto tardarías aproximadamente (tiempo total de cómputo) en calcular todas las posibles similitudes? Menos de:

(a) un minuto (b) una hora (c) un día (d) un mes (e) un año

(8 pts) **11.** Supongamos que nos interesa encontrar todos los pares de documentos con similitud de jaccard mayor a 0.8 ¿Cuáles de las siguientes son verdaderas?

(a) Si usamos LSH con número de bandas y hashes adecuados, siempre capturamos todos los pares de documentos con similitud mayor a 0.8.

(b) Para un número total k fijo de hashes, hacemos bandas mas grandes si queremos capturar documentos de similitud más baja.

(c) Para un número total k fijo de hashes, hacemos más bandas si queremos capturar documentos de similitud más baja.

(d) En LSH, siempre es mejor utilizar un número total de hashes más grande.

(5 pts) **12.** Considera tres vectores $v_1 = (1, 3)$, $v_2 = (-2, -2)$ y $v_3 = (1, -3)$ escogidos al azar. Usando la familia sensible a la localidad de proyecciones (hiperplanos aleatorios para similitud coseno), ¿cuál la firma del elemento $x = (1, 1)$?

(a) $(1, -1, -1, 1)$ (b) $(1, -1, -1)$ (c) $(1, -1)$ (d) 1

(8 pts) **13.** ¿Cuáles son razones por las que es buena idea centrar las calificaciones por usuario en un sistema de recomendación?

(a) Hacer más comparables las calificaciones entre usuarios.

(b) Hacer más rápido el cálculo de similitud entre artículos.

(c) La similitud coseno es una mejor medida de similitud si utilizamos las calificaciones centradas por usuario.

(d) Controlar parte de la heterogeneidad en el uso de la escala por los usuarios.

(8 pts) **14.** Selecciona todas lo que apliquen. Al aplicar la DVS (descomposición en valores singulares) a una matriz de datos X:

(a) Es necesario que los datos de X estén estandarizados

(b) La DVS se puede utilizar como una técnica de reducción de dimensionalidad

(c) Frecuentemente, si las unidades de las columnas son distintas y/o la escala de las columnas es muy distinta puede ser conveniente normalizar por columna antes de aplicar la DVS.

(d) Al aplicar la DVS, las dimensiones que más contribuyen a la aproximación de los datos son las que tienen los valores singulares más chicos.

(e) La DVS se puede utilizar para aproximar la matriz de datos con otra matriz de rango bajo.

(5pts) **15.** Considera la matriz

```
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
[3,]    7    8    9
```

que queremos aproximar mediante una descomposición UV^t, donde U y V tienen 1 columna. Si usamos mínimos cuadrados alternados, tomamos inicialmente $V = (5, 5, 5)^t$, y calculamos la primera iteración de $U = (x, y, z)^t$, entonces x es igual a:

(a) 0 (b) 2/5 (c) 1 (d) 7/5 (e) Ninguno de los anteriores

(5pts) **16.** Para hacer componentes principales de una matriz de datos X:

(a) Centramos todos los renglones por su media y aplicamos la descomposición en valores singulares (DVS)

(b) Centramos todas las columnas por su media y aplicamos la DVS

(c) Centramos los datos con la media general de la tabla y aplicamos la DVS

(8pts) **17.** Considera una gráfica no dirigida con las aristas a-b, a-c, a-d, d-e. ¿Cuáles de las siguientes son verdaderas si usamos la medida de centralidad de eigenvector para medir importancia de los nodos?

(a) Algunos nodos tienen importancia igual a 0

(b) El nodo d es el más importante

(c) Los nodos c, b y e tienen la misma importancia

(d) Los nodos c y b tienen la misma importancia

(8pts) **18.** Considera una gráfica no dirigida con las aristas a-b, b-c, c-d, d-e y e-a. ¿Cuáles de las siguientes son verdaderas si usamos la medida de betweeness para medir importancia de los nodos?

(a) Algunos nodos tienen importancia igual a 0

(b) El nodo d es el más importante

(c) Los nodos c, b y e tienen la misma importancia

(d) Los nodos c y b tienen la misma importancia