

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). 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 consultas 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 estará disponible en forma pública en el grupo de la materia.

"Machines take me by surprise with great frequency." (A. Turing)

#	1	2	3.1	3.2	4	5	6	7	Entrega Hojas:
Corr									Total:
Puntos	/15	/15	/5	/5	/15	/15	/15	/15	/100

Nombre:
Padrón:
Corregido por:

1) Dado el siguiente mensaje seleccionar entre las siguientes la clave más adecuada para aplicar cifrado por Vigenere :

Mensaje: "ELQUEPOQCOCOMEPOCOCOMPRÁ"

Claves:
 - AAAAA
 - 23
 - CT
 - ABT

¿Por qué considera que es la mas adecuada? (***) (10 pts)
 Explique un método para criptoanalizar un criptograma cifrado con vigenere. (**) (5 pts)

2) Genere un clasificador bayesiano naive para determinar si el siguiente mensaje de sms es spam o no:

"compre 0km descuento" D

basándose en el siguiente set de entrenamiento:

auto 0km venta - SPAM.
 5min descuento gol - NO SPAM
 super lanzamiento 0km - SPAM
 compro super pan - NO SPAM
 super descuento auto - SPAM

(***) (15 pts) 0,00048 \rightarrow cerca de underflow!

3.1) ¿Por qué en una base de datos de documentos es muchas veces recomendable desnormalizar los datos, duplicando información? (*) (5 pts)

3.2) ¿En qué casos usaría una base de datos de grafos en lugar de una para documentos? (*) (5 pts)

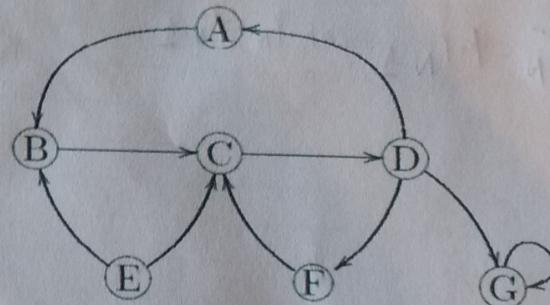
4) Utilizando RDF haga una representación de los metadatos de un partido de fútbol de la copa del mundo teniendo en cuenta los siguientes recursos:

Partido; Equipo; Jugador; Director Técnico; Sede.

Para ello tengamos en cuenta la siguiente frase:

"Argentina e Iran se batieron a duelo el 21 de Junio de 2014 en el Estadio Mineirão en Belo Horizonte. El resultado del Partido fue 1-0 a favor de Argentina con Gol de Lionel Messi. El equipo argentino fue dirigido por A. Sabella mientras el conjunto Irani por C. Queiroz" (***) (15 pts)

5) Usando $b=0.8$ ejecutar pagerank para el siguiente grafo hasta la convergencia indicando el ranking final de cada nodo. (**) (15 pts)



6) A partir de los siguientes puntos en el plano muestre el funcionamiento del algoritmo K-means con K=3 hasta que no haya cambios. En cada iteración grafique los puntos, centroides y clusters generados en el plano. Usar los puntos en negrita (**A1,A4,A7**) como centroides iniciales. (***) (15 pts)

A1: (2,10) A2: (2,5) A3: (8,4) A4: (5,8) A5: (7,5) A6: (6,4)
 A7: (1,2) A8: (4,9)

7) En un arbol B+ con capacidad para dos claves por nodo realizar las siguientes operaciones . A=Alta y B=Baja. A(11), A(24), A(4), A(12), A(38) A(36) A(17) B(17) B(11) B(12) B(4) (***) (15 pts)

$$\begin{aligned}
 & \frac{L+L}{2} = 0,0008 \\
 & \frac{L+2}{2} = 0,0009 \\
 & \frac{L+4}{2} = 0,001 \\
 & \frac{L+6}{2} = 0,0011 \\
 & \frac{L+8}{2} = 0,0013
 \end{aligned}$$

⑤	A	B	C	D	E	F	6
A	0	0	$\frac{1}{3}$	0	0	0	
B	1	0	0	0	$\frac{1}{2}$	0	0
C	0	1	0	$\frac{1}{4}$	$\frac{1}{2}$	0	
D	0	0	1	0	0	0	
E	0	0	0	0	0	0	
F	0	0	0	$\frac{1}{3}$	0	0	
6	0	0	0	$\frac{1}{3}$	0	0	1

$\rightarrow (\text{Pre } \beta)$

$$\beta = 0.8$$

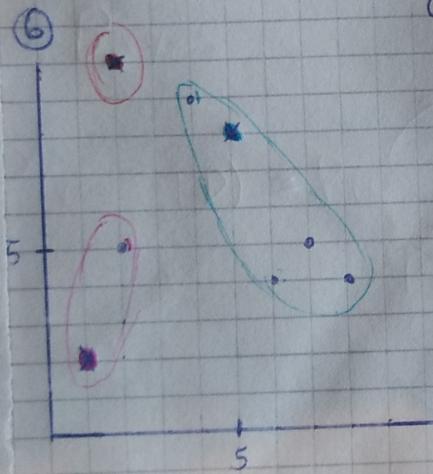
$$M_{\text{new}} \xrightarrow{\beta} \beta M + (1-\beta) \begin{pmatrix} \frac{1}{17} & & \\ & \ddots & \\ & & \frac{1}{17} \end{pmatrix}$$

~~Repetitio~~ ~~Repetitio~~

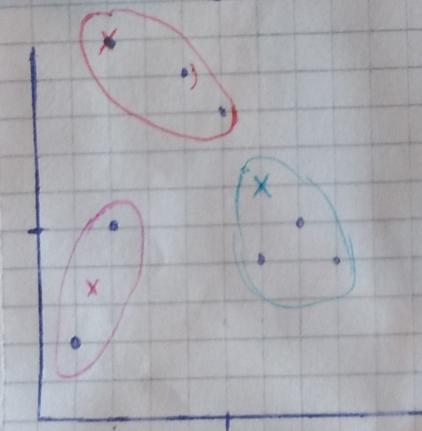
$$M = \begin{matrix} 0,03 & 0,03 & 0,03 & 0,30 & 0,03 & 0,03 & 0,03 \\ 0,83 & 0,03 & 0,03 & 0,03 & 0,43 & 0,03 & 0,03 \\ 0,03 & 0,83 & 0,03 & 0,03 & 0,43 & 0,83 & 0,03 \\ 0,03 & 0,03 & 0,83 & 0,03 & 0,03 & 0,03 & 0,03 \\ 0,03 & 0,03 & 0,03 & 0,03 & 0,03 & 0,03 & 0,03 \\ 0,03 & 0,03 & 0,03 & 0,30 & 0,03 & 0,03 & 0,03 \\ 0,03 & 0,03 & 0,03 & 0,30 & 0,03 & 0,03 & 0,03 \end{matrix}$$

$$r_0 = (1_7 \dots 1_7)^T$$

10. $r_1 = Mr_0 ; Mr_1 = Mr_1 \dots \boxed{r_N = Mr_{N-1} / r_N = r_{N-1}}$ convergence
centerades



$$\begin{aligned} C1 &\rightarrow 2,1 \\ C2 &\rightarrow (4,8) + (5,8) + (6,4) + (7,5) + (8,4) / 5 = (6,6) \\ C3 &\rightarrow (2,1) + (4,2) = \left(\frac{3}{2}, \frac{7}{2}\right) \end{aligned}$$

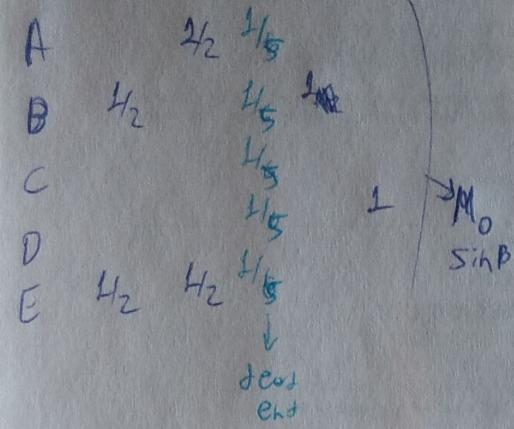


$$\begin{aligned} C1 &\rightarrow (2,1) + (4,9) + (5,8) / 3 = (3,6; 9) \\ C2 &\rightarrow (6,4) + (7,5) + (8,4) / 3 = (7,4; 3) \\ C3 &\rightarrow \left(\frac{3}{2}, \frac{7}{2}\right) \end{aligned}$$

\rightarrow No mas cambios

To convergence
& Beyond!

⑥ A B C D E



$$\beta = 0.85$$

$$M_0 = \frac{B h_0}{\sin \beta} + (L - B) \left(\frac{L/5}{L/5} \right)$$

$$= 0.03$$

$$\therefore 0.03$$

$$M_0 = \begin{pmatrix} 0.08 & 0.455 & 0.2 & 0.03 & 0.03 \\ 0.455 & 0.03 & 0.2 & 0.88 & 0.03 \\ 0.03 & 0.03 & 0.2 & 0.03 & 0.03 \\ 0.03 & 0.03 & 0.2 & 0.03 & 0.88 \\ 0.455 & 0.455 & 0.2 & 0.03 & 0.03 \end{pmatrix}$$

$$r_0 = (H_5 \cdot H_5 \ H_5 \ H_5 \ H_5)^T$$

$$r_1 = K_{00}^{-1} M_0 r_0 = \begin{pmatrix} 0.149 \\ 0.319 \\ 0.064 \\ 0.023 \\ 0.239 \end{pmatrix}$$

⑦

	OBL	MAT	AVAT
A	5	5	2
B	2	4	3
C	4	4	0
D	4	4	0

Normalize
by row col

1.3	0.7	-1
-1.7	0	0
-0.3	1	1
0.3	-0.3	0
Normal	2.16	0.82

sin(jA)	-0.43	-0.86	1
---------	-------	-------	---

$$\frac{\langle P, \text{AVAT} \rangle}{\|P\| \|AVAT\|}$$

$$r_{D, \text{AVAT}} = \frac{r_{D, \text{OBL}} \cdot s_{\text{OBL, AVAT}} + r_{D, \text{MAT}} \cdot s_{\text{MAT, AVAT}}}{s_{\text{OBL, AVAT}} + s_{\text{MAT, AVAT}}} = \frac{4 \cdot (-0.43) + 4 \cdot (-0.86)}{-0.43(-0.86)} = 4$$

Organización de Datos (75.06) Segundo Cuatrimestre de 2014. Gran Examen por promoción. [2014_2c_Promoción]

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). Este enunciado debe ser entregado junto con el parcial si quiere una copia del mismo puede bajarla del grupo de la materia. Si tiene dudas o consultas levante la mano, está prohibido hablar desde el lugar, fumar o cualquier actividad que pueda molestar a los demás.

"Lo importante no es haber llegado, sino disfrutar del camino." (D. Lamas)

#	1	2	3	4	5	6	7	8	9	10	Grd	Hojas:
Corr												Total:
Puntos	/10	/10	/10	/10	/10	/10	/10	/10	/10	/10	/20	/100

Nombre:
Padrón:
Corregido por:

- 1) Clustering: Tenemos dos centroides: (0,0) y (100,40). Dados los siguientes puntos indique para cual de ellos cambiaria el centroide al cual se lo asigna según usemos la distancia Manhattan o la distancia euclídea.

(53,15) (51,15) (50,18) (52,13)

A B C D

- 2) Arboles y Signatures: Se usan tres funciones de hashing en 0.5 para construir un signature file. Se dan los resultados de las funciones para dos palabras "a" (0,2,3) "b" (0,1,5). Si tenemos un documento cuyo signature es 110101 entonces:

- a) El documento puede o no tener "a"
- b) El documento puede o no tener "b"
- c) El documento no tiene "a"
- d) El documento no tiene "b"

(Marcar todas las opciones correctas)

- 3) Streams: Se usan tres funciones de hashing 0..4 para el algoritmo count-min. Para las siguientes palabras se da el resultado de las tres funciones de hashing. "casa" (2,1,2) "canasta" (0,3,3) "kilo" (4,1,0) "alfa" (3,3,0). Si los filtros son los siguientes:

$$F1 = [0,2,3,1,3] \quad F2 = [1,2,2,4,0] \quad F3 = [3,0,1,1,1]$$

¿Cuál de las cuatro palabras estimamos como la más frecuente?

- 5) Redes Sociales: Zoilo tiene 4 amigos: Armando, Barbara, Claudio y Diana. De estos Barbara y Claudio son amigos pero los demás no se conocen. ¿Cuál es el coeficiente de clustering de Zoilo? $2.1/4.3 = 1/2$

- 7) Recomendaciones: Se sabe que el promedio global de todas las calificaciones es de 2.28. El usuario "Ariel" ha realizado las siguientes calificaciones: (1,4,2,1,1) la película "Interstellar" tiene las siguientes calificaciones (5,4,1,3). ¿Cuál sería la estimación de la calificación de Ariel para Interstellar?

$$R_{A,I} = \mu + (\mu - \text{Pr}_{\text{on}}(I)) + (\mu - \text{Pr}_{\text{on}}(\frac{I}{2})) + \dots$$

$$2.28 + (2.28 - 1.8) + (2.28 - 3.25) = 2.74$$

- 9) Metadatos: Escriba en RDF, notación Turtle, una representación de la frase "Facebook, Twitter e Instagram son Redes Sociales". Juan es usuario de Facebook e Instagram.

- 6) Page Rank: Dados los siguientes links: (B,H) (H,C) (C,H) (C,B) (B,C). Usando $b=0.85$ indique el page rank de B,C y H luego de 3 iteraciones.

- 8) Clasificación: Aplicamos SVM lineal a nuestro set de entrenamiento y tenemos una precisión del 93%, sin embargo al aplicarlo al set de pruebas la precisión baja al 62%. ¿Qué podemos hacer? (Indicar todas las opciones correctas)

- a) Recolectar mas datos → Síempre!!!
- b) Usar una función kernel
- c) Aumentar el parámetro "C" X) Mayor C, Mayor OF
- d) Disminuir el parámetro "C" ✓

- 10) NoSQL: Explique qué características tiene una base de datos para documentos.

	A	B	C	D
C1	55 .08	53 .16	53 .24	53 .60
C2	72 .23	74 .01	72 .63	75 .07

M E

$$\text{def eucl}(x,y) = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2}$$

$$\text{def man}(x,y) = |x_1-y_1| + |x_2-y_2|$$

(También se puede hacer
a ojo!)

→ M: (A B C D) ()
E: (B C D) (A)
J → (53,15)
(51,15.3)

M E

C1	C1	C2	C1	C1	C2	C1	C2	C2
✓	✓	✓	✓	✓	✓	✓	✓	✓

coh dist EUCL, en la Segunda
iter D pasa de C1 a C2!

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). Este enunciado debe ser entregado junto con el parcial si quiere una copia del mismo puede bajarla del grupo de la materia. Si tiene dudas o consultas levante la mano, está prohibido hablar desde el lugar, fumar o cualquier actividad que pueda molestar a los demás.

"It ends tonight" (Neo, The Matrix)

#	1	2	3	4	5	6	7	8	9	10	Grd	Hojas:
Corr												Total:
Puntos	/10	/10	/10	/10	/10	/10	/10	/10	/10	/10	/20	/100

Nombre:

Padrón:

Corregido por:

1) Clustering: Sean los siguientes puntos en dos dimensiones: (5,1) (5,3) (4,4) (9,4) (10,3) (11,6). Aplicar K-Means comenzando con los centroides (8,1) y (7,5). (***) (10 pts)

2) Arboles y Signatures: Dar las siguientes altas y bajas sobre un árbol B con capacidad máxima para 2 (dos) claves por nodo. A(28), A(3), A(21), A(39), B(3), A(11), A(51), A(42), B(39), B(42). (**) (10 pts) Con este ejercicio despedimos a este tema de nuestra materia para siempre. So long, and good bye.

3) Streams: Dado el siguiente stream: {1,3,2,3,3,5,2,1} indique una función de hashing de la forma $h(x) = ax + b \pmod{32}$ de forma tal que el algoritmo de Flajolet Martin se aproxime de la mejor forma posible al cálculo del momento de orden 0 del stream. Usar 5 bits y tomar los bits a derecha para el algoritmo. (****) (10 pts)

4) qsdhxhb hrdhxhb lqe dc当地 ahjhr av jhxv algv dc当地 ihaa sqdc vrx sqaa nlq xlir ahjhr av jhxv algv (****) (10 pts)

5) Redes Sociales: Sean las siguientes relaciones de amistad en una RS tipo Facebook (grafo no dirigido): (A,B) (A,C) (B,D) (C,D) (D,E) (D,F) (E,C) (E,F). a) Calcular el coeficiente de clustering promedio de la red. b) Calcule el betweenness de cada nodo. (*) (10 pts)

6) Page Rank: Dados los siguientes links en un grafo dirigido (A,B) (B,D) (D,A) (C,A) (D,E) se pide calcular el PageRank de cada nodo usando $b=0.8$ y realizando 3 (tres) iteraciones a partir del vector $(1/5, 1/5, 1/5, 1/5, 1/5)$. (***) (10 pts)

7) Recomendaciones: tenemos las calificaciones del usuario "Alice" para cinco películas: (3,1,5,2,?). Usando el coeficiente de correlación de Pearson indique cuál es la semejanza de A con los usuarios "Bob" y "Claire" y luego estime la calificación de Alice para la quinta película. Bob: (2,1,4,1,2) Claire: (4,2,5,3,5). (**) (10 pts)

8) Sean los siguientes puntos en dos dimensiones y sus labels: (3,1,+1) (1,3,+1) (5,4,+1) (2,4,-1) (4,2,-1). Queremos clasificarlos usando un SVM. Indicar qué tipo de kernel usaría y con qué parámetros. (**) (10 pts)

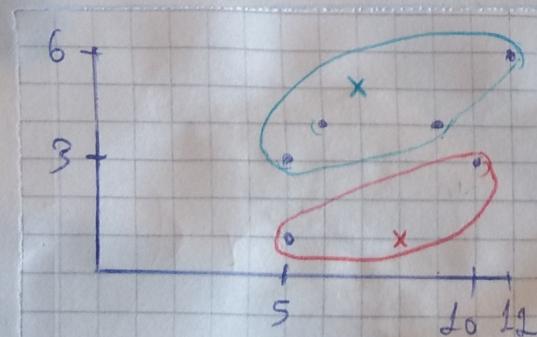
9) Metadatos: e cuenta con una colección de datos relacionados con publicaciones de libros representados en turtle. El siguiente es un ejemplo de como se encuentran guardados los mismos:

10) NoSQL: De un ejemplo en el cuál sea conveniente usar una base de datos para grafos para datos que en su forma natural no tienen forma de grafo, por ejemplo texto o información tabular. Justifique adecuadamente. (****) (10 pts)

pb:Fundacion a pb:book ;
dc:title "Fundacion" ;
dc:creator pb:IsaacAsimov ;
dc:date "1951".

b:IsaacAsimov a pb:person ;
pb:nationality "Estadounidense" ;
pb:name "Isaac Asimov".

Se pide escribir una consulta en SPARQL que liste, todos los libros publicados en la década del 50, publicados por autores estadounidenses. Para cada uno, se pide mostrar título del libro, año de publicación, y nombre de autor, ordenado por el título. (***) (10 pts)



$$C_1 \rightarrow (5,3+6,4+9,4+11,6)/4 = (7,75; 4,25)$$

$$C_2 \rightarrow (10,3+5,2)/2 = (7,5; 2)$$

Loop until
infinity

convergence.

LARGOS.
JUNO QUE
ME SALEN

③ 1 3 2 3 3 5 2 1

$$M^0(S) = 4 \quad (\# \text{events} \text{ } \text{diseños})$$

$$M^0(S) = 2 \xrightarrow{\text{Fijo}} \begin{array}{l} \max \\ \# \text{Os} \end{array} \& \text{la derecha de } h(x)$$

Flotante
Máximo

Busco $r=2$

Busco $h(x)/h(x) \forall x \in [1, 5] \quad \text{cond}(x) \leq ax + b \leq 32$

$$h(x) = \underbrace{\begin{array}{r} 001000 \\ 011000 \\ 101000 \\ 111000 \end{array}}_{\text{Alfabeto}} \quad \text{y } h(x) \neq \underbrace{\begin{array}{r} 010000 \\ 110000 \\ 100000 \\ 000000 \end{array}}_{\text{Número}}$$

8
24
16
0

$$h(x) = x \cdot 32$$

$$h(1) = 00001 \quad r=0$$

$$h(2) = 00010 \quad r=1$$

$$h(3) = 00011$$

$$h(4) = 00100$$

$$h(5) = 00101$$

$$\boxed{r=2} \rightarrow 2^r = 4 \checkmark$$

$\frac{1}{1000000}$

	0	1	2	3	4	5	6	7
A	3	1	5	2	0			
B	2	1	4	1	2			
C	4	2	5	3	5			

Norm.

$U-U$
(XF^{1/2})

$$0.25 \quad -1.75 \quad 2.25 \quad -0.75$$

$$0 \quad -1 \quad 2 \quad -1 \quad 0$$

$$0.2 \quad -1.8 \quad 2.2 \quad -0.8 \quad 1.2$$

	Norma	Sigma
1	2.96	1
2	2.45	0.97
3	2.61	0.84

$$r_{A,5} = \frac{r_{B,5} \cdot S_{BA} + r_{C,5} \cdot S_{CA}}{S_{BA} + S_{CA}} = 3,39 \approx \boxed{3}$$

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). Este enunciado debe ser entregado junto con el parcial si quiere una copia del mismo puede bajarla del grupo de la materia. Si tiene dudas o consultas levante la mano, está prohibido hablar desde el lugar, fumar o cualquier actividad que pueda molestar a los demás.

"Even dragons have their endings" (The Hobbit)

#	1	2	3	4	5	6	7	8	Grd	Hojas:
Corr										Total:
Puntos	/12.5	/12.5	/12.5	/12.5	/12.5	/12.5	/12.5	/12.5	/20	/100

Nombre:
Padrón:
Corregido por:

1) Clustering: Sean los siguientes puntos en dos dimensiones: (2,10), (2,5), (8,4), (5,8), (7,5), (6,4), (1,2), (4,9). Usando los puntos subrayados como centroides iniciales y la distancia euclídea aplicar K-Means hasta la convergencia. Dibujar en el plano el resultado de cada iteración. (**) (12.5 pts)

Ejercicio repetido → 1C2014!!

3) Streams: Dado el siguiente stream: [1,3,4,3,4,2,2,4,2,2]. Calcular el momento de orden 2 del stream (número sorpresa) (*) (2.5 pts) luego realizar una estimación usando AMS a partir del quinto elemento del stream es decir considerando [2,2,4,2,2] usar el promedio entre estimadores para la estimación final (*) (10 pts)

5) Redes Sociales: Tenemos un grafo con m nodos y k aristas. Queremos saber si el grafo responde a las características de una Red Social. Indique que tipo de análisis haría para determinar esto, que variables analizaría y que valores debería esperar si el grafo respondiera al comportamiento de una Red Social. (***) (12.5 pts)

7) Existen 3 especies de flores distintas: "Iris Setosa", "Iris Virginica", "Iris Versicolor". Queremos almacenar mediante RDF con notación Turtle los datos de recolección de flores de estas especies. Por cada flor registramos su especie, el tamaño de los pétalos, sépalos y el lugar en donde fue recolectada la flor. En primer lugar definir una ontología basada en RDF Schema (6 pts) (**) finalmente dar un ejemplo en donde se describa la recolección de al menos 4 flores. (6.5 pts) (*)

2) Recomendaciones: Conocemos las calificaciones que tres usuarios han hecho sobre 5 películas: U1=(4,2,5,2,2) U2=(2,3,4,1,5), U3=(5,1,4,2,?). Queremos estimar la calificación del usuario 3 para la película 5 usando semejanza user-user y desviaciones sobre el promedio global. (***) (12.5 pts)

4) Desencriptar el siguiente criptograma (****) (10 pts) e indicar a que libro pertenece la frase desencriptada (*) (2.5 pts)

LYTSTKUVPLYKVEZKRTZPPRKVEKAXPCNHLVLLPAKVIUPGTLYKVE
XPCBTSLHKVZXCUCHZZAKVI UPGTLYKVEKAXPCZPROCLKL
VPLHZNHXULYTUPGTLYKVEEXPNTSHALT

6) Page Rank: Dados los siguientes links: (A,C), (C,D), (D,E), (F,A), (E,A), (A,B), (F,B). En primer lugar plantear el sistema M*V inicial y calcular el PR para la primera iteración (6 puntos) (**) luego indique sin hacer cuentas cual va a ser el nodo con mayor y menor PR final justificando (2 pts) (*) finalmente indique cual sería el link a agregar que mas posiciones mejoraría al nodo F en su PR (nos referimos el PR final), justificar adecuadamente su propuesta. (4.5 puntos) (****)

8) Tenemos los siguientes datos acerca de partidos de fútbol americano disputados por un cierto equipo:

Juega de	Mes	Rwin?	RESULTADO
Local	Octubre	Si	Win
Visitante	Octubre	No	Win
Local	Noviembre	No	Win
Local	Noviembre	Si	Lose
Visitante	Diciembre	No	Lose
Visitante	Diciembre	Si	Lose
Local	Diciembre	No	Win

Rwin indica si el rival tiene record ganador.

Construya un árbol de decisión para predecir el resultado en función de los otros tres atributos. (***) (10 pts)

Sugiera alguna mejora al modelo construido (****) (2.5 pts)

③ 1 3 4 3 4 2 2 4 2 2

$$R_{\text{col}}: 1^2 + 4^2 + 2^2 + 3^2 = 30$$

Ans de 22422, con 2 estimadores agrupados en 1 grupo de 2

2, X X X 4

4, 1

$$M_2^2(S) = 10(2 \cdot 4 - 1) = 70$$

$$M_2^2(S) = 10(2 \cdot 2 - 1) = 10$$

$$\text{Prob} = 40 \rightarrow \text{Media} = \underline{\underline{40}}$$

8) Pron. Ponderado
Vicorias

Quiero clasificar \rightarrow Divido cada Poso segün
Victorias con atributo mayor ponderación

$$\hookrightarrow GI(\text{atributo}) = H(\text{ent. crosificación}) - H(\text{atributo})$$

$$H(\text{res}) = H(4/7; 3/7) = -\sum P_i \log_2(P_i) = 0.985$$

$$H(\text{Juegode}) = H(4/7; 3/7)$$

$$\hookrightarrow \text{Prom. Ponder} = \left(H(\text{Local}) \cdot \frac{4}{7} + H(\text{Vis}) \cdot \frac{3}{7} \right) = 0.857$$

$$GI(\text{Juegode}) = 0.128$$

$$H(3/4; 1/4) = 0.871 \quad H(4/3; 2/3) = 0.918$$

\downarrow Local \downarrow Local
With Lose

$$\underline{H(\text{mes})} \rightarrow \text{Prom. Ponder} = H(2/2; 0/2) \cdot \frac{2}{7} + H(4/2; 4/2) \cdot \frac{2}{7} + H(4/3; 2/3) \cdot \frac{3}{7}$$

$$= 0 + 1 \cdot \frac{2}{7} + 0.918 \cdot \frac{3}{7} = 0.679$$

$$GI(\text{mes}) = 0.306$$

$$H(\text{rwin}) \rightarrow H(4/3; 2/3) \cdot \frac{3}{7} + H(3/4; 1/4) \cdot \frac{4}{7} = 0.857$$

$$GI(\text{rwin}) = 0.128$$

Primer split \rightarrow Mes. $\rightarrow H = 0.679$

OCT \rightarrow GANA

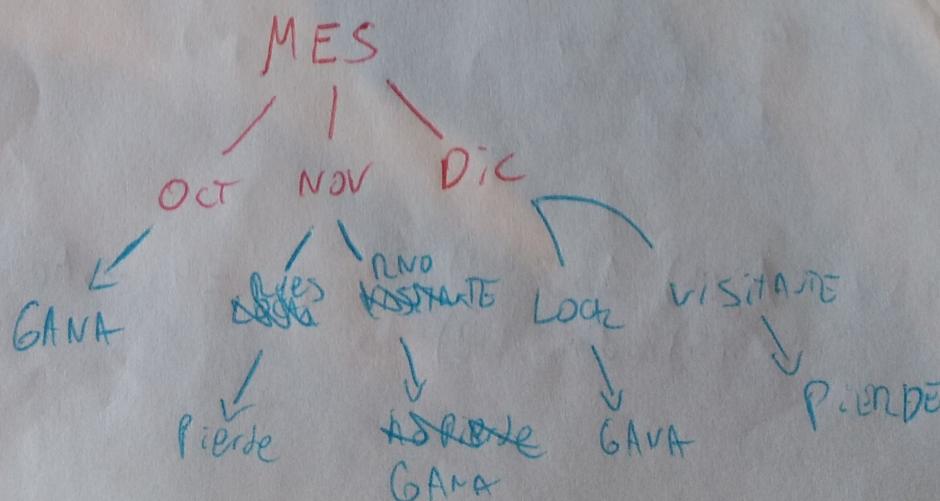
1

$$\text{NOV} \rightarrow GI(\text{Juegode}) = 0.679 - \left[\frac{2/2 \cdot H(4/2; 4/2)}{2/2} + 0 \right] = 0.679 - 0.321$$

$$\underline{GI(\text{rwin})} = 0.679 - \left[\frac{4/2 \cdot H(4/2; 4/2)}{4/2} + \frac{1/2 \cdot H(4/2; 4/2)}{1/2} \right] = 0.679 - 0.679$$

$$\text{DIC} \rightarrow GI(\text{Juegode}) = 0.679 - \left[\frac{1/3 \cdot H(0; 0)}{1/3} + \frac{2/3 \cdot H(0; 2/2)}{2/3} \right] = 0.679$$

$$\underline{GI(\text{rwin})} = 0.679 - \left[\frac{1/3 \cdot H(0; 2)}{1/3} + \frac{2/3 \cdot H(1/2; 1/2)}{2/3} \right] = \text{mas chico...}$$



(2)

A	4	2	5	2	2
B	2	3	4	1	5
C	5	1	4	2	⊗

$$\mu = 3$$

$$\text{Pr}_{\text{on}}(C) = 3$$

$$\text{Pr}_{\text{on}}(i_5) = 3.5$$

$\Gamma_{U,i}$ Por desviación: $\mu + \underbrace{\left(\text{Pr}_{(U)} - \mu \right)}_{\delta_U} + \underbrace{\left(\text{Pr}_{(i)} - \mu \right)}_{\delta_i} + \dots$

↓ usamos U periódico
Pr on probabilidad tensión

$$\Gamma_{C,15} = \frac{3 + (3-3) + (3.5-3)}{b=3.5} + \frac{\Gamma_{B,15}^{-b} S_{B,C} + \Gamma_{A,15}^{-b} S_{A,C}}{S_{B,C} + S_{A,C}}$$

Normalizo U-U (xF10)

						Norma	Sin C	$\rightarrow \frac{\langle X_f \rangle}{\langle X_f \rangle + \langle Y_f \rangle}$
1	-1	2	-1	-1	2.83	0.78		
-1	0	1	-2	2	3.16	0.10		
2	-2	1	-1	⊗	3.16	1		

$$\Gamma_{C,15} = 3.5 + \frac{(2-3.5)(0.78) + (5-3.5)(0.10)}{0.78+0.10} = 2.34 \approx 2$$

5

• Qué características tiene una red social?

Sí, relaciones.

- sección d
- Diametro bajo
 - Coeficiente de clustering alto
 - Caudal entre nodos bajo (Small world)
 - Comp. compleja grande
 - Distr. grados: Power-law

Comparo cuatro modelos

→ Erdos-Renyi: Nodos con prob de optener e

El diametro no sirve para comparar

↳ \log_{10}

→ Barabasi-Alberti: Nodos se crean por

preferencia de conectar

El diametro no sirve

↳ $\log_{10}(\log_{10}(N))$

los dist. de grado son power-laws con dep, sirven.

→ Watts Strogatz: Por cada e con $P_{add} = \frac{1}{2}$, intercambios aleatorios

busco q/ alto clust, bajo diam.

diam bajo ✓

alt clust ✓

6

→ OJO, no hay teletransportación

$$\begin{pmatrix} 0 & 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{6} \\ \frac{1}{6} \\ \frac{1}{6} \\ \frac{1}{6} \\ \frac{1}{6} \end{pmatrix} = t_1 = \begin{pmatrix} 0.28 \\ 0.29 \\ 0.11 \\ 0.19 \\ 0.19 \\ 0.028 \end{pmatrix}$$

NOTA

Por Jean-ew

r_0

Mejor $\rightarrow A$ (el que más se repite)
Peor $\rightarrow F$ (el que menos se repite)

Conectar F con A

(A \rightarrow F)

y hace mucho por venir de seguir
popu

Organización de Datos 75.06. Segundo Cuatrimestre de 2016. The one and only examen por Promoción.

Importante: Antes de empezar complete nombre y padrón en el recuadro. Lea bien todo lo que se requiere un mínimo de 60 puntos (60 puntos = 4). Este enunciado debe ser entregado ~~junto~~ con el parcial si quiere una copia del mismo puede bajarla del grupo de la materia. Si tiene dudas o consultas levante la mano, esta prohibido hablar desde el lugar, fumar o cualquier actividad que pueda molestar a los demás.

"Everything has to come to an end, sometime" (El Mago de Oz)

#	1	2	3	4	5	6	7	8	Entrega Hojas:
Corrección									Total:
Puntos	/15	/15	/15	/10	/10	/10	/10	/15	/100

Nombre:
Padrón:
Corregido por:

<p>1) Se usa el algoritmo Count-Min con 3 (tres) filtros de 8 (ocho) posiciones cada uno. El estado de los filtros es el siguiente:</p> <p>[0,2,4,2,0,0,3,5] [1,0,2,6,1,2,0,4] [0,0,3,3,3,2,4,1]</p> <p>Indicar cuáles de las siguientes afirmaciones son verdaderas justificando adecuadamente sus respuestas.</p> <p>a) La cantidad total de elementos del stream es 16. ✓ <i>A la lista array = 16</i> b) No puede tratarse de un stream con 16 elementos diferentes. c) Puede existir un elemento con frecuencia 5. F <i>El array 3 tiene total menor a 5</i> (***) (15 pts)</p>	<p>2) Dada la siguiente matriz de utilidad representando la calificación de usuarios y películas:</p> <table border="1" data-bbox="742 672 1389 986"> <thead> <tr> <th></th><th>U1</th><th>U2</th><th>U3</th><th>U4</th></tr> </thead> <tbody> <tr> <td>M1</td><td>5</td><td>4</td><td>1</td><td>4</td></tr> <tr> <td>M2</td><td>4</td><td>5</td><td>2</td><td>4</td></tr> <tr> <td>M3</td><td>1</td><td>4</td><td>4</td><td>?</td></tr> <tr> <td>M4</td><td>2</td><td>4</td><td>4</td><td>1</td></tr> </tbody> </table> <p>a) Normalizar la matriz restando a cada columna su promedio. Luego Estimar la calificación faltante usando collaborative-filtering user-user. (***)(10 pts) b) ¿Qué modalidad de collaborative-filtering genera resultados mas previsibles o conservadores? ¿User-User o Item-Item? Justifique. (****) (5 pts)</p>		U1	U2	U3	U4	M1	5	4	1	4	M2	4	5	2	4	M3	1	4	4	?	M4	2	4	4	1
	U1	U2	U3	U4																						
M1	5	4	1	4																						
M2	4	5	2	4																						
M3	1	4	4	?																						
M4	2	4	4	1																						
<p>3) En cada uno de los siguientes casos sugiera que algoritmo de clustering usaría: (***)</p> <p>a) 1000 puntos, 3 clusters de diferentes densidades y formas complejas. (5 pts) <i>Jerárquico y espacial</i> b) 1340 millones de puntos, 168 clusters de diferentes densidades y forma regular. (5 pts) c) 20.000 millones de puntos, no sabemos la cantidad de clusters, densidad variable y formas complejas. (5 pts)</p>	<p>4) Usamos SimRank para recomendar usuarios a seguir en twitter, sin embargo el algoritmo es demasiado "conservador", los usuarios piensan que las recomendaciones realizadas son muy obvias y por lo tanto no permite descubrir usuarios interesantes a seguir. Para hacer que nuestro algoritmo sea un poco mas audaz planteamos la posibilidad de aumentar o reducir el parámetro beta. ¿Cuál de las dos opciones haría que el algoritmo genere recomendaciones un poco mas interesantes? (****) (10 pts) <i>(1 - β) → tejer</i> <i>Bajar beta de vez en cuando se tejer</i></p>																									
<p>5) Dar un ejemplo de una Red en donde el nodo con mayor centralidad por PageRank tenga coeficiente de clustering 0. Analice si este tipo de red es probable en una Red social del mundo real. (10 pts) (**)</p>	<p>6) Se entrena un algoritmo de clasificación para determinar a que categoría corresponde una noticia. El algoritmo funciona muy mal con el set de test, todas las noticias que contienen la palabra "Alemania" son clasificadas en la categoría "conflictos bélicos" lo cual evidentemente no tiene sentido. ¿Qué está pasando? ¿Cuál sería su recomendación para solucionar el problema? (***)(10 pts)</p>																									
<p>7) ¿Cuál es la capital de Turkmenistán? <i>oetebhieopcidsnkhwiencwewjuelei</i></p>	<p>8) El campeonato mundial de ajedrez de 2016 entre Magnus Carlsen (noruega) y Sergey Karjakin (Rusia) lleva desarrolladas 6 partidas todas ellas resultaron en tablas. Diseñar un vocabulario RDF para representar esta información. (****) (15 pts)</p>																									

ES más estable (no contiene errores, tiempo/presión)

5
4
3
2
1

OJO!

Traspagos Por Comodidad
Mental

(el mejor tipo de
comodidad)

$$\frac{\langle u_i, v_i \rangle}{\|u_i\| \|v_i\|}$$

U	I	1	2	3	4
A		5	4	1	2
B		4	5	4	4
C		1	2	4	4
D		4	4	1	1

Norm
 \rightarrow
 $U - U$
(xP1a)

	2	1	-2	-1	$\ x\ $	$S_{ih}(u_i, v_i)$
	-0.25	0.75	-0.25	-0.25	3.16	0.65
	-1.75	-0.25	1.25	1.25	2.60	-0.78
	1	1		-2	2.45	1

¡decido usar los 2 users más similares: A, B!

$$R_{D,3} = R_{A,3} \cdot S_{AD}^{0.65} + R_{B,3} \cdot S_{BD}^{0.47}$$

$$\frac{S_{AD}^{0.65} + S_{BD}^{0.47}}{= 2.26 \approx [2]}$$

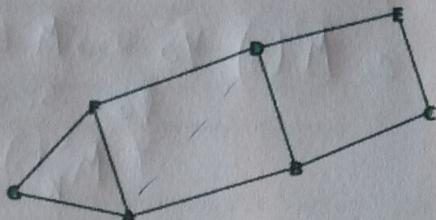
Importante: Antes de empezar complete nombre, padrón y dirección de mail en el recuadro. Lea bien todo el enunciado antes de empezar. Para aprobar se requiere un mínimo de 60 puntos (60 puntos = 4). Por favor resuelva los ejercicios en los espacios destinados para esto siempre que sea posible. Si desea agregar hojas adicionales puede hacerlo al final. Una copia del enunciado estará disponible para ser bajada del grupo de la materia. Si tiene dudas o consultas levante la mano, está prohibido hablar desde el lugar, fumar o cualquier actividad que pueda molestar a los demás.

"Your focus determines your reality." - Qui-Gon Jinn

#	1	2	3	4	5	6	7	8	Entrega Hojas:
Corrección									Total:
Puntos	/15	/15	/15	/15	/10	/15	/10	/5	/100

Nombre:
Padrón:
Email:
Corregido por:

1) Dado el siguiente grafo representando una Red Social



Indicar:

- a) De acuerdo al modelo de Preferential attachment, cuál o cuáles son las aristas con mayor probabilidad de agregarse a la red.
 b) ¿Cuál es el coeficiente de clustering promedio de la red?
 c) ¿Cuáles son los nodos con mayor betweenness (sin hacer cuentas)?
 (***) (15 pts) $\theta \text{ y } D$

agregar a quien
más tenga. The Rich Get
Richer

$C=2$

2) Dados los puntos:

	orig	$\vec{t} + \vec{L}: (1,1,1)$	$\vec{t} + \vec{2}: (2,5; 0; 0)$	$(2, -2, -2)$	
$X_1 = (-2, -2)$ \rightarrow Clase +1	$1, 2, 2$	$2, 0$	$-1 \quad X$	$1 \quad \checkmark$	
$X_2 = (0, -2)$ \rightarrow Clase +1	$1, 0, -2$	$2, 0$	$-3 + 2 \quad \checkmark$	$2 \quad \checkmark$	
$X_3 = (4, 4)$ \rightarrow Clase -1	$1, 4, 4$	$-2, 0$	$-1 \quad X$	$-1 \quad \checkmark$	
$X_4 = (4, 5)$ \rightarrow Clase -1	$1, 4, 5$	$-1, 0$		$-2 \quad \checkmark$	

$\text{Pref}(x) < 0 \text{ si } w^T x > 0$
 $\text{Pref}(x) < 0 \text{ si } w^T x \leq 0$

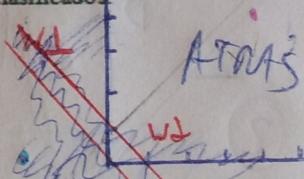
$$W_1 = (1, 1, 1) + 0.5(2, -2, -2) \quad W_2 = (2, 5; 0, 0) + 0.5(1, 4, 4) - 1$$

a) Entrenar un perceptrón utilizando $W_0 = (1, 1, 1)$ y $\alpha = 0.5$, hasta que todos los puntos queden correctamente clasificados.

b) Graficar los puntos y la separación de clases encontrada en base al vector de pesos resultante.

c) Indicar si existe alguna mejor separación de clases y, en caso afirmativo, indicar cómo podría encontrarla.

(***) (15 pts) Mejor? Clasif. doble bien.
 Pero se podría hacer SVM...



3) Dados los siguientes streams:

E A B D C A E C B D \rightarrow L_1

E A E B C E E D B E \rightarrow L_2

Estimar el número sorpresa utilizando AMS con 5 estimadores, y explicar qué conclusiones pueden obtenerse de la comparación de los resultados para ambos streams. (**) (15pts)

4) Realizar topic rank del tópico 1 del siguiente grafo:

A pertenece al tópico 1 y linka con B, C y E

$$M_0 = \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ \frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} \\ \frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 2 & \frac{1}{2} \end{pmatrix}$$

B pertenece al tópico 1 y linka con A y E

C pertenece al tópico 2 y linka con A, C y D

D pertenece al tópico 1 y linka con E

E pertenece al tópico 2 y no tiene links

$$M_0 = BM_0 + (1-B) \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Considerar beta=0.3, Realizar 3 iteraciones y rankear los resultados. Determinar si puede considerar que los resultados son adecuados y por qué
 (***)(15 pts)

$$M_0 = \begin{pmatrix} 0.23 & 0.58 & 0.46 & 0.23 & 0.32 \\ 0.16 & 0.00 & 0.23 & 0.23 & 0.37 \\ 0.23 & 0.46 & 0.46 & 0.23 & 0.37 \\ 0.23 & 0.58 & 0.46 & 0.23 & 0.32 \\ 0.23 & 0.22 & 0.83 & 0.73 & 0.24 \end{pmatrix}$$

$$\begin{pmatrix} 0.23 & 0.38 & 0.33 & 0.23 & 0.29 \\ 0.33 & 0.23 & 0.23 & 0.23 & 0.24 \\ 0.1 & 0 & 0.1 & 0 & 0.06 \\ 0.23 & 0.23 & 0.33 & 0.23 & 0.29 \\ 0.1 & 0.15 & 0 & 0.3 & 0.06 \end{pmatrix}$$

No. Beta muy

Chico!

$$T_1 = M_0 \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \end{pmatrix}$$

$$R_1 = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

③ EABDCAECBD

$$r_{real} = 2^2 + 2^2 + 2^2 + 2^2 + 2^2 = 20$$

Estimador ← Letra
correspondiente

E, L2	$M_2^2(S) = (2 \times 10) - 1)h$
A, L2	$M_A^2(S) = (2, 2 - 1) \cdot 10 = 30$
B, L2	$M_B^2(S) = 30$
D, L2	$M_D^2(S) = 30$
C, L2	$M_C^2(S) = 30$

$$\begin{aligned} M_2^2(S) &= 30 \\ M_A^2(S) &= 30 \\ M_B^2(S) &= 30 \\ M_D^2(S) &= 30 \\ M_C^2(S) &= 30 \end{aligned}$$

↓ Prom

$$M_2^2(S_{\text{real}}) = \text{Mediana}([30, 30, 30, 30, 30]) = 30$$

(Notación es $M^2(S)$, Yo uso $M^2(S)$)

para screen

y $M_K^2(S)$ para estimar

E A E B CEE DBE

$$r_{real} = 1^2 + 2^2 + 1^2 + 1^2 + 5^2 = 32$$

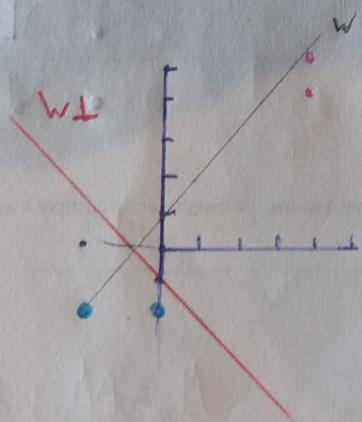
$$E, L2 \times 4S \rightarrow M_E^2(S) = 10(2.5 - 1) = 90$$

$$\begin{aligned} A, L & M_A^2(S) = 10 \\ B, L2 & M_B^2(S) = 30 \\ C, L & M_C^2(S) = 10 \\ D, L & M_D^2(S) = 10 \end{aligned}$$

$$\begin{aligned} M^2(S) &= \text{Med}([10, 10, 10, 30, 90]) \\ &= 30 \end{aligned}$$

prom

②



$$w = 1, -2, -2 \rightarrow y = -\frac{3}{2}x + 1$$

$$w \perp = -1, 2, -2 \rightarrow y = \frac{3}{2}x - 1$$

5) Sea la siguiente matriz de utilidad entre usuarios (números romanos) y libros (letras):

	A	B	C	D	E	F
A	3	0	5	0	3	3
B	0	5	4	0	2	4
C	5	4	3	5	2	2
D	0	3	5	2	2	5
E	3	2	2	2	5	5
F	3	4	4	5	5	5

NOTA:
SIM
OK

Realizando colaborative filtering de tipo item-item determinar que libro se le debe recomendar en primer lugar al usuario II. Tomar para la recomendación los 2 libros mas semejantes. (*) (10pts)

6) Dado el siguiente conjunto de documentos y sus correspondientes categorías:

D1('nintendo'): "super mario luigi peach toad world"

D2('capcom'): "chunli ken ryu super world edition"

D3('nintendo'): "super mario world"

D4('capcom'): "chunli ken bison super"

D5('nintendo'): "mario super world"

D6('capcom'): "super street fighter ultra"

$$P(N|D) = \frac{3}{6} \cdot \left(\frac{3+1}{12+14} \cdot \frac{3+1}{12+14} \cdot \frac{0+1}{12+15} \right) = 0.172$$

$$P(C|D) = \frac{3}{6} \cdot \left(\frac{3+1}{14+14} \cdot \frac{1+1}{14+14} \cdot \frac{0+1}{14+15} \right) = 0.124$$

Construir un clasificador naive bayes paso a paso e indicar dado el documento "super world yoshi" el score de que el mismo pertenezca a la categoría 'nintendo' o a la categoría 'capcom'. ¿Cuál es la categoría que indicará el clasificador para el documento? (*****) (15 pts)

7) Dados los siguientes puntos: (1,2) (1,3) (1,5) (2,3) (2,8) (3,3) (4,1) (4,9) (7,8) (9,9) (12,2) (13,4)

Agrupar utilizando Clustering Jerárquico, indicando la cantidad de clusters resultantes y el criterio utilizado para definir dicho número. Representar el dendrograma mostrando como se agrupan los distintos elementos en los clusters obtenidos. (**) (10pts)

8) Completar las fases del algoritmo mapper:

1) Aplicar un lente topológico a los datos.

2)

3)

4) Generar un grafo en donde cada _____ es un nodo y las aristas representan _____

(*) (5pts)

5) Normalizar I-I (x C0/S) NO CREO QUE ESTE BIEN ESTO.

-0,3	0,3	0,5		
0,6	-0,7	0,5	-0,5	-0,7
1,7	0,3	1,5	.	.
0	.	.	0,3	
-1,3	0	-1,5	0,3	
$\ X\ $	2,16	0	0.81	2.12

Sih
coh A

1 0,24

-0,1 -0,22

OJO:

SOLO BUSCO
Semejanzas de
los yd catif
(C,E,F) con
los no catif (A,B,D)

$$r_{II,A} = r_{II,C} \cdot S_{AC} + r_{II,E} \cdot S_{AE} = \frac{5}{S_{AC} + S_{AE}} = 5$$

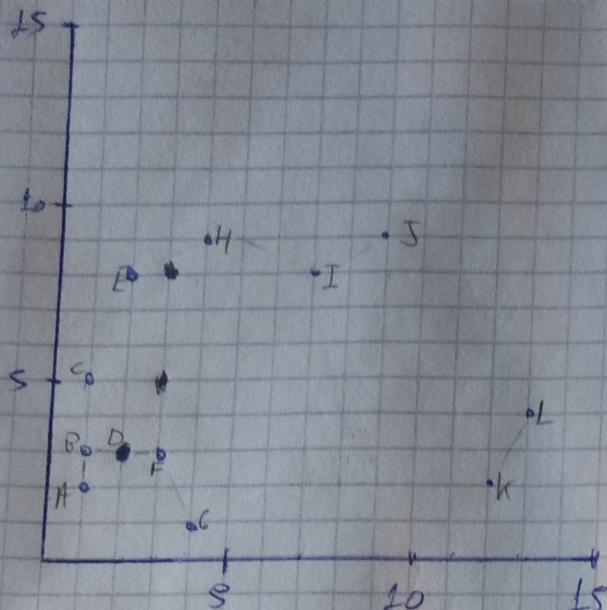
Sih
coh B

1 0,26 1 0,26

$$r_{II,B} = r_{II,F} \cdot S_{AF} = \frac{0}{S_{AF}} = 0$$

NOTA

$$r_{II,D} > r_{II,C} \cdot S_{CD} + r_{II,F} \cdot S_{DF} = \frac{4}{S_{CD} + S_{DF}}$$



Clustering jerárquico

tomando distancias

entre clusters

como la min distancia

entre los cuales tienen

puntos de ellos.

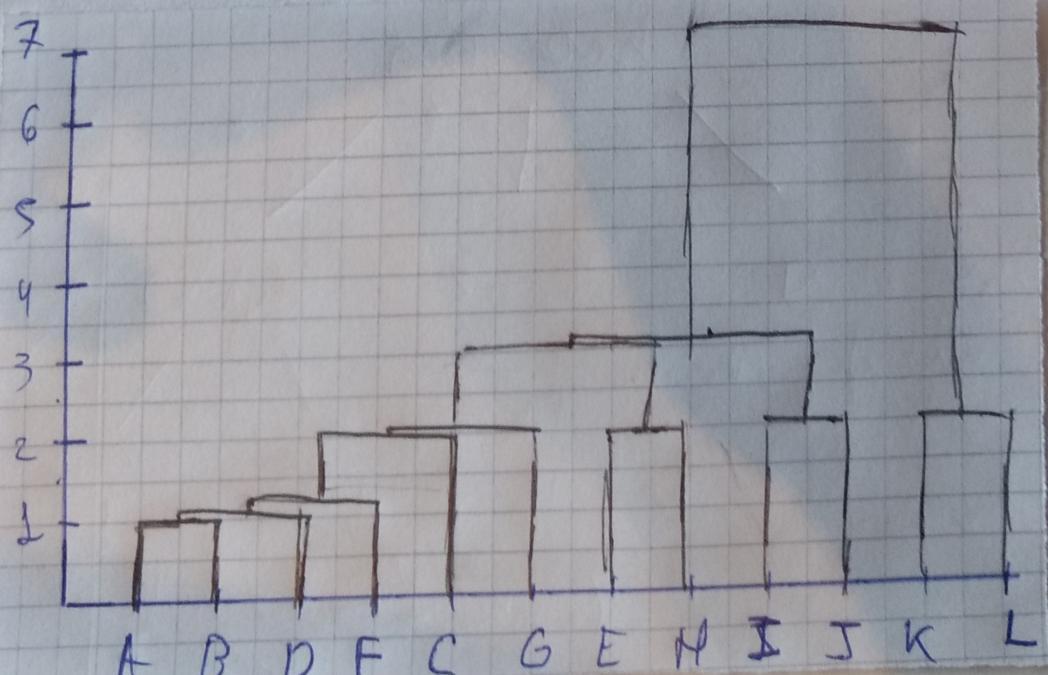
Si hoy en plante, se empieza
la menor suma de puntos

$$(e_i = 2,4 \times \frac{3,4}{1,4} \checkmark)$$

$$\begin{aligned}
 O &\rightarrow (A, B) \rightarrow (A, B, D) \rightarrow (A, B, D, F) \rightarrow (A, B, D, F, C) \\
 &\rightarrow (A, B, D, F, C, G) \rightarrow (A, B, D, F, C, G, E, H) \\
 &\rightarrow (A, B, D, F, C, G, E, H) (I, J) \rightarrow (A, B, D, F, C, G, E, H, I, J, K, L) \\
 &\rightarrow (A, B, D, F, C, G, E, H, I, J) (K, L) \xrightarrow{\text{Posibles Soluciones}} \\
 &\rightarrow (A, B, D, F, C, G, E, H, I, J, K, L)
 \end{aligned}$$

Si hoy en plante, se empieza
el menor de los
que tienen mas puntos

$$\begin{array}{l}
 \cancel{(A, B, D, F, C, G, E, H)} \\
 \cancel{(I, J, K, L)} \\
 \cancel{(A, B, D, F, C, G, E, H, I, J)} \\
 \cancel{(K, L)}
 \end{array}$$



Importante: Antes de empezar complete nombre y padrón en el recuadro. I
mínimo de 60 puntos (60 puntos = 4) y tener un 50% de los puntos
entregado junto con el parcial si quiere una copia del mismo puede bajar
entregado junto con el parcial si quiere una copia del mismo puede bajar
prohibido hablar desde el lugar, fumar o cualquier actividad que pueda
aprovecharse para aprobar.

lanciado antes de empezar. Para aprobar se requiere un
menos 4 de los 7 ejercicios. Este enunciado debe ser
en la materia. Si tiene dudas o consultas levante la mano, esta
más. Los exámenes que no resuelven puntos suficientes para
aprovecharse para aprobar.

"Nobody panics when things go according to plan. Even if... it is horrifying" - The Joker, The Dark Knight

#	1	2	3	4	5	6	7	8	Entrega Hojas:
Corrección									Total:
Puntos	/15	/15	/15	/15	/10	/15	/15	/10	/100+10

Nombre:
Padrón:
Corregido por:

1) Dada la siguiente matriz de links M:

	A	B	C	D	E	F
A						
B	1/3		1		1	
C	1/3					
D		1/2				
E		1/2		1/2		
F	1/3			1/2		

- a. Dibujar el grafo dirigido asociado a la misma indicando el peso de las aristas. A partir del mismo indicar todos los problemas que hacen que no podamos aplicar PageRank SIN Teletransportación sobre este. (8pts)
b. Indicar paso a paso cómo construir una única matriz de PageRank que incluya la teletransportación, considerando un beta de 0.85. (7pts)

(***)

2) Dados los siguientes datos:

Persona	Estudios	Trabaja?	Estado Civil	Préstamo?
1	Universitarios	SI	Casado	SI
2	Primarios	SI	Soltero	NO
3	Universitarios	SI	Casado	SI
4	Universitarios	NO	Casado	NO
5	Primarios	SI	Divorciado	SI
6	Secundarios	NO	Divorciado	NO
7	Universitarios	SI	Divorciado	SI
8	Universitarios	SI	Soltero	NO
9	Secundarios	NO	Divorciado	NO
10	Secundarios	SI	Soltero	SI
11	Universitarios	SI	Soltero	NO
12	Primarios	SI	Casado	SI

a. Armar un árbol de decisión seleccionando en cada split el atributo que mayor ganancia de información nos da. (8pts) *Arbol de decisión*

b. Explicar cómo hacer para evitar el overfitting y aplicarlo al árbol del punto anterior. (**) (7 pts)

3) Se desea aplicar clustering

espectral sobre el siguiente grafo. Los 4 primeros autovalores de la matriz laplaciana (ordenados de menor a mayor) son:

$$\lambda_1=0.2, \lambda_2=0.27, \lambda_3=0.55 \text{ y } \lambda_4=3$$

y sus autovectores:

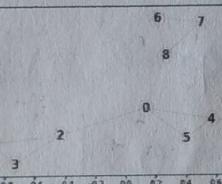
$$v_1=[-0.33, -0.33, -0.33, -0.33, -0.33, -0.33]$$

$$v_{2,3}=[0.00, -0.44, -0.33, -0.44, 0.00, 0.00, 0.44, 0.44, 0.33]$$

$$v_4=[0.25, -0.28, -0.12, -0.28, 0.55, 0.55, -0.28, -0.28, -0.12]$$

$$v_5=[-0.07, 0.72, -0.07, -0.66, 0.15, -0.09, 0.03, 0.03, -0.07]$$

Determinar cuántos y cuáles autovectores utilizar para detectar las 3 comunidades del grafo. Justificar (***). (15 pts)



4) Spotify registra con 1 o 0 si al usuario le ha gustado o no una canción. Tenemos la siguiente matriz en donde conocemos los gustos de los usuarios A, B, C, D y E para 6 canciones. Se pide estimar si al usuario "A" le van a gustar o no las canciones 5 y 6 usando collaborative-filtering de tipo user-user tomando los 2 vecinos más cercanos para realizar la estimación

A 1 0 0 1 ??
B 0 1 1 0 1 0
C 1 1 0 0 0 1
D 1 0 1 1 1 0
E 1 0 0 1 0 1

(***) (15 pts)

5) a. Construir un filtro de bloom de 16 bits ($m=16$) que contenga los caracteres C y D, para el universo de 5 ($n=5$) caracteres A, B, C, D, E considerando las siguientes funciones h_1 y h_2 . Tener en cuenta que la probabilidad de falso positivo para la construcción es de 0.2.

	A	B	C	D	E
H1	13	8	4	4	3
H2	15	5	9	15	7

Una vez creado, explicar cómo se realiza la resolución con el filtro de las consultas A, B y D, indicando que se infiere usando el mismo y por qué.

(5pts) *A, esto es verdad? No.*

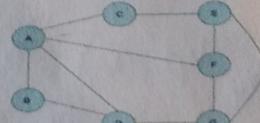
b. Sabiendo que el m que se utilizó es óptimo, indicar si la cantidad de funciones de hashing utilizadas es óptima, justificando su resolución. (5 pts)

(**) *B, m. D, 5, puede ser, o no*

6) Dados los nodos A, B, C, D, E, F y G, y las aristas A-B, A-C, A-D, A-F, B-D, C-E, D-G, E-F, F-G, E-G.

- a. Muestre la distribución de grados. (3pts)
b. Indique cuál es el diámetro de la red. (3pts)
c. Calcule el coeficiente de clustering promedio. (4pts)
d. Compare dichos valores con los valores característicos de una red social y analice las semejanzas y diferencias. (5pts)

(***)



Notas: Puede haber en este hoja

Diametro:
Ley de los 2/3:
Koltz = $\frac{m}{n} \log_2(2)$
Koltz = 0.9682

B-A-C-E

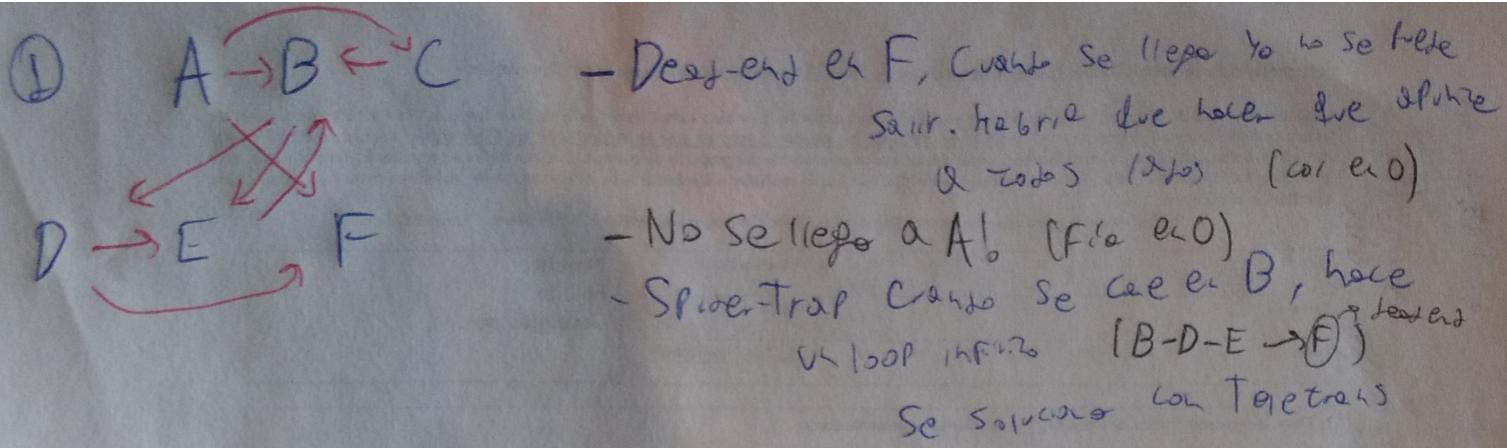
3

7) Partimos de imágenes RGB de 20x20 pixeles que queremos clasificar entre 3 clases posibles con una red neuronal convolucional, aplicamos una capa de convolución de tipo "same" con 5 filtros de 3x3. Luego aplicamos Max Pooling de 4x4 con stride=4. Luego una capa fully-connected y finalmente una capa softmax. (****)

- a. Diagramar el modelo de red. ¿Cuántas neuronas tienen las dos últimas capas del modelo? (5pts)
b. ¿Cuántos parámetros debemos entrenar en total? (10pts)

8) Bonus: Explique en qué consiste el mecanismo conocido como "transfer learning". (**) (10 pts)

- Distri \rightarrow Power Law X
-Diametro \rightarrow Bajo \downarrow (cuanto es menor, mejor)
-Geo Clust \rightarrow Alto X



$$M_{\text{can}} = \beta M + (1-\beta) \begin{pmatrix} 1/6 & & \\ & \ddots & \\ & & 1/6 \end{pmatrix}$$

$H(\text{presario}) = H(6/12; 6/12) = 1 \rightarrow$ equidistribución

$$H(\text{escudos}) = H(\text{cuer}) \cdot 6/12 + H(\text{pri}) \cdot 3/12 + H(\text{sec}) \cdot 3/12 = 0.96$$

$$\downarrow \qquad \qquad \qquad \downarrow \qquad \qquad \qquad \downarrow$$

$$H(3/6; 3/6) = 1 \quad | \quad H(1/3; 4/3) = 0.92 \quad | \quad H(4/3; 2/3) = 0.92$$

$$H(\text{trabajos}) = 9/12 \cdot H(6/0; 3/0) + 3/12 H(0; 3/3) = 0.69$$

$$H(\text{escudos}) = 4/12 \cdot H(3/4; 3/4) + 4/12 H(4/4; 3/4) + 4/12 H(2/4; 2/4) = 0.87$$

$$GI(\text{actr}) = H(\text{presario}) - H(\text{actr}) \rightarrow \text{Mayor GI} = \text{Escritor}$$

CTSAO

$$b_3 H(\text{escudos}) = 3/4 \cdot H(2/3; 4/3) + 1/4 H(2; 0) + 0 = 0.69$$

$$H(\text{trabajos}) = 3/4 H(3/3; 0) + 1/4 (0; 4) = 0 \rightarrow \text{SPLIT}$$

Soltero

$$b_3 H(\text{escudos}) = 3/4 H(0; 2) + 1/4 H(0; 2) + 1/4 H(2; 0) = 0 \rightarrow \text{SPLIT}$$

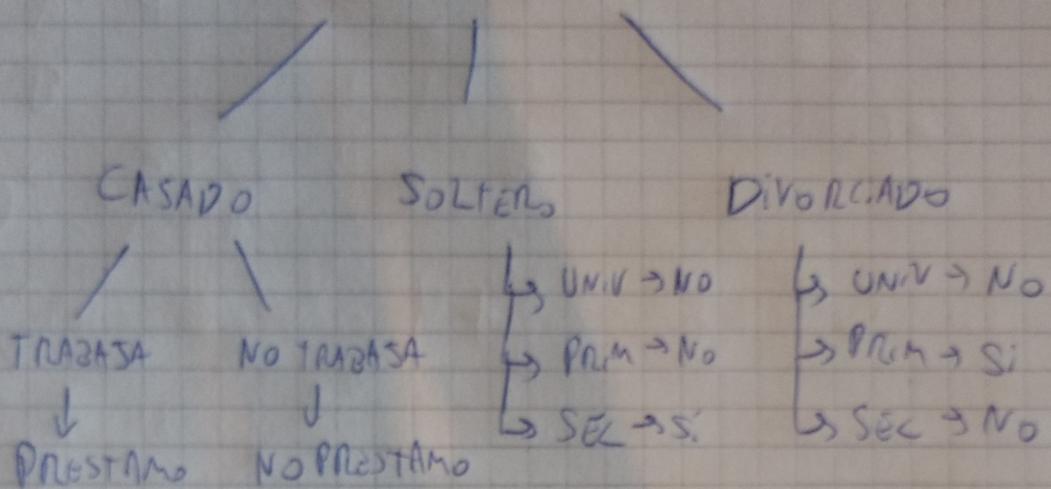
$$H(\text{trabajos}) = \text{No vale la pena.}$$

Divorcado

$$b_3 H(\text{escudos}) = 1/4 H(0; 2) + 1/4 H(2; 0) + 2/4 H(0; 2) = 0$$

$$H(\text{trabajos}) = \text{No vale la pena}$$

Deco Mayor Presccho?
CONSTRUCCIÓN



(4)

A 1 0 0 1 0
 B 0 1 1 0 1 0
 C 1 1 0 0 0 1
 D 1 0 1 1 1 0
 E 1 0 0 1 0 1

SISTEMA

Morganento
~~(-1)~~
~~(+1)~~

$\frac{1}{2} - \frac{1}{2} - \frac{1}{2} \frac{1}{2} \star \star$
 $-\frac{3}{6} \frac{3}{6} \frac{3}{6} - \frac{3}{6} \frac{3}{5} \frac{3}{5}$
 $\frac{2}{5} \frac{4}{5} - \frac{3}{5} \frac{1}{5} \frac{1}{5} \frac{2}{5}$

↓ Norma, 20 U-U
 (XF(1))

	NORMA	SIM. CONT.
$\frac{1}{2} - \frac{1}{2} - \frac{1}{2} \frac{1}{2} \star \star$	1	1
$-\frac{1}{2} \frac{1}{2} \frac{1}{2} - \frac{1}{2} \frac{1}{2} - \frac{1}{2}$	1,22	-0,82
$\frac{1}{2} \frac{1}{2} - \frac{1}{2} - \frac{1}{2} \frac{1}{2} \frac{1}{2}$	1,22	0
$\frac{1}{3} - \frac{1}{3} \frac{1}{3} \frac{1}{3} \frac{1}{3} - \frac{1}{3}$	1,15	0,43 -
$\frac{1}{2} - \frac{1}{2} - \frac{1}{2} \frac{1}{2} \frac{1}{2} \frac{1}{2}$	1,22	0,82 -

$$r_{AS} = r_{0,5} S_{AD} + r_{0,5} \cdot S_{AE} = \frac{0,344 \rightarrow 0}{0,656 \rightarrow 1}$$

$$r_{A,G} = r_{0,5} S_{AD} + r_{0,5} S_{AE} = 0,256 \rightarrow 1$$