Organización de Datos 75.06. Segundo Cuatrimestre de 2020. Examen por promoción

Importante: Lea bien todo el enunciado antes de empezar. Para aprobar se requiere un mínimo de 60 puntos (60 puntos = 4). Si tiene dudas o consulta estaremos disponibles vía meet, pero tengan en cuenta que solo se contestarán dudas de enunciado, y no deben compartir por esa vía nada relacionado con la resolución. Está prohibido realizar cualquier actividad que pueda molestar a los demás. El criterio de corrección de este examen estará disponible en gradescope.

"It doesn't matter what we want. Once we get it, then we want something else ." — Lord Baelish – Game of Thrones

#	1	2	3	4	5	Entrega Hojas:
Corrección						Total:
Puntos	/20	/20	/20	/20	/20	/100

Nombre:
Padrón:
Corregido por:

1) Sean los siguientes puntos en dos dimensiones: x1=(0,0), x2=(8,0), x3=(16,0), x4=(0,6), x5=(8,6), x6=(16,6). Sobre estos puntos queremos usar K-Means, con la distancia euclidea, para encontrar 3 clusters. Sabemos que el resultado del algoritmo depende de la forma en la que elijamos los tres centroides iniciales. En base a esto responder:

a) ¿Cuántas configuraciones iniciales posibles existen? (5 pts)
b) ¿Cuántas clusterizaciones finales son posibles? (es decir aquellas clusterizaciones en las cuáles K-Means ya ha convergido). (5 pts)
c) ¿Cuál es el máximo número de pasos que podemos necesitar desde cualquiera de las inicializaciones iniciales posibles hasta cualquiera de las posibles clusterizaciones.? (10 pts)

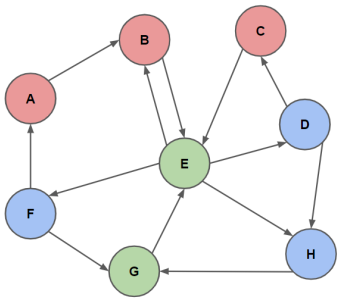
2) Dada la siguiente matriz que representa si los usuarios 1 a 6 les gustaron o no las series A-F:

		Items					
		A	B	C	D	E	F
Usuarios	1	Si	No		Si	Si	
	2	No			No	Si	Si
	3		No	No	Si		Si
	4	No		Si	No	No	
	5	Si	No	Si			Si
	6	No	Si	No		No	Si

Usar la semejanza de Jaccard y collaborative filtering user-user con k (cantidad de vecinos) = 3 para estimar la probabilidad de que al usuario 6 le guste la serie " D".

(20pts)

3) Dadas las siguientes páginas, sabemos que las mismas fueron asociadas con tres tópicos distintos (A, B, C en rojo para educación, D, F, H en azul relacionadas con economía y E, G en verde asociadas con tecnología), utilizar topic rank para calcular el ranking de cada una, sabiendo que se quiere favorecer las páginas relacionadas con economía, indicadas en el grafo con el color azul. (20 pts)



4) Sabemos que tenemos una construcción de Count-Min de 3 filtros (F1, F2 y F3) de los cuales conocemos dos

F1 = [4,0,8,0,0,0,5,3]

F2 = [0,0,3,3,4,4,6,0]

Dados los siguientes candidatos para F3, cuál es el correcto y por que. (5 pts)

A = [5,0,1,3,2,4,4,1]

B = [5,4,1,0,1,6,0,4]

Considerando el F3 válido, cuál sería el resultado de la función de hashing para lograr la mayor estimación posible con los filtros F1, F2 y F3 para un cierto elemento "X" y cuál sería su resultado. (15 pts)

5) Tenemos información sobre precios de venta de propiedades en el país:

(fecha, tipo_propiedad, m2_totales, m2_cubiertos, habitaciones, dormitorios, baños, cocheras, provincia, ciudad, poblacion_ciudad, cant_escuelas_cercanas, servicios, estado_propiedad, antiguedad, precio)

Se quiere generar un modelo de predicción del precio de las propiedades usando XGBoost.

a) Indique el proceso de feature engineering que realizaría, indicando el detalle de las features que utilizaria, las tranformaciones que realizaria a cada columna y cuales datos agregaria. Indique una fila completa del set de datos final.

b) Indique qué cambios debería realizar al feature engineering si el modelo a utilizar es una red neuronal.