Mini-Control 3

CC5213 – Recuperación de Información Multimedia
Profesor: Juan Manuel Barrios
Fecha entrega: 30 de noviembre de 2023

Debe resolver los siguientes ejercicios en un documento o en papel. No es necesario programar ni debe entregar código fuente.

Pregunta 1 (TF-IDF, semana 09)

Se tiene un dataset con **10.000** documentos de texto. A continuación, se muestran dos documentos de ese dataset:

D_1 = Compré una casa nueva en Constitución

D_2 = En mi casa no quieren una nueva constitución

La siguiente tabla muestra un resumen de las palabras en el dataset:

Id	Palabra	Documentos en los que aparece
1	casa	1.000
2	compré	1.000
3	constitución	100
4	en	1.000
5	mi	100
6	no	100
7	nueva	10
8	quieren	100
9	una	10

Suponga que un usuario escribe la siguiente consulta de texto:

Q = Una nueva constitución

Utilizando el modelo Bag-of-Words y TF-IDF deberá calcular la similitud de la consulta **Q** con ambos documentos $\mathbf{D_1}$ y $\mathbf{D_2}$ y señalar el documento más similar. Específicamente realice los siguientes pasos:

a. Calcule el IDF para el vocabulario relevante.

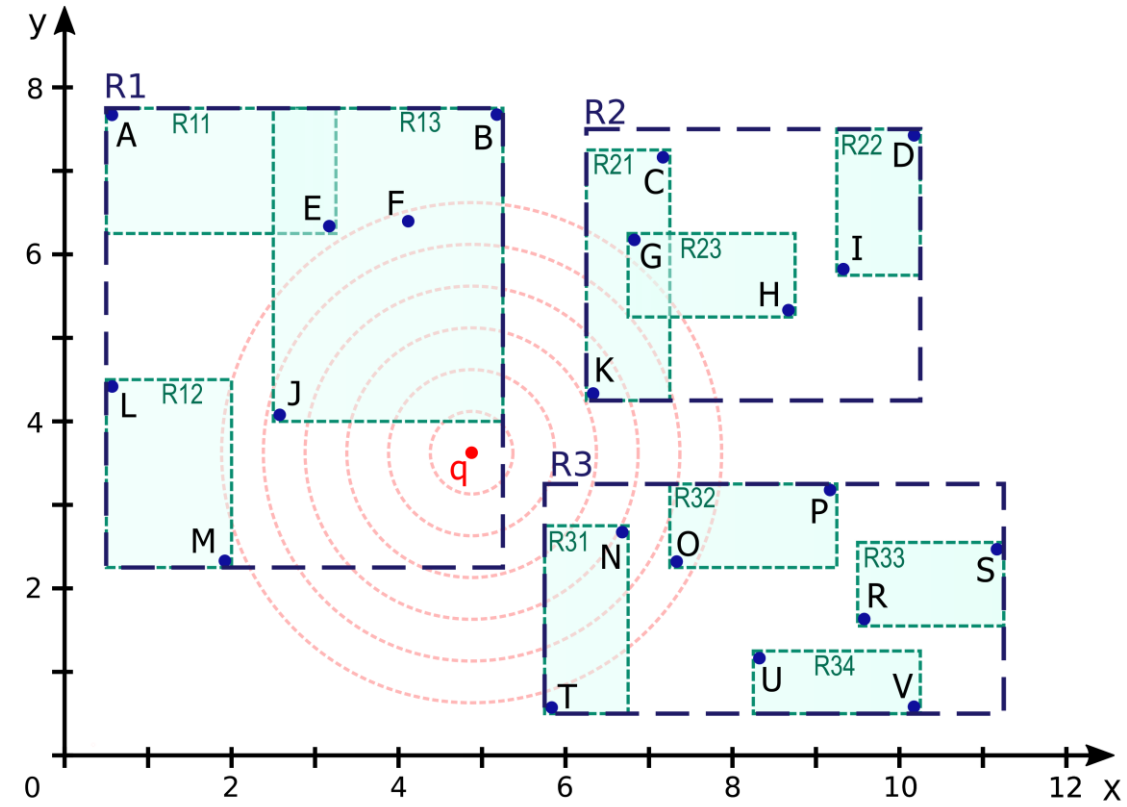
b. Calcule el vector TF-IDF normalizado de $\mathbf{D_1}$, $\mathbf{D_2}$ y **Q**.

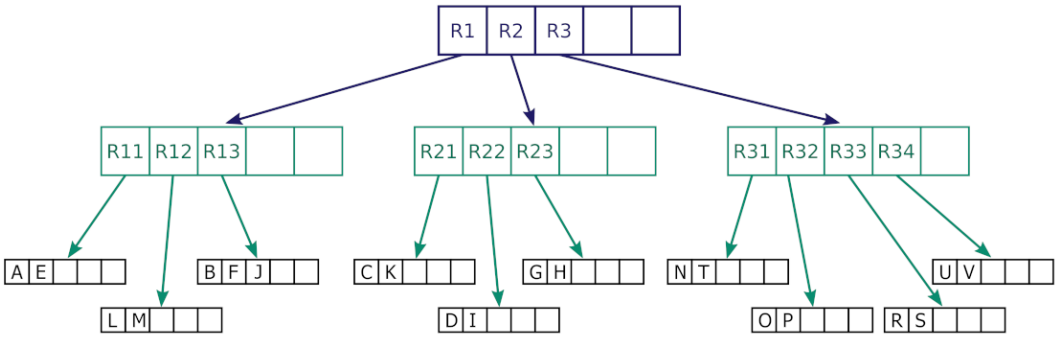
c. Calcule la similitud coseno entre **Q**, $\mathbf{D_1}$ y $\mathbf{D_2}$ y señale el documento más similar a **Q**.

Utilice las fórmulas vistas en el curso (“Slides 09.1-Descriptores de texto-TF-IDF.pdf“) usando **logaritmos en base 10**.

Pregunta 2 (R-Tree, semana 12)

Se tiene un conjunto de 21 vectores de dos dimensiones. Los vectores fueron indexados por un R-Tree según el siguiente diagrama:





Para el objeto **q** de la figura anterior se desea buscar los **dos vecinos más cercanos (2-NN)** según distancia Euclidiana. Como ayuda visual, se muestran varios círculos concéntricos a **q**.

Explique cómo funcionará el algoritmo de **búsqueda k-nn por prioridad** para la búsqueda **2-NN** de **q**. Específicamente señale:

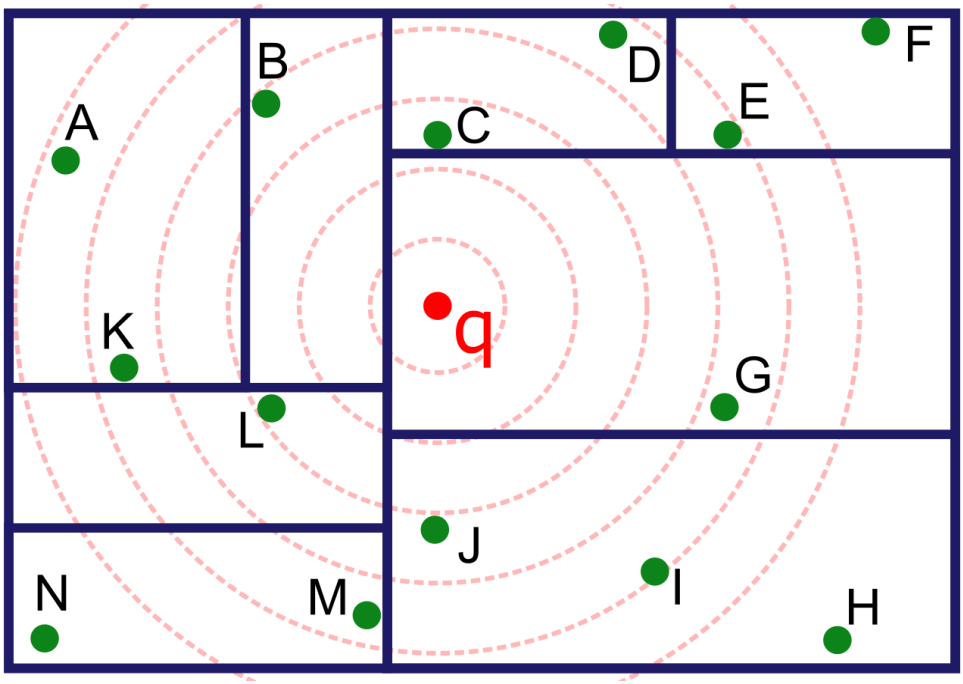
a. Las regiones espaciales visitadas, el orden en que se visitan y la evolución de los candidatos durante la búsqueda de los 2-nn.

b. Señale el número de veces que se evaluó alguna función de distancia para poder resolver la búsqueda 2-NN de q.

c. Señale el número de veces que se evaluará una función de distancia si se resuelve la misma búsqueda anterior pero utilizando el algoritmo linear scan.

Pregunta 3 (k-d tree, semana 12)

En la siguiente figura se muestran 14 vectores de dos dimensiones. Los vectores fueron indexados por un k-d tree cuyos nodos hoja corresponden a las regiones que se muestran en la figura.



Se desea realizar una **búsqueda aproximada** del vecino más cercano a **q** según distancia Euclidiana. El parámetro ***c*** corresponde al número máximo de nodos hoja a visitar durante la búsqueda aproximada.

Señale el nombre del vector que se encontrará como vecino más cercano si:

a. La búsqueda del NN se restringe a **c** = 1
b. La búsqueda del NN se restringe a **c** = 2
c. La búsqueda del NN se restringe a **c** = 3
d. La búsqueda del NN se restringe a **c** = 4
e. La búsqueda del NN se restringe a **c** = 5

Como ayuda visual, se muestran varios círculos concéntricos a **q**.

Pregunta 4 (PCA, semana 13)

Se tiene el conjunto **R** de 9 imágenes donde a cada imagen se le calculó un descriptor de contenido de 4 dimensiones. El promedio de los descriptores de contenido de **R** es:

$$\bar{x} = \begin{pmatrix} 5 \\ 2 \\ 3 \\ 9 \end{pmatrix}$$

Llamaremos **S** al conjunto de descriptores que se obtiene de restar $\bar{x}$ de los descriptores de **R**, es decir:

$$S = \{y \,|\, \forall x \in R,\ y = x - \bar{x}\}$$

Notar que **S** es un conjunto de descriptores centrados (promedio cero).

Al calcular las covarianzas entre las coordenadas de los vectores de **S** se obtiene la siguiente matriz de covarianza ***C***:

$$C = \begin{pmatrix} 14.3 & 3.6 & 6.7 & -2.0 \\ 3.6 & 18.8 & 1.5 & 0.6 \\ 6.7 & 1.5 & 17.1 & -2.2 \\ -2.0 & 0.6 & -2.2 & 30.3 \end{pmatrix}$$

Los 4 valores propios de ***C*** son los siguientes:

$$\lambda_1 = 8.5 \quad \lambda_2 = 31.4 \quad \lambda_3 = 23.9 \quad \lambda_4 = 16.8$$

Los 4 vectores propios asociados a cada valor propio de ***C*** son:

$$v_1 = \begin{pmatrix} -0.8 \\ 0.2 \\ -0.5 \\ -0.2 \end{pmatrix} \quad v_2 = \begin{pmatrix} 0.2 \\ 0 \\ -0.6 \\ 0.8 \end{pmatrix} \quad v_3 = \begin{pmatrix} 0.6 \\ 0.3 \\ -0.6 \\ -0.5 \end{pmatrix} \quad v_4 = \begin{pmatrix} 0 \\ -0.9 \\ -0.3 \\ -0.2 \end{pmatrix}$$

Se desea utilizar el método PCA para proyectar los descriptores de **S** desde 4 dimensiones a **una única dimensión** creando el conjunto **T**.

a. Señale la matriz de transformación ***W*** que se debe aplicar sobre **S** para crear el conjunto **T**.

b. Señale la "cantidad de información" que se mantiene (según PCA) al proyectar los descriptores de **S** en **T**. Justifique.

Se tiene una imagen de consulta a la que se le calculó su descriptor de contenido ***q***:

$$q = \begin{pmatrix} 9 \\ 12 \\ 5 \\ 17 \end{pmatrix}$$

El conjunto **T** se compone de los siguientes 9 descriptores de una dimensión, obtenidos luego de proyectar con PCA los descriptores de contenido de las imágenes de **R**:

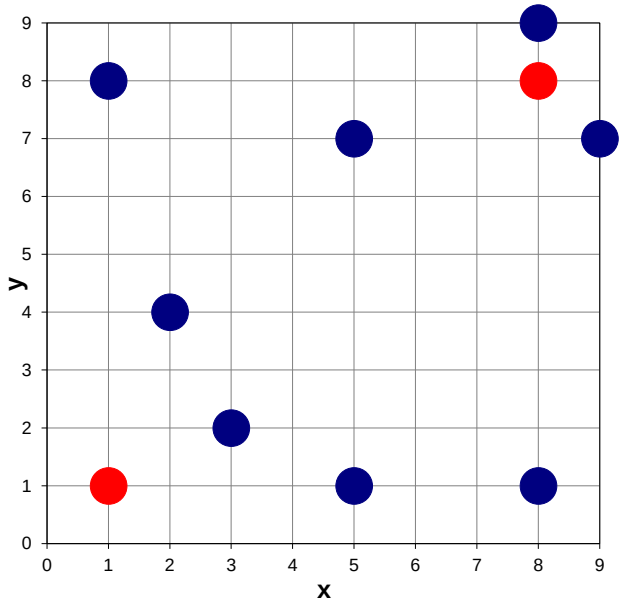
$$\begin{array}{lll} a = \begin{pmatrix} -13.6 \end{pmatrix} & d = \begin{pmatrix} 0.7 \end{pmatrix} & g = \begin{pmatrix} 7.4 \end{pmatrix} \\ b = \begin{pmatrix} -7.8 \end{pmatrix} & e = \begin{pmatrix} 2.5 \end{pmatrix} & h = \begin{pmatrix} 11.9 \end{pmatrix} \\ c = \begin{pmatrix} -2.1 \end{pmatrix} & f = \begin{pmatrix} 5.9 \end{pmatrix} & i = \begin{pmatrix} 13.8 \end{pmatrix} \end{array}$$

c. Señale el vector de ***T*** que será el vecino más cercano a ***q***, cuando ***q*** es proyectado a una dimensión con la transformación PCA descrita anteriormente.

Pregunta 5 (Índices Métricos, semana 14)

El conjunto ***R*** contiene ocho vectores de dos dimensiones ***R*={a, …, h}**. El conjunto ***Q*** contiene dos vectores que forman el conjunto de consulta ***Q*={q_1, q_2}**.

R	**x**	**y**
a	5	1
b	9	7
c	3	2
d	5	7
e	8	9
f	2	4
g	8	1
h	1	8



Q	**x**	**y**
q_1	1	1
q_2	8	8

Se desea localizar para cada objeto de ***Q*** su vecino más cercano en ***R*** de acuerdo a la distancia **Manhattan** o **L_1**. Se utilizará el enfoque métrico para acelerar las búsquedas. Suponga que se seleccionó el conjunto de pivotes: **P = { a }.**

a. Calcule la Tabla de Pivotes para **P**.

b. Resuelva la búsqueda exacta del vecino más cercano para **q_1** y para **q_2** usando cotas inferiores según **P**.

c. Calcule la complejidad interna y externa del índice al resolver ambas búsquedas. ¿Se realizaron más o menos cálculos que en un linear scan? Justifique.

Entrega:

- Puede desarrollarlo en papel y enviar una foto, o puede desarrollarlo en formato digital (planilla, documento u otro) y exportarlo a .pdf.
- El plazo máximo de entrega es el **jueves 30 de noviembre de 2023** hasta las 23:59 por U-Cursos.
- Será posible volver a enviarlo una vez más durante el semestre (en fecha por definir).

El mini-control es *individual* y debe ser de su autoría. En caso de detectar copia o plagio se asignará nota 1.0 a los involucrados.