



Analítica y Visualización de Datos

Dr. Miguel Jesús Torres Ruiz

Noviembre, 2022



Distancia de Hamming ⁽¹⁾

- **Definición**

- La distancia de Hamming se define como: $D_H(x, y) = \sum_{i=1}^p \rho(x^{(i)}, y^{(i)})$ y considerando la métrica discreta:

$$\rho(x, y) = \begin{cases} 0 & \text{si } x = y \\ 1 & \text{Otro caso} \end{cases}$$

- Entonces, la distancia de Hamming produce la cantidad de valores de características que no coinciden.
 - Para características binarias, la distancia de Hamming es igual a la distancia de Manhattan.
- Sin embargo, la distancia de Hamming no está asociada con una norma porque la condición $\|\alpha \cdot x\| = |\alpha| \cdot \|x\| \forall \alpha \in \mathbb{R}, x \in \mathbb{R}^p$ no se cumple.
- Las variantes de la distancia de Hamming usan funciones modificadas ρ para especificar similitudes entre características individuales.
 - Por ejemplo, si las características son páginas web (escala nominal), entonces ρ podría ser menor para pares de páginas con contenido similar y mayor para pares de páginas con contenido bastante diferente.



Distancia de Hamming ⁽²⁾

- **Definición**

- La distancia de Hamming se utiliza en procesamiento de señales y telecomunicaciones.
 - Contar el número de bits corruptos en la transmisión de un mensaje de una longitud determinada.
- Permite cuantificar la diferencia entre dos secuencias de símbolos.
- Es una distancia en el sentido matemático, considerando dos secuencias de símbolos de la misma longitud y asocia el número de posiciones donde difieren las dos secuencias.
- Para comparar las secuencias de longitudes variables o cadenas de caracteres que pueden sufrir no solo sustituciones, sino también inserciones o borrados, se utiliza la **distancia de Levenshtein**.

- **Ejemplo:**

- $\alpha = (0\ 0\ 0\ 1\ 1\ 1\ 1)$; $\beta = (1\ 1\ 0\ 1\ 0\ 1\ 1) \therefore d_H = 1 + 1 + 0 + 0 + 1 + 0 + 0 = 3$
- La distancia entre α y β es igual a 3 porque **3 bits difieren**.
- La distancia de Hamming entre **(1 0 1 1 1 0 1)** y **(1 0 0 1 0 0 1)** es **2**
- La distancia de Hamming entre **(2 14 3 8 96)** y **(2 23 3 7 96)** es **3**
- La distancia de Hamming entre "**r a m e r**" y "**c a s e s**" es **3**



Similitud y Disimilitud ⁽¹⁾

- **Similitud**

- Es una medida numérica que define qué tan parecidos son dos objetos de datos...
- Es más alta cuando los objetos son más parecidos.
- A menudo caen en el rango $[0,1]$.

- **Disimilitud**

- Es una medida numérica que define qué tan diferentes son dos objetos de datos...
 - Cuanto más bajo es el valor es cuando los objetos son más parecidos.
 - La disimilitud mínima suele ser 0.
 - El límite superior varía.
- En resumen, la **proximidad** se refiere a una similitud o disimilitud.



Medidas de Similitud ⁽¹⁾

- En búsqueda y/o recuperación de información:
 - Una medida de similitud puede representar la similitud entre dos **documentos**, dos **consultas** o un documento y una consulta.
 - Es posible clasificar (**ranking**) los documentos recuperados en el orden de supuesta importancia.
 - Una medida de similitud es una función que calcula el **grado de similitud** entre un par de objetos (documentos).
 - Hay un gran número de medidas de similitud propuestas en la literatura,
 - Porque la **mejor** medida de similitud **no existe** (¡todavía!).



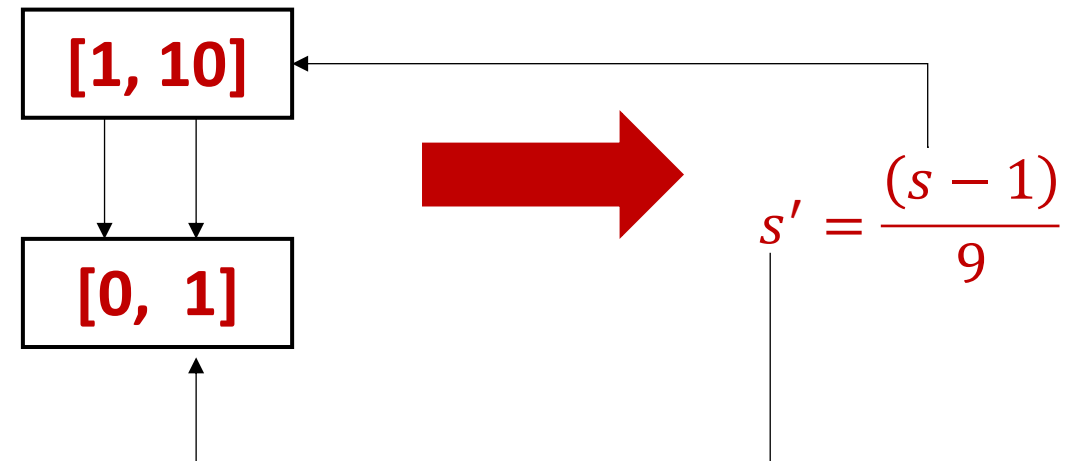
Medidas de Similitud ₍₂₎

- Conversión de valores de similitud:

- Por ejemplo:
 - 1 muestra similitud incompatible
 - 10 muestra similitud absoluta

- En general podemos utilizar:

$$s' = \frac{(s - \min_s)}{(\max_s - \min_s)}$$





Medidas de Similitud ₍₃₎

- Definición
 - Una función s se llama medida de **similitud** o proximidad si para toda $x, y \in \mathbb{R}^p$.
 - $s(x, y) = s(y, x)$
 - $s(x, y) \leq s(x, x)$
 - $s(x, y) \geq 0$
 - La función s se llama medida de **similitud normalizada** si:
 - $s(x, x) = 1$
- Cualquier medida de disimilitud d se puede usar para definir una medida de similitud correspondiente s y viceversa. Por ejemplo, usando una función positiva monótonicamente decreciente $f(0) = 1$ tal como:

$$s(x, y) = \frac{1}{1 + d(x, y)}$$



Modelo de Espacio Vectorial ⁽¹⁾

- Es un modelo algebraico utilizado para el **filtrado**, **recuperación**, **indexado** y cálculo de **relevancia** de **información**.
- Puede ser utilizado en datos de una manera formal mediante el uso de **vectores** (que pueden ser identificadores, o por ejemplo términos de búsqueda) en un **espacio lineal multidimensional**.
- Muchas de las tareas de recuperación de información como la **búsqueda**, **agrupamiento** o **categorización** de documentos tienen como primer objetivo procesar documentos en lenguaje natural.
- El problema que surge es que los algoritmos que pretenden resolver estas tareas necesitan representaciones internas **explícitas** de los **documentos**.



Modelo de Espacio Vectorial ₍₂₎

- En el área de recuperación de información normalmente se usa una expresión vectorial, donde las **dimensiones** del **vector** representan términos, frases o conceptos que aparecen en el documento.
- En este aspecto la representación más adoptada es la conocida como **bolsa de palabras**: que es una colección de documentos compuesta por **n** documentos indexados y **m** términos representados por una matriz documento-término de **$n \times m$** .
- Donde los **n** vectores renglón representan los **n** documentos; y el valor asignado a cada componente refleja la **importancia** o **frecuencia ponderada** que produce el término, frase o concepto **t_i** en la representación semántica del documento **j** .

$$d_j(\omega_{1j}, \omega_{2j}, \dots, \omega_{mj})$$



Modelo de Espacio Vectorial ⁽³⁾

- Donde m es la cardinalidad del diccionario (una lista de términos únicos que aparecen en un conjunto de documentos) y $0 \leq \omega_{ij} \leq 1$ representa la contribución del término t_i para la representación semántica del documento d_j .
- En esta representación vectorial de documentos el éxito o fracaso se basa en la ponderación o peso de los términos.
- Aunque existen muchos trabajos sobre técnicas de ponderación de términos, en realidad no hay un consenso sobre cuál método es el mejor.
- También hay que destacar que el espacio de renglones de la **matriz documento-término** determinan el contenido semántico de la colección de documentos. Sin embargo, una combinación lineal de dos vectores-documento no representa necesariamente un documento viable de la colección.
- Mediante el modelo espacio vectorial se pueden explotar las **relaciones geométricas** entre dos vectores documento (y términos) a fin de expresar las similitudes y diferencias entre términos.



Modelo de Espacio Vectorial ⁽⁴⁾

- Si bien el **rendimiento** de un sistema de recuperación de información depende en gran medida de las medidas de similitud entre documentos, la ponderación de términos desempeña un papel fundamental para que esa similitud entre documentos sea más confiable.
- Por ejemplo, mientras que una representación de documentos basada solo en las **frecuencias** o apariciones de términos no es capaz de representar adecuadamente el contenido semántico de los documentos, la representación de **términos ponderados** (aplicación de métodos de normalización a la matriz documento-término) hace frente a errores o incertidumbres asociadas a la representación simple de documentos.



Modelo de Espacio Vectorial ⁽⁵⁾

- Implementación del modelo
- Una colección de n documentos indexados por m términos puede ser representada por una matriz A de dimensión $n \times m$, donde cada elemento a_{ij} es usualmente definido por una frecuencia ponderada del término i en el documento j cuyo objetivo principal es mejorar el rendimiento en la recuperación de información.
- Entendiendo como rendimiento la **habilidad** de recuperar información **relevante** y descartar información **irrelevante**.

Matriz documento-término simple, donde cada columna representa un **término** en la colección, cada renglón un **documento** y cada celda o elemento de la matriz la **ocurrencia** en el documento



	Término 1	Término 2	Término 3
Documento 1	1	0	0
Documento 2	0	0	1
Documento 3	1	1	1
Documento 4	0	1	0

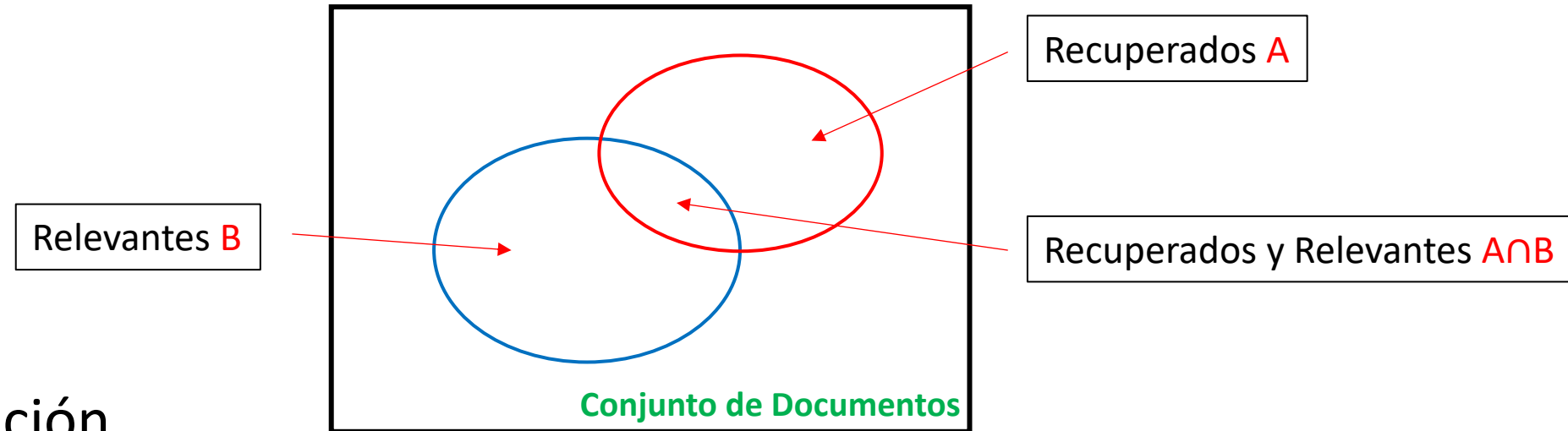


Modelo de Espacio Vectorial ⁽⁶⁾

- A partir de la matriz, se puede observar que el **Término 1** aparece en el **Documento 1 y 3**, pero no en los otros dos documentos. Se demuestra así que cada renglón de la matriz de 4×3 puede ser representado en un espacio de tres dimensiones.
- Entonces cada elemento a_{ij} de la matriz documento-término A queda definido como: $a_{ij} = l_{ij} * g_i * d_j^{-1}$
- Donde: l_{ij} es el peso local del término i en el documento j , el cual mide la importancia de dicho término en el documento; g_i es el peso global del término i en la colección de documentos y d_j es el factor de normalización para el j -ésimo documento.
- **Peso local:** mide la importancia del término i en el documento j y solo depende de las frecuencias en el documento y no de otros documentos.
- **Peso global:** Son aquellos que toman información de la colección de documentos para obtener el peso de un término en un documento.



Modelo de Espacio Vectorial (7)



- Medidas de evaluación

- $Precisión = \frac{\text{Documentos relevantes Devueltos}}{\text{Total de Documentos Devueltos}}$

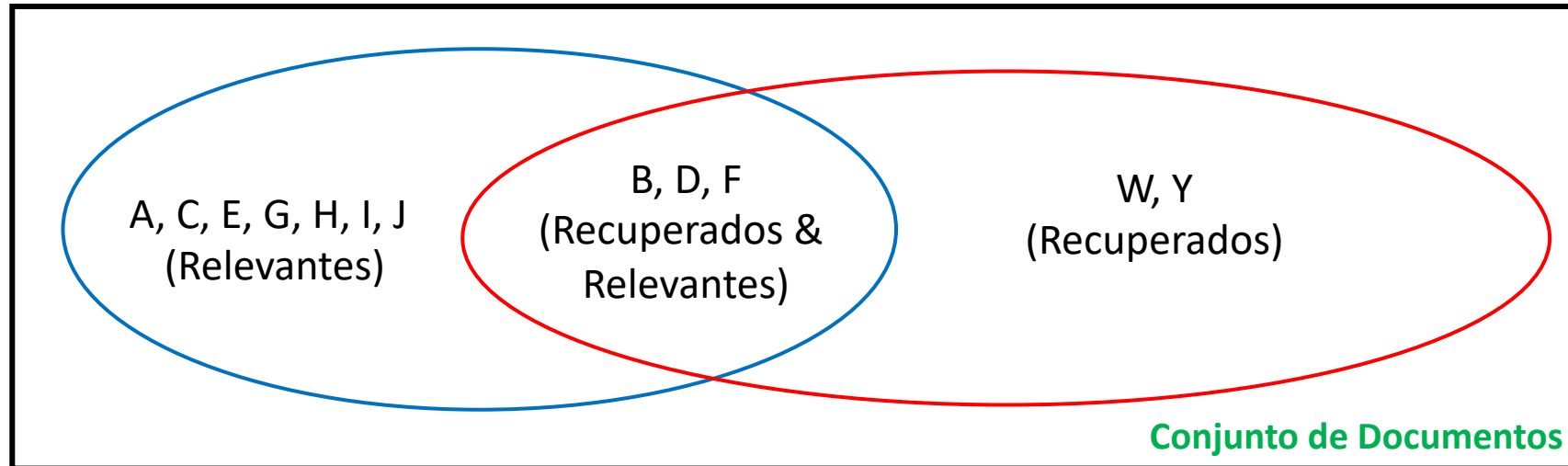
- $P(A, B) = \frac{|A \cap B|}{|A|}$

- $Recall = \frac{\text{Documentos relevantes Devueltos}}{\text{Total de Documentos Relevantes}}$

- $R(A, B) = \frac{|A \cap B|}{|B|}$



Modelo de Espacio Vectorial ₍₈₎



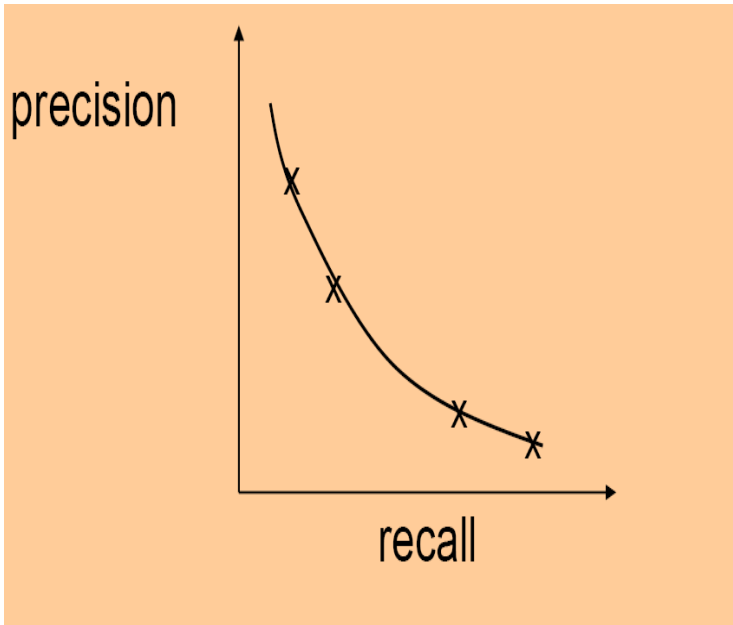
- Conjunto de documentos

- $|A| = \{recuperados\} = \{B, D, F, W, Y\} = 5$
- $|B| = \{relevantes\} = \{A, B, C, D, E, F, G, H, I, J\} = 10$
- $|A \cap B| = \{recuperados\} \cap \{relevantes\} = \{B, D, F\} = 3$
- $Precision = \frac{3}{5} = 60\%$
- $Recall = \frac{3}{10} = 30\%$



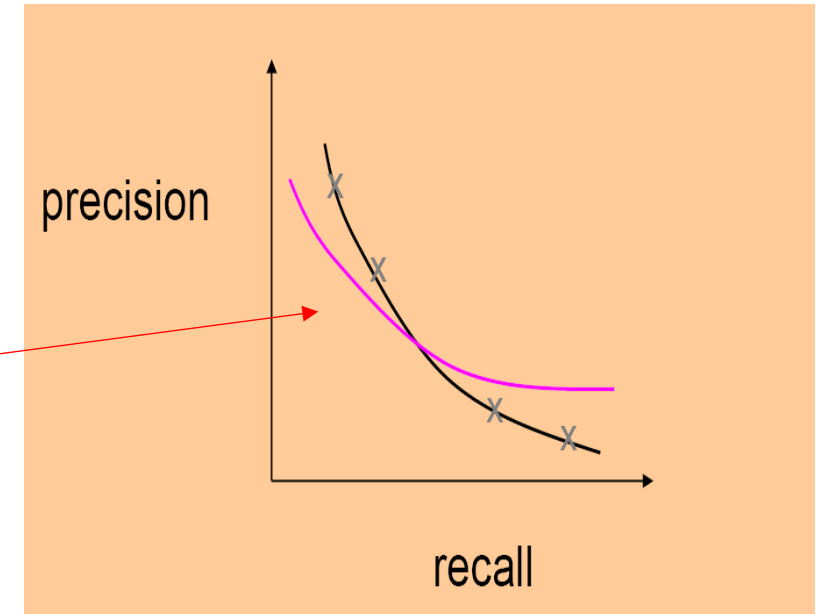
Modelo de Espacio Vectorial ₍₉₎

- Siempre existe una compensación entre las medidas *precision* y *recall*.
- Por tanto, se mide el *precision* en diferentes niveles de *recall*.



Una consulta

Difícil determinar cuál de estos
dos resultados hipotéticos es
mejor



Dos consultas



Definición de medidas de similitud ⁽¹⁾

- Consideremos primero las similitudes entre los vectores de características binarias.
 - Un par de vectores de características binarias pueden considerarse similares si muchos coinciden.
 - Esta **coincidencia** se puede representar mediante la operación producto, por lo que el **producto escalar** de los vectores de características es un candidato razonable para una medida de similitud.
 - También para características de valores reales no negativos $x, y \in (\mathbb{R}^+)^p$. Entonces, las medidas de similitud se pueden definir en función de productos escalares que se pueden normalizar de diferentes maneras.
 - Medida de Similitud del Coseno
 - Medida de Similitud Dice
 - Medida de Similitud Jaccard (Tanimoto)
 - Medida de Similitud de Sobreposición (Overlap)



Definición de medidas de similitud (2)

- Estas expresiones no están definidas para vectores de características **cero** porque los **denominadores** son cero. Entonces, la similitud debe definirse explícitamente para este caso como cero.
- Por ejemplo, la similitud del coseno es **invariable** frente a la escala (positiva) de los vectores de características y , por lo tanto, considera la distribución relativa de las características, cumpliendo con:
 - $s(c \cdot x, y) = s(x, y)$
 - $s(x, c \cdot y) = s(x, y)$
 - Para toda $x, y \in \mathbb{R}^p$ y $c > 0$
- Por ejemplo, considerar dos recetas de pasteles:
 - **Receta 1.** 3 huevos, $1 \frac{1}{2}$ tazas de azúcar, $1 \frac{1}{2}$ tazas de harina y $\frac{1}{2}$ taza de mantequilla.
 - **Receta 2.** 6 huevos, 3 tazas de azúcar, 3 tazas de harina y 1 taza de mantequilla.
- Obviamente, ambas recetas dan el mismo resultado, pero la segunda rinde el doble de pastel que la primera.
- Por tanto, siguiendo una expectativa intuitiva, la **similitud del coseno** entre las dos recetas es igual a **uno**.
- Entonces, podemos concluir que estas medidas cuantifican la similitud entre vectores de características (**filas de la matriz de datos**). Pero para cuantificar la similitud entre características (**columnas de la matriz de datos**), se utiliza la **correlación**.
- Si la matriz de datos es transpuesta (las filas y las columnas se intercambian), la correlación también se puede usar como una forma alternativa de cuantificar la similitud entre los vectores de características.



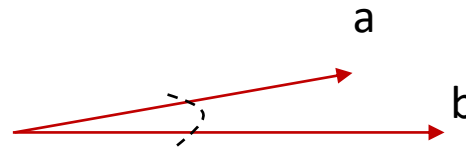
Similitud del Coseno ⁽¹⁾

- En minería de datos, la similitud del coseno se refiere a la **distancia con dimensiones** que representan las **características** de los objetos de datos en un conjunto de datos.
- También conocida como **distancia coseno**, y se define como la medida de la **magnitud** de la diferencia entre dos individuos, usando el valor coseno del ángulo entre dos vectores en un espacio vectorial.
- En la similitud del coseno, los objetos de datos en un conjunto se procesan como **vectores**.
- Cuanto más **cercano** es el valor del coseno a **1**, más cerca está el ángulo a **0 grados**; es decir, los **dos vectores** son **más similares**, lo que se denomina como “similitud del coseno”.

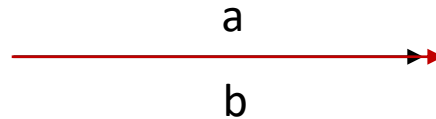


Similitud del Coseno (2)

- Por ejemplo, en la figura se puede apreciar que el **ángulo** entre los dos vectores ***a*** y ***b*** es muy **pequeño**. Por tanto, se puede decir que el vector ***a*** y el vector ***b*** tienen una gran **similitud**.



- En casos extremos, cuando los **vectores *a* y *b*** coinciden completamente. Entonces ***a* y *b*** son iguales, lo que significa que los datos representados por los vectores ***a* y *b*** son completamente **similares o iguales**.



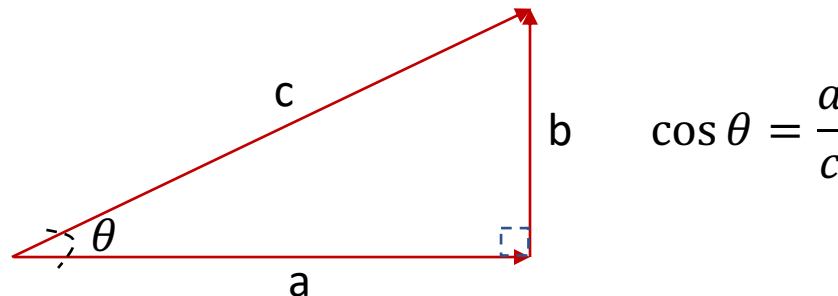


Similitud del Coseno (3)

- Pero que pasa si el ángulo entre los vectores a y b es **grande**, o se ubica en la dirección opuesta, se puede decir que el vector a y el vector b tienen una **baja similitud**, o que el conjunto de datos representado por los vectores a y b no es básicamente similar.



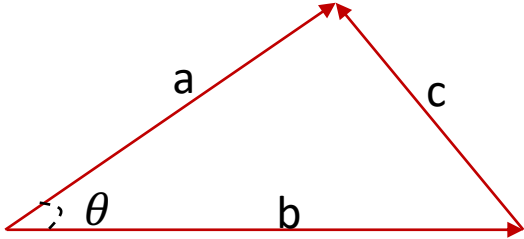
- Por tanto, la teoría de la similitud del coseno del espacio vectorial es un método para calcular la **similitud de los individuos**, a partir del siguiente análisis.
- Cuando se piensa en la fórmula del coseno, el método de cálculo más básico es θ .





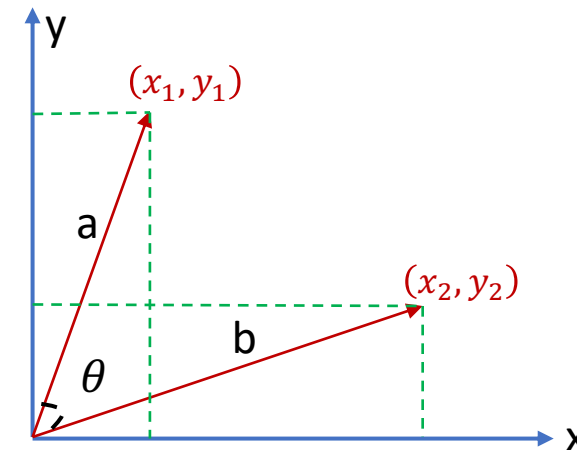
Similitud del Coseno ⁽⁴⁾

- Pero la fórmula anterior, es solo aplicable para **triángulos rectángulos**, y en los triángulos no rectángulos, la fórmula para calcular el coseno del ángulo **a** y **b** sería:


$$\cos \theta = \frac{a^2 + b^2 - c^2}{2ab}$$

- Por ejemplo, en el triángulo representado por el vector, suponiendo que el vector **a** es (x_1, y_1) y el vector **b** es (x_2, y_2) , el teorema del coseno se puede reescribir de la siguiente forma a partir de la figura.

$$\begin{aligned} \cos \theta &= \frac{a \cdot b}{||a|| \cdot ||b||} = \frac{(x_1, y_1) \cdot (x_2, y_2)}{\sqrt{x_1^2 + y_1^2} \cdot \sqrt{x_2^2 + y_2^2}} \\ &= \frac{x_1 x_2 + y_1 y_2}{\sqrt{x_1^2 + y_1^2} \cdot \sqrt{x_2^2 + y_2^2}} \end{aligned}$$





Similitud del Coseno ⁽⁵⁾

- Para el caso, si los vectores **a** y **b** no son bidimensionales, el método de cálculo del coseno anterior sigue siendo correcto. Suponiendo que **a** y **b** son dos vectores de n -dimensiones, **a** es **B** y **b** es **A**, entonces el coseno del ángulo entre **a** y **b** es igual a:

$$\cos \theta = \frac{\sum_{i=1}^p x_i y_i}{\sqrt{\sum_{i=1}^p (x_i)^2 \sum_{i=1}^p (y_i)^2}} = \frac{a \cdot b}{||a|| \cdot ||b||}$$

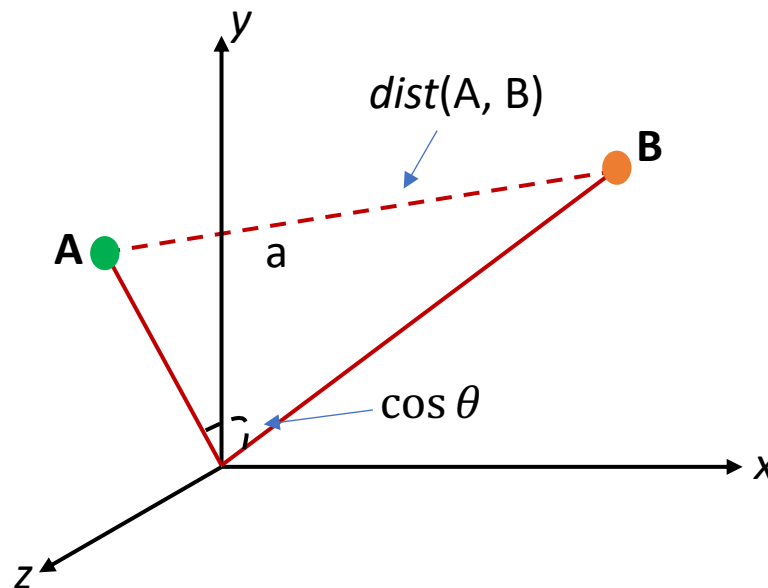
$$\text{sim}(x, y) = \frac{\sum_{i=1}^p x_i y_i}{\sqrt{\sum_{i=1}^p (x_i)^2 \sum_{i=1}^p (y_i)^2}}$$

- Por tanto, mientras más **cerca** esté el valor del coseno a **1**, más **cerca** estará el ángulo a **0 grados**, es decir, los **dos vectores** son más **similares** y el ángulo es igual a 0. Esto indica que los dos vectores son iguales, y se denomina **similitud del coseno**.



Similitud del Coseno ⁽⁶⁾

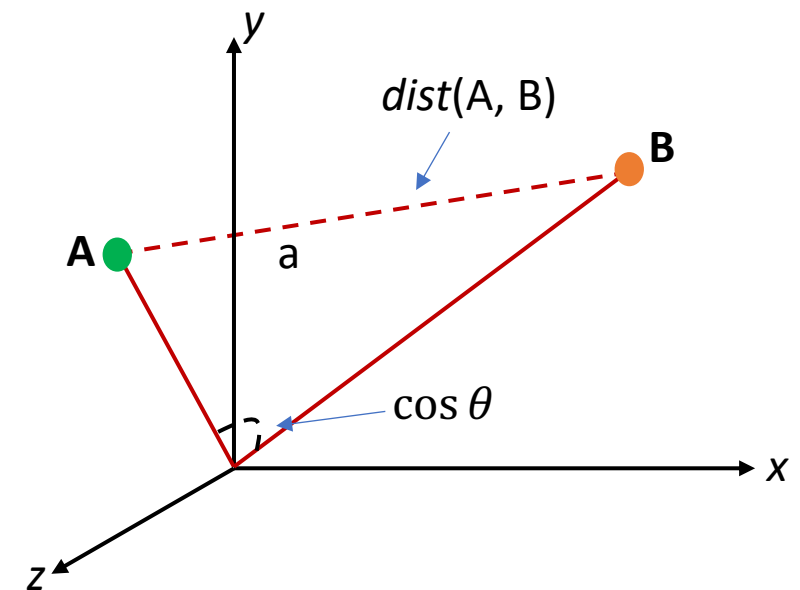
- Por otra parte, la distancia del coseno utiliza el valor del coseno del ángulo entre los dos vectores como una medida de la **diferencia** entre los dos **individuos**. En comparación con la distancia euclidiana, la distancia del coseno presta más atención a la **diferencia** en la **dirección** de los **vectores**.
- Utilizando un sistema con **coordenadas tridimensionales**, se puede observar la diferencia entre la distancia euclidiana y la del coseno.





Similitud del Coseno ⁽⁷⁾

- La distancia euclidiana mide la **distancia absoluta** de cada punto en el espacio y está directamente relacionada con las **coordenadas** de posición de cada punto.
- La distancia del coseno mide el **ángulo del vector espacial**, que se refleja más en la diferencia de la dirección, no de la ubicación, pero si mantiene la posición del punto A sin cambios y el punto B lejos del origen del eje de coordenadas en la dirección original.
- Entonces la distancia del coseno ($\cos\theta$) permanece igual (**porque el ángulo no cambia**) y la distancia entre los puntos A y B si está cambiando.
- Esta es la diferencia entre la distancia euclidiana y la distancia del coseno.





Similitud del Coseno ⁽⁸⁾

- La distancia euclidiana y del coseno tienen diferentes **métodos de cálculo** y diferentes **características de medición**, por lo que son adecuadas para diferentes modelos de análisis de datos, por ejemplo:
 - La distancia **euclidiana** puede reflejar la **diferencia absoluta** de las características numéricas individuales, por lo que se usa más para aquellos análisis que requieren reflejar la diferencia del valor numérico de la dimensión, como el uso de indicadores de comportamiento del usuario para analizar la similitud o diferencia en el valor del usuario.
 - La distancia del **coseno** se usa más para **distinguir** la **diferencia** de la **dirección**, pero no es sensible al valor absoluto. Se utiliza para distinguir la **similitud** y la diferencia de interés por la calificación del contenido por parte del usuario. Al mismo tiempo, se corrigen los posibles estándares de medición entre usuarios.



Similitud del Coseno ⁽⁹⁾

- Ejemplo de cálculo de la similitud de coseno para encontrar similitud en texto:
 - **Texto 1.** Julie loves me more than Linda loves me
 - **Texto 2.** Jane likes me more than Julie loves me
 - ¿Cómo calcular la similitud entre los dos textos?
 - Queremos saber en qué medida se parecen estos textos, únicamente en términos de **recuento de palabras** (e ignorando el orden de las mismas). Comenzamos haciendo una lista de ambos textos.
- me Julie loves Linda than more likes Jane
- Se procede en obtener la **frecuencia** que aparece cada una de estas palabras en cada texto.

Palabra	Freq-T ₁	Freq-T ₂
me	2	2
Jane	0	1
Julie	1	1
Linda	1	0
likes	0	1
loves	2	1
more	1	1
than	1	1



Similitud del Coseno ₍₁₀₎

- Sin embargo, no nos interesan las palabras en sí; solo nos interesan esos **dos vectores verticales** de frecuencia.
 - Por ejemplo, hay dos casos de “**me**” en cada texto. Vamos a decidir lo **cerca** que están estos dos textos **entre sí** calculando una función de esos dos vectores, particularmente el coseno del ángulo entre ellos.
 - a: [2, 0, 1, 1, 0, 2, 1, 1]
 - b: [2, 1, 1, 0, 1, 1, 1, 1]
 - Aplicando la fórmula la **similitud del coseno**

$$sim(x, y) = \frac{\sum_{i=1}^p x_i y_i}{\sqrt{\sum_{i=1}^p (x_i)^2 \sum_{i=1}^p (y_i)^2}} = sim(x, y) = 0.822$$

- Estos vectores son de **8 dimensiones**. Una virtud de usar la similitud del coseno es que convierte una cuestión que va más allá de la capacidad humana para visualizarla, en una que puede serlo.
- En este caso se podría pensar esto, como un **ángulo de unos 35 grados**, que está a cierta “**distancia**” de 0 o de la concordancia perfecta.

Palabra	Freq-T ₁	Freq-T ₂
me	2	2
Jane	0	1
Julie	1	1
Linda	1	0
likes	0	1
loves	2	1
more	1	1
than	1	1



Proyecto 5 ₍₁₎

Proyecto 5. Desarrollar un programa en R que permita calcular la similitud del coseno a partir del conjunto de datos: Abstract COVID Papers.csv

Instrucciones:

- 1. Analizar el conjunto de datos que contiene 3 atributos: “title”, “abstract” y “url”.**
- 2. Revisar el corpus de datos y los metadatos directamente de la página: <https://www.kaggle.com/datasets/anandhuh/covid-abstracts>**
- 3. Seleccionar 20 abstracts y aplicarles la similitud del coseno. Cada uno debe seleccionar diferentes títulos para que no sean iguales. El dataset está compuesto por 10,000 instancias.**
- 4. Posteriormente, *rankear* los abstracts por grado de similitud y visualizarlos por el título.**