Taller de similaridad de documentos usando la medida del coseno

Raúl Ernesto Gutierrez de Piñerez Reyes

1. Enunciado del problema

Dado el conjunto de términos-documentos:

	Doc1	Doc2	Doc3
car	27	4	24
auto	3	33	0
insurance	0	33	29
best	14	0	17

Figura 1:

Entonces podemos calcular la similaridad de un query q con un documento d de la siguiente manera:

$$Sim(q,d) = \frac{\Sigma_{i=1}^{N} \overrightarrow{q}_i \overrightarrow{d_i}}{\sqrt{\Sigma_{i=1}^{N} \overrightarrow{q}_i^2 \overrightarrow{d}_i^2}}$$

Donde N es el número de términos. Sea el query q un vector (13,0,0,22) para los términos *car, auto, insure y best* respectivamente y los vectores para los documentos Doc1 (27,3,0,14), Doc2 (4,33,33,0) y Doc3 (24,0,29,17) de la matriz mostrada en la Figura 1. Use el criterio de similaridad del coseno e implemente la función $Sim(q,d)$ sobre los términos *car* (eje X) y *best* (eje Y) para encontrar el documento más relevante (valor mayor del coseno) y el menos relevante (menor valor del coseno) según el query. Implementar además una rutina que muestre los vectores de cada documento y el query en el plano cartesiano sobre los términos *car* y *best*.

2. Generalidades

- Lenguaje: Python,
- Valor del taller: 20 %
- Tiempo de entrega: 10 días