Taller 11

Ejercicio Naybe Bayes

Usando la técnica Naybe Bayes construya un clasificador que determine si un texto es sobre deportes o no. Los datos de entrenamiento tienen 5 oraciones:

Texto	Etiqueta
Un juego inolvidable	Deportes
Se acabó el discurso	No deportes
Partido muy limpio	Deportes
Un juego limpio pero olvidable	Deportes
Fue un discurso cerrado	No deportes

Dado que Naive Bayes es un clasificador probabilístico, calcule la probabilidad de que la frase "Un juego muy cerrado" pertenezca a la clase Deportes y la probabilidad de que no sea Deportes.

Ejercicio Knn

Usando la técnica Knn (k nearest neighbor) clasifique los textos de ejemplo en dos clases de temas

- a (amor) y
- c (crimen).

Para ello suponga que cada documento (D_i) luego de ejecutar las tareas de pre-procesamiento contiene las siguientes frecuencias de términos

	Clase a			Clase c	
Términos	Documento			Documento	
	D1	D2	D3	D4	D5
novio	10	8	7	0	1
beso	5	6	4	1	0
inspector	2	0	0	12	8
asesino	0	1	0	20	56

La tarea es clasificar dos nuevos documentos los cuales contiene las frecuencias listadas continuación

Términos	Documento	
	D6	D7
Amor	5	1
Beso	6	0
Inspector	2	12
Asesino	0	4

Antes de comparar los vectores y calcular las distancias normalice los vectores. Use como medida la distancia euclidiana

Determine las clases asignadas a los nuevos documentos para valores de k = 1 y k = 3, usando la técnica de votos por mayoría y suma ponderada.

Recuerde que la técnica de votos por mayoría asigna la siguiente puntuación a cada clase c, dado un documento d que debe clasificarse:

$$score(c, d) = \sum_{d_t \in S_k(d)} I_c(d_t)$$

Donde:

- Sk(d) es el conjunto de los k documentos adyacentes más cercanos de un nuevo documento d.
- Se define la función Ic: Sk → {0;1} para una clase c y un documento d como Ic(dt) = 1 si la clase de dt es c, de lo contrario Ic(dt) = 0.

El clasificador asigna a la clase c del conjunto de clases C con la puntuación más alta:

$$\hat{c} = \arg\max_{c \in C} score(c, d)$$

La técnica de suma ponderada asigna la siguiente puntación a cada clase c, dado un documento d

$$score(c, d) = \sum_{d_t \in S_k(d)} I_c(d_t) \cos(\vec{v}(d_t), \vec{v}(d))$$

Donde:

cos (v(dt), v(d)) es la medida del coseno entre los vectores que representan los documentos

Tarea adicional

Reemplace la frecuencia de palabras con los pesos tfn * idf correspondientes en lugar de las frecuencias.