



# Procesamiento de Lenguaje Natural

## Vectores Semánticos

Mauricio Toledo-Acosta  
[mauricio.toledo@unison.mx](mailto:mauricio.toledo@unison.mx)

Departamento de Matemáticas  
Universidad de Sonora



## Section 1

# Introducción







## Section 2

# Modelo BOW



# La matriz term-document

Procesamiento  
de Lenguaje  
Natural

Introducción

Modelo BOW

Modelo  
TF-IDF

La métrica  
angular

- La Revolución Francesa fue un período de grandes **cambios** políticos y sociales en **Europa**.
- El Imperio Romano dominó gran parte de **Europa** durante siglos, expandiéndose por toda **Europa**.
- La paella es un plato tradicional de España que lleva **arroz**, mariscos y verduras.
- El sushi es una comida japonesa hecha con **arroz** y **pescado** crudo, acompañado de algas.

Texto	Europa	cambios	arroz	pescado
1	1	1	0	0
2	2	0	0	0
3	0	0	1	0
4	0	0	1	1

Visualización



Texto	Europa	cambios	arroz	pescado
1	1	1	0	0
2	2	0	0	0
3	0	0	1	0
4	0	0	1	1

El modelo BOW asigna a cada documento el vector correspondiente a la fila. El vector de cada palabra es su columna.

¿Cuántas filas y cuántas columnas hay en una matriz BOW?

## Visualización



### Ventajas:

- Simplicidad e interpretabilidad
- Fácil implementación computacional
- Eficiente para conjuntos de datos grandes
- Base fundamental para modelos más avanzados
- Compatible con algoritmos de machine learning

### Desventajas:

- Pérdida del orden de las palabras, contexto y semántica
- No captura relaciones entre términos (sinonimia, polisemia, etc.)
- Matriz grande y *sparse*
- Sensible a stop words
- No considera la importancia relativa de términos



- El gato come ratones y juega con el perro. El perro duerme al lado y come.
- El gato come pescado.
- El perro ladra fuerte y come.
- El código tiene un error.
- El programa ejecuta código.

- gato
- come
- ratones
- juega
- perro
- duerme
- lado
- pescado
- ladra
- fuerte
- código
- error
- programa
- ejecuta





## Problemas de la representación BOW (por conteos):

- Las palabras muy comunes (“el”, “de”, “y”) dominan los vectores. No necesariamente las stopwords.
- Estas palabras aportan poco sobre el contenido real del documento.

Idea central del modelo TF-IDF:

- No todas las palabras son igual de importantes.
- Una palabra es relevante si aparece *mucho* en un documento pero aparece en *pocos* documentos del corpus.



# La matriz TF-IDF

Procesamiento  
de Lenguaje  
Natural

Introducción

Modelo BOW

Modelo  
TF-IDF

La métrica  
angular

- La Revolución Francesa fue un período de grandes **cambios** políticos y sociales en **Europa**.
- El Imperio Romano dominó gran parte de **Europa** durante siglos, expandiéndose por toda **Europa**.
- La paella es un plato tradicional de España que lleva **arroz**, mariscos y verduras.
- El sushi es una comida japonesa hecha con **arroz** y **pescado** crudo, acompañado de algas.

Texto	Europa	cambios	arroz	pescado
1	0.043	0.301	0	0
2	0.043	0	0	0
3	0	0	0.090	0
4	0	0	0.060	0.301

Los valores TF-IDF ponderan la importancia de cada término según su frecuencia en el documento y su rareza en el corpus.



# Cálculo del TF-IDF

Procesamiento  
de Lenguaje  
Natural

Introducción

Modelo BOW

Modelo  
TF-IDF

La métrica  
angular

- $TF\text{-}IDF = TF * IDF$
- TF (Term Frequency): Frecuencia del término en el documento
- IDF (Inverse Document Frequency): Rareza del término en el corpus

$$TF(t, d) = \frac{\text{frecuencia del término } t \text{ en documento } d}{\text{total de términos en documento } d}$$

$$IDF(t) = \log \left( \frac{\text{total de documentos}}{\text{documentos que contienen término } t} \right)$$

$$TF\text{-}IDF(t, d) = TF(t, d) * IDF(t)$$



### Ejemplo: *Europa* en Texto 2

### Ejemplo: *arroz* en Texto 3



### Desventajas:

- No captura relaciones semánticas entre términos
- Matriz grande y *sparse*
- Sensible a la longitud del documento
- No resuelve problemas de polisemia o sinonimia
- Depende de la calidad del corpus de entrenamiento





# ¿Cómo medir la distancia entre documentos?

- Doc 0: “buen hotel, economico, muy recomendable hotel, mi hotel favorito”
- Doc 1: “buen hotel, economico”
- Doc 2: “buen automovil economico”

Representamos cada documento como vector BOW.

	automovil	buen	economico	favorito	hotel	recomendable
0	0	1	1	1	3	1
1	0	1	1	0	1	0
2	1	1	1	0	0	0

¿A qué documento se parece más el Doc 1? Calculemos distancias



# ¿Cómo medir la distancia entre documentos?

- Con distancia Euclidiana: Doc 1 es más cercano a Doc 2.

La distancia euclidiana es sensible a la longitud de los documentos: Doc 0 es más largo.

- Con la distancia angular (coseno):

$$d_{\theta}(u, v) = \arccos \left( \frac{u \cdot v}{|u| \cdot |v|} \right)$$
$$\text{sim}(u, v) = \frac{u \cdot v}{|u| \cdot |v|}$$

Doc 1 es más cercano a Doc 0.



# Conclusión

## Procesamiento de Lenguaje Natural

### Introducción

### Modelo BOW

### Modelo TF-IDF

### La métrica angular

La similitud coseno, o métrica angular, suele ser la métrica ideal para medir distancias entre representaciones de documentos (incluso en modelos basados en redes neuronales).

Dos documentos son similares si sus vectores apuntan en una dirección similar, sin importar su longitud. La distancia coseno se centra en el contenido de palabras. De esta forma, ignora la longitud del documento y captura mejor la similitud temática.