



Procesamiento de Lenguaje Natural

Vectores Semánticos

Mauricio Toledo-Acosta
mauricio.toledo@unison.mx

Departamento de Matemáticas
Universidad de Sonora



Section 1

Introducción





Section 2

Modelo BOW



- ¿Cómo representamos texto de forma que un algoritmo pueda procesarlo?
- Los modelos matemáticos trabajan con números, no con palabras.

La idea central del modelo BoW:

- Un documento se representa como un vector numérico.
- Cada posición del vector corresponde a una palabra del vocabulario.
- El valor indica cuántas veces aparece esa palabra en el documento.

“Es una bolsa donde echamos todas las palabras del documento, ignorando su orden”



La matriz term-document

Procesamiento
de Lenguaje
Natural

Introducción

Modelo BOW

Modelo
TF-IDF

La métrica
angular

Ejercicio en
clase

- 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



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?

◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ ↺ 🔍 ↻ 8/21



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



Section 3

Modelo TF-IDF



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

- 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.





Cálculo del TF-IDF

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

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

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

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



Cálculo del TF-IDF

Ejemplo: *Europa* en Texto 2

- $TF = 2/14 = 0.143$ (aparece 2 veces de 14 palabras totales)
- $IDF = \log(4/2) = \log(2) = 0.301$
- $TF-IDF = 0.143 * 0.301 = 0.043$

Ejemplo: *arroz* en Texto 3

- $TF = 1/9 = 0.111$ (aparece 1 vez de 9 palabras totales)
- $IDF = \log(4/1) = \log(4) = 0.602$
- $TF-IDF = 0.111 * 0.602 = 0.067$



Ventajas y Desventajas

Procesamiento de Lenguaje Natural

Introducción

Modelo BOW

Modelo
TF-IDF

La métrica
angular

Ejercicio en
clase

Ventajas:

- Mejora la representación al ponderar términos
- Reduce el impacto de términos muy frecuentes
- Identifica términos relevantes por documento
- Mejor rendimiento en recuperación de información
- Fácil de implementar y computacionalmente eficiente

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?

- 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

Ejercicio en clase

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.



Section 5

Ejercicio en clase



Ejercicio BOW

Documentos:

- 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.

Palabras a considerar:

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