Procesamiento de Lenguaje Natural

Vectores Semánticos

Mauricio Toledo-Acosta
mauricio.toledo@unison.mx

Departamento de Matemáticas
Universidad de Sonora

Section 1

Introducción

Vectores Semánticos

Representación numérica del lenguaje natural

¿Por qué representar texto numéricamente?

- Los algoritmos de ML requieren features numéricas
- Necesidad de cuantificar la semántica y similitud
- Búsqueda y recuperación eficiente de información
- Análisis automatizado de grandes volúmenes de texto

Evolución de las representaciones

- **Modelos tradicionales**: BOW, TF-IDF (representaciones locales)
- **Modelos modernos**: Word2Vec, GloVe, BERT (representaciones densas)
- **LLMs**: GPT, LlaMa, Qwen, Claude, DeepSeek.

- De representaciones dispersas a embeddings densos
- De contar palabras a capturar significado contextual

Section 2

Modelo BOW

La matriz term-document

- 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

Modelo BOW

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 y Desventajas

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

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 consideradas:

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

Section 3

Modelo TF-IDF

La matriz TF-IDF

- 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	0.301	0.602	0	0
2	0.602	0	0	0
3	0	0	0.301	0
4	0	0	0.301	0.602

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

- TF-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 de *Europa* en Texto 2:

- TF $= 2/7 = 0.286$ (aparece 2 veces de 7 palabras totales)
- IDF $= \log(4/2) = \log(2) = 0.301$
- TF-IDF $= 0.286 * 0.301 = 0.086$

Ventajas y Desventajas

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