Procesamiento de Lenguaje Natural

Vectores Semánticos

Mauricio Toledo-Acosta
mauricio.toledo@unison.mx

Departamento de Matemáticas
Universidad de Sonora

Section 1

Introducción

Section 2

Ejemplos ilustrativos

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.

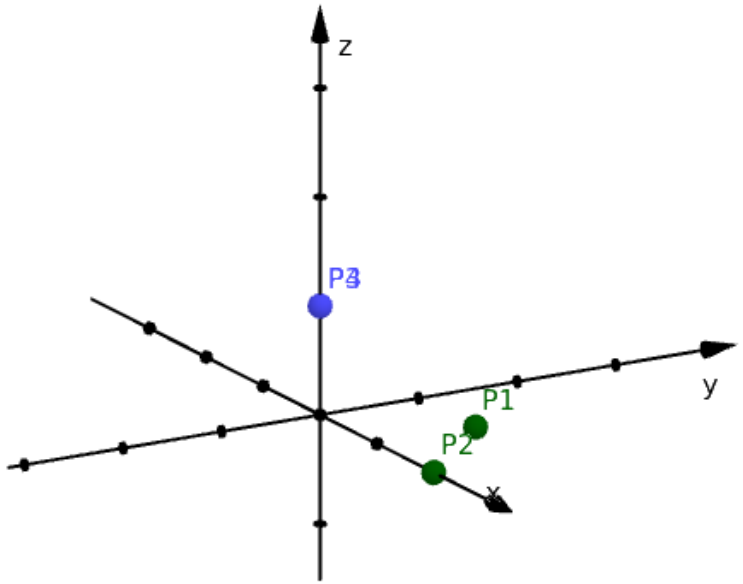
Texto	Europa	cambios	arroz	pescado
1	1	1	0	0
2	2	0	0	0
3	0	0	1	0
4	0	0	1	1

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.



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

Subsection 1

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 = $0.286 * 0.301 = 0.086$

Section 3

LSA

Latent Semantic Analysis

LSA (Latent Semantic Analysis)

Técnica de procesamiento de lenguaje natural usada en Topic Modelling para descubrir temas en textos, es decir, identificar temas ocultos en un conjunto de documentos.

Fundamentos teóricos

- **Matriz Término-Documento**: Representación numérica de textos, tipicamente BOW o TF-IDF.
- **SVD**: Reducción de dimensionalidad para capturar relaciones semánticas.
- **Espacio semántico latente**: Representación compacta de palabras y documentos.

SVD (Descomposición en Valores Singulares)

- **Propósito**: Reducir la dimensionalidad conservando la estructura semántica.
- **Ecuación**:

$$A_{m\times n} = U_{m\times m}\Sigma_{m\times n}V^T_{n\times n}$$

- **Proceso**: Se calculan las matrices simétricas AA^T y A^TA.
- **Componentes**: U son vectores propios de AA^T, V son vectores propios de A^TA, Σ contiene $\sqrt{\lambda_i}$.
- **Reducción**: Se conservan solo los k valores singulares más grandes: $A \approx U_k\Sigma_k V_k^T$.

Espacio semántico latente

- **Concepto**: Representación de palabras y documentos en un espacio de menor dimensión.
- **Ventaja**: Captura relaciones semánticas entre términos y documentos.
- **Ejemplo**: Palabras como "coche" y "automóvil" estarán cerca.

Proceso de LSA

- **Preprocesamiento**: Tokenización, eliminación de stopwords, etc.
- **Matriz Término-Documento**: Creación y ponderación (TF-IDF).
- **SVD**: Aplicación y reducción de dimensionalidad.
- **Interpretación**: Identificación de temas latentes.

Ventajas de LSA

- Captura relaciones semánticas entre palabras.
- Reduce el ruido en grandes conjuntos de datos.
- Simple y fácil de implementar.

Limitaciones de LSA

- Dificultad para interpretar temas explícitamente.
- Depende del preprocesamiento y parámetros.

Aplicaciones de LSA

- Recuperación de información.
- Clasificación de textos.
- Análisis de sentimientos.
- Recomendación de contenido.

Comparación con otras técnicas

- **LSA/NMF**: Basado en álgebra lineal (SVD).
- **LDA**: Basado en probabilidad (distribuciones de temas).
- **Clustering**: Basados en geometría y distancias.