



Procesamiento de Lenguaje Natural

Vectores Semánticos

Mauricio Toledo-Acosta
mauricio.toledo@unison.mx

Departamento de Matemáticas
Universidad de Sonora



Section 1

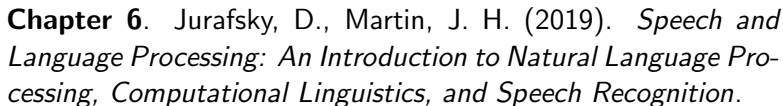
Introducción



the.language.nerds

Me speaking English:

I have literally no idea what this word translates to in my native language but I've seen it being used in similar context so I'm just gonna use it here and pray that it does mean what I think it means.





Palabras que aparecen en contextos similares tienden a tener significados similares. Este vínculo entre la similitud en la distribución de las palabras y la similitud en su significado se denomina **hipótesis distribucional**.

Esta hipótesis fue formulada en los años 50 por lingüistas como Joos (1950), Harris (1954) y Firth (1957), que observaron que las palabras sinónimas (como oculista y oftalmólogo) tendían a aparecer en el mismo entorno (por ejemplo, cerca de palabras como ojo o examinado).

A word is characterized by the company it keeps

Más información



Ejemplo

Procesamiento
de Lenguaje
Natural

Introducción

¿Qué es el **Ongchoi**?



- **Ongchoi** es delicioso **salteado con ajo**.
- **Ongchoi** es excelente **sobre arroz**.
- ...**ongchoi** **hojas** con salsas saladas...

- ...la espinaca salteada con ajo sobre arroz...
- ...los tallos de acelga y las hojas son deliciosos...
- ...la col rizada y otras verduras de hoja saladas...

◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ ↺ 🔍 ↻ 6/26



La **semántica vectorial** son modelos que buscan aprender representaciones del significado de las palabras directamente a partir de su distribución en los textos.

La idea de la semántica vectorial es representar una palabra como un punto en un espacio semántico multidimensional. Los vectores que representan palabras suelen denominarse **embeddings**, porque la palabra está incrustada en un espacio vectorial concreto.

La cercanía de embeddings da cuenta de diversos fenómenos, además de la similitud de palabras.

Ejemplo



What's the meaning of life?
LIFE

Un buen modelo semántico debería decirnos que:

- Algunas palabras tienen significados similares (gato es similar a perro).
- Algunas palabras son sinónimas o antónimas (frío – caliente).
- Algunas palabras tienen connotaciones positivas (feliz) mientras que otras tienen connotaciones negativas (triste).
- Algunas palabras como comprar, vender y pagar ofrecen perspectivas diferentes sobre el mismo acontecimiento de compra subyacente (Si te compro algo, me lo has vendido, y te he pagado).



Similitud de Palabras

No todas las palabras tienen muchos sinónimos, sin embargo, la mayoría de ellas tienen muchas palabras similares (gato – perro). La noción de similitud entre palabras es muy útil en diversas tareas semánticas, por ejemplo, decidir si dos oraciones significan cosas parecidas.

- Podemos obtener las similitudes entre palabras de listas pre-definidas (por ejemplo, [SimLex-999](#)).

vanish	disappear	9.8
behave	obey	7.3
belief	impression	5.95
muscle	bone	3.65
modest	flexible	0.98
hole	agreement	0.3

- Podemos aprender las similitudes a partir de co-ocurrencias.



- Modelos basados en conteos: El significado de una palabra está dado en términos de ocurrencias en documentos.
 - Bag of Words (BOW)
 - Term Frequency - Inverse Document Frequency (TF-IDF)
- Modelos basados en redes neuronales:
 - Clásicos: Word2Vec, GloVe, ...
 - LLMs: GPT, LLaMA, DeepSeek, Gemma, Qwen, ...



- Modelos basados en conteos:
 - + Modelo sencillo y simple
 - + Interpretabilidad
 - Vectores raros (*sparse*)
 - Alta dimensionalidad (del orden de miles o más)
- Modelos basados en redes neuronales:
 - + Vectores densos
 - + Menor dimensionalidad (del orden de cientos)
 - Algunos pueden ser computacionalmente caros de obtener



- La **similitud coseno** es un valor entre -1 y 1 y está dada por

$$\text{sim}(u, v) = \frac{u \cdot v}{\|u\| \|v\|} = \cos(\theta_{uv})$$

donde $u \cdot v$ es el producto punto y $\|u\|$ es la norma del vector. Podemos normalizar $\|u\| = \|v\| = 1$ y tenemos

$$\text{sim}(u, v) = u \cdot v$$

En general, no usamos la distancia Euclidiana.

Ejemplo



- La distancia angular es un valor entre 0 y π dado por

- En ocasiones, nos referimos a la métrica coseno como

En general, no usamos la distancia Euclidiana.

◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ ↺ 🔍 ↻ 13/26

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.

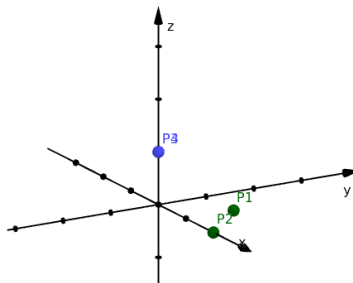
Texto	Europa	cambios	arroz	pescado
1	1	1	0	0
2	2	0	0	0
3	0	0	1	0
4	0	0	1	1



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.





Preguntas

Procesamiento
de Lenguaje
Natural

Introducción

- ¿Qué obtenemos si sumamos las filas?
- ¿Qué obtenemos si sumamos las columnas?
- ¿Qué tamaño tiene la matriz anterior?
- ¿Qué palabras tenderán a dominar la matriz?
- ¿Qué significa que dos palabras sean similares usando estas representaciones vectoriales?



¿Qué significa que dos palabras sean similares usando estas representaciones vectoriales?

Las palabras similares tienen vectores similares porque suelen aparecer en documentos similares. La matriz término-documento nos permite representar el significado de una palabra por los documentos en los que suele aparecer.



la matriz término-término (también denominada matriz palabra-palabra o matriz término-contexto) es la matriz de tamaño $|V| \times |V|$ donde las columnas y filas están etiquetadas por palabras. La entrada ij es el número de veces que la palabra i (objetivo) y la palabra j (contexto) coinciden en algún contexto (ventana) en algún corpus de entrenamiento.

Doc₁: *I go to school every day by bus.*

Doc₂: *I go to theatre every night by bus.*



Usemos una ventana de contexto de 2 palabras.

Doc₁: *I go to school every day by bus.*

Doc₂: *I go to theatre every night by bus.*

	bus	by	day	every	go	i	night	school	theatre	to
bus	0	2	1	0	0	0	1	0	0	0
by	2	0	1	2	0	0	1	0	0	0
day	1	1	0	1	0	0	0	1	0	0
every	0	2	1	0	0	0	1	1	1	2
go	0	0	0	0	0	2	0	1	1	2
i	0	0	0	0	2	0	0	0	0	2
night	1	1	0	1	0	0	0	0	1	0
school	0	0	1	1	1	0	0	0	0	1
theatre	0	0	0	1	1	0	1	0	0	1
to	0	0	0	2	2	2	0	1	1	0



Las matrices término-documento se definieron originalmente como un medio de encontrar documentos similares para la tarea de recuperación de información documental. Dos documentos que son similares tenderán a tener palabras similares, y si dos documentos tienen palabras similares sus vectores columna tenderán a ser similares.

IR es la tarea que consiste en buscar, localizar y presentar información que coincida con la consulta de búsqueda o la necesidad de información de un usuario.



Doc₁: ... se necesitan dos ingredientes: agua y harina. Mezclar los ingredientes comenzando con la harina ...

Doc₂: Calentar los ingredientes en una sartén con agua...

Doc3: *Las playas en México tienen agua templada...*

◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ ↺ 🔍 ↻ 21/26



...

Doc₃: *Las playas en México tienen agua templada...*

	Doc ₁	Doc ₂	Doc ₃
agua	1	2	1
ingredientes	1	2	0
...

Con la métrica euclidiana, el más similar a Doc_1 es Doc_3 . Con la métrica coseno, es Doc_2 .



Documentos:

Palabras consideradas:



TF-IDF

- **Term-frequency:** Para un término t en un documento d :

$$\text{tf}_{t,d} = \begin{cases} 1 + \log_{10} \text{count}(t, d), & \text{si } \text{count}(t, d) > 0 \\ 0, & \text{si } \text{count}(t, d) = 0 \end{cases}$$

- **Inverse document frequency:** Para un término t en una colección de N documentos:

$$\text{idf}_t = \log_{10} \left(\frac{N}{\text{df}_t} \right)$$

donde df_t es el número de documentos donde aparece el término t .

$$\text{TF-IDF}(t, d) = \text{tf}_{t,d} \cdot \text{idf}_t$$



Doc₃: *Las playas en México tienen agua templada...*

◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ ↺ 🔍 ↻ 24/26



BOW:

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

TF-IDF:

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	0.074	0	0.22	0.28
good	0	0	0	0
fool	0.019	0.021	0.0036	0.0083
wit	0.049	0.044	0.018	0.022



Aplicaciones Adicionales

Una aplicación adicional de estos modelos es la extracción de features del texto para tareas de Machine Learning. Es importante reflexionar sobre qué rasgos del texto captan estas features.

- Clasificación (Análisis de Sentimientos, etc.).
- Segmentación (Detección de tópicos, etc.).
- Similitud de documentos (IR, Autoría, etc.).