Procesamiento de Lenguaje Natural

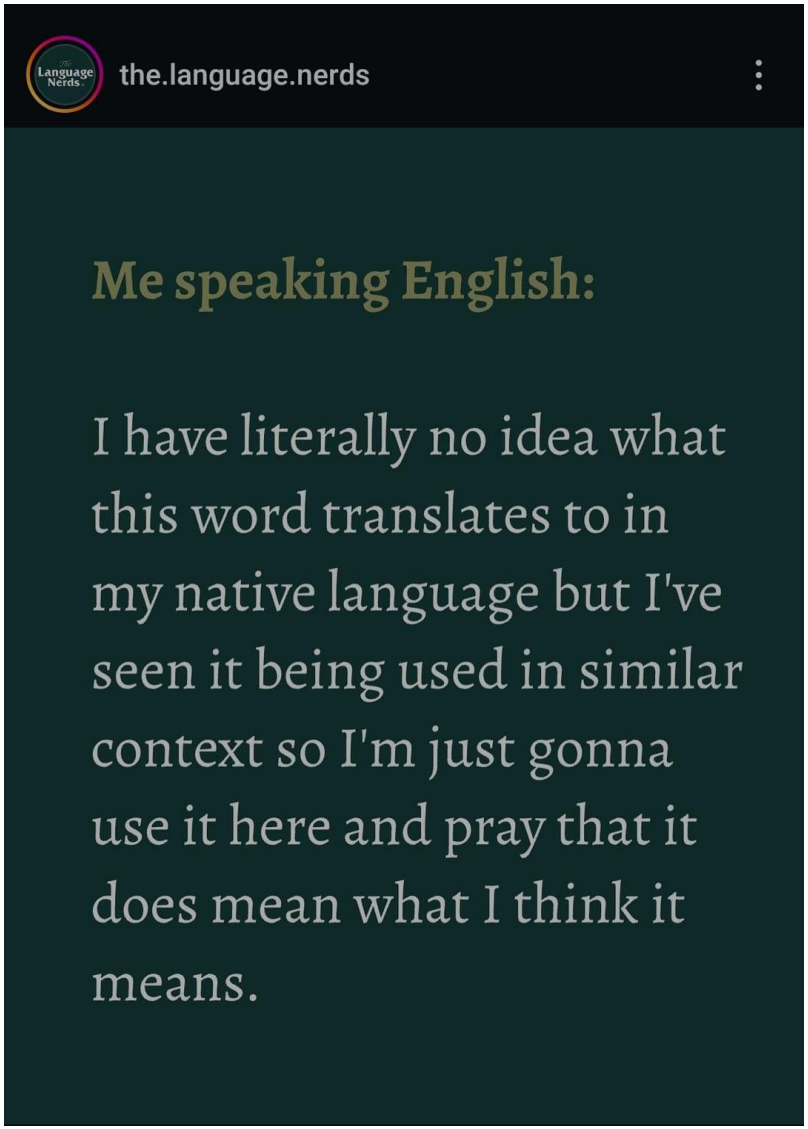
Vectores Semánticos

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.

Referencias

Chapter 6. Jurafsky, D., Martin, J. H. (2019). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*.

La Hipótesis Distribucional

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 oftamologo) 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

¿Qué es el **Ongchoi**?

Ejemplo

Veamos algunos contextos en los que ocurre **ongchoi**:

- **Ongchoi** is delicious sauteed with garlic.
- **Ongchoi** is superb over rice.
- ...**ongchoi** leaves with salty sauces...

Ahora, veamos algunos contextos de los contextos anteriores:

- ...spinach sauteed with garlic over rice...
- ...chard stems and leaves are delicious...
- ...collard greens and other salty leafy greens

Respuesta

La semántica vectorial

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 cercania de embeddings da cuenta de diversos fenómenos, además de la smilitud de palabras.

Ejemplo

¿Por qué es útil la semántica vectorial?

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 predefinidas (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 de semántica vectorial

- Módelos 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)
- Módelos basados en redes neuronales:
  - Clásicos: `Word2Vec`, `GloVe`, ...
  - LLMs: `GPT`, `LLaMA`, `Jamba`, `Gemma`, ...

Modelos de semántica vectorial

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

¿Cómo medimos la similitud/distancia entre vectores?

Consideremos dos embeddings $u, v \in \mathbb{R}^D$.

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

$$\text{sim}(u, v) = \frac{u \cdot v}{\|u\|\|v\|} = \cos(u0v)$$

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

¿Cómo medimos la similitud/distancia entre vectores?

Consideremos dos embeddings $u, v \in \mathbb{R}^D$.

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

$$d_\theta(u, v) = \arccos(\text{sim}(u, v))$$

- En ocasiones, nos referimos a la métrica coseno como

$$d_{\cos}(u, v) = 1 - \text{sim}(u, v)$$

En general, no usamos la distancia Euclidiana.

Ejemplo

La matriz term-document

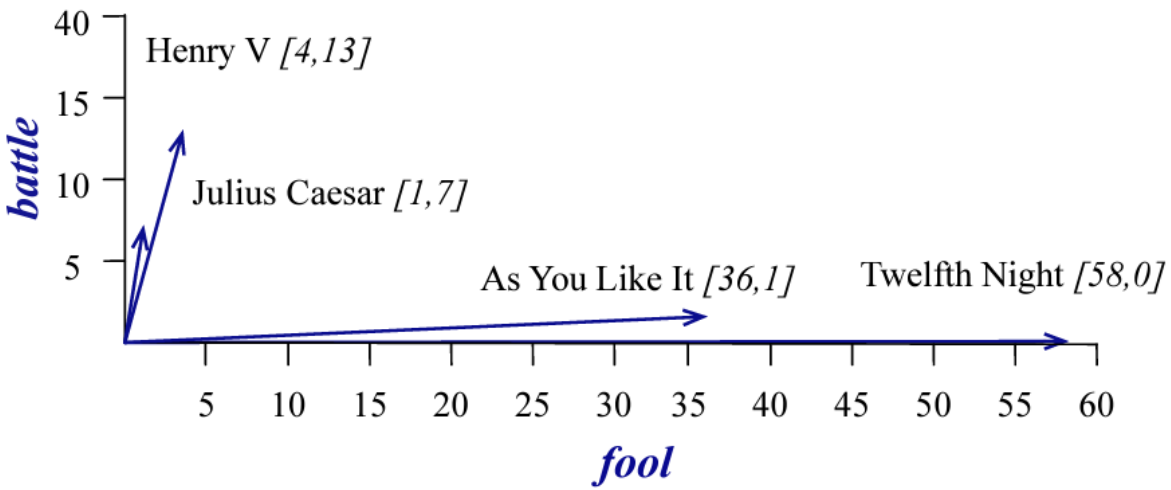
En una matriz término-documento, cada fila representa una palabra del vocabulario y cada columna representa un documento de alguna colección de documentos.

	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

Modelo 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

El modelo BOW (bag of words) asigna a cada documento el vector correspondiente a la columna. El vector de cada palabra es su fila.



Preguntas

- ¿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?

Vectores de palabras

¿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 term-term

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.

Doc_1:	*I go to school every day by bus.*
Doc_2:	*I go to theatre every night by bus.*

La matriz term-term

Usemos una ventana de contexto de 2 palabras.

Doc_1:	*I go to school every day by bus.*
Doc_2:	*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

Information Retrieval

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.

Information Retrieval

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.

Métrica Coseno vs Distancia Euclidiana

¿Por qué la similitud coseno (o métrica angular) captura mejor la similitud?

Doc_1:	*... se necesitan dos ingredientes: agua y harina. Mezclar los ingredientes comenzando con la harina ...*
Doc_2:	*Calentar los ingredientes en una sarten con agua...*
Doc_3:	*Las playas en México tienen agua templada...*

	Doc_1	Doc_2	Doc_3
agua	1	2	1
ingredientes	1	2	0
...	...	...	...

Métrica Coseno vs Distancia Euclidiana

$\texttt{Doc}_1$:	*... se necesitan dos ingredientes: agua y harina. Mezclar los ingredientes comenzando con la harina ...*
$\texttt{Doc}_2$:	*Calentar los ingredientes en una sarten con agua...*
$\texttt{Doc}_3$:	*Las playas en México tienen agua templada...*

	$\texttt{Doc}_1$	$\texttt{Doc}_2$	$\texttt{Doc}_3$
agua	1	2	1
ingredientes	1	2	0
...	...	...	...

Con la métrica euclidiana, el más similar a $\texttt{Doc}_1$ es $\texttt{Doc}_3$. Con la métrica coseno, es $\texttt{Doc}_2$.

TF-IDF

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

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

- **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.

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

Ejemplo

Doc_1:	*... se necesitan dos ingredientes: agua y harina. Mezclar los ingredientes comenzando con la harina ...*
Doc_2:	*Calentar los ingredientes en una sarten con agua...*
Doc_3:	*Las playas en México tienen agua templada...*

	Doc_1	Doc_2	Doc_3
agua	...	...	...
ingredientes	...	...	...
harina	...	...	...
México	...	...	...

Ejemplo

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).