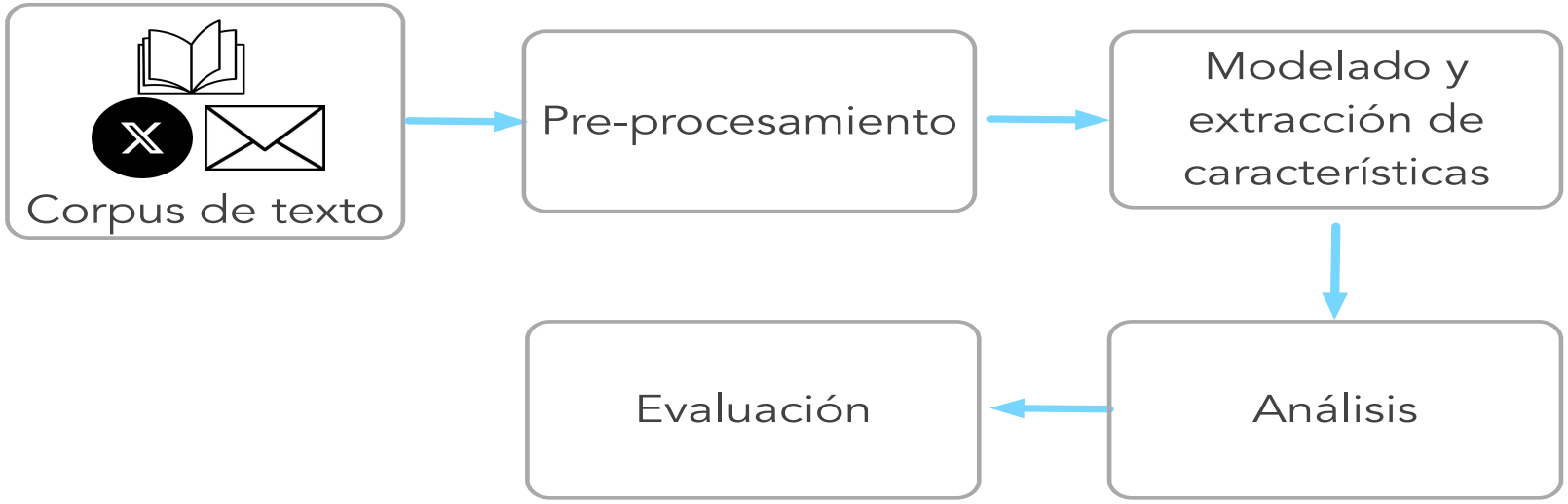
UNIVERSIDAD HEMISFERIOS

Procesamiento del lenguaje

Ph.D. Janneth Chicaiza Espinosa

Modelado de lenguaje y extracción de características

Unidad 2



OBJETIVO: Aplicar técnicas de representación de textos y extracción de características para resolver tareas de NLP comunes, considerando la naturaleza de los textos y los objetivos de la tarea específica a implementar.

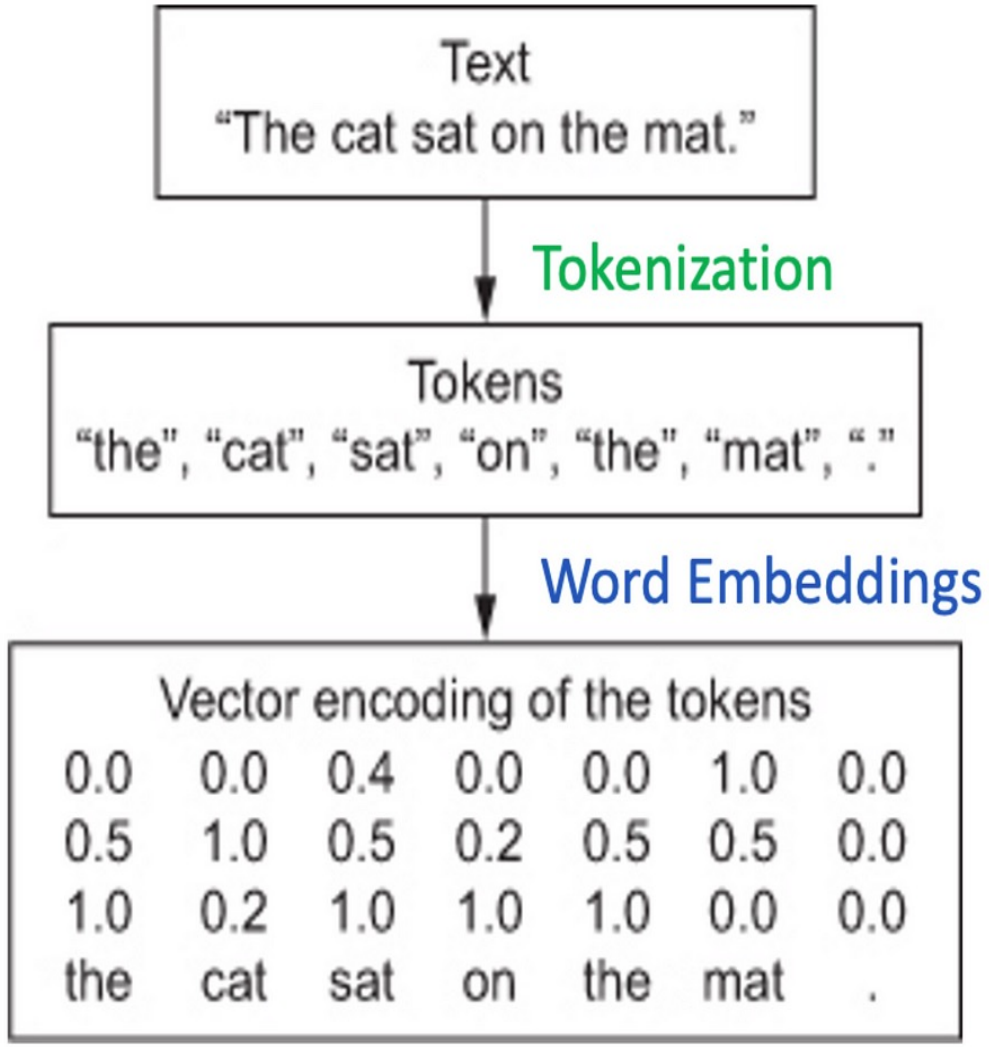
Las palabras como símbolos discretos

- En el enfoque clásico, cada palabra se trata como un **símbolo atómico,** es decir, es una unidad indivisible y única.
- En este enfoque, las palabras no tienen características internas que puedan utilizarse para compararlas entre sí. Es decir, dos palabras son **iguales** o son **distintas**.
- Esto significa que **no se pueden calcular similitudes o diferencias entre palabras de manera efectiva**, ya que cada palabra es considerada totalmente distinta de cualquier otra, sin importar su significado o similitud en contexto.
- Para superar estas limitaciones, se desarrollaron representaciones vectoriales de palabras (como Word2Vec) en las que cada palabra se representa como un vector en un espacio multidimensional, lo que permite calcular distancias y similitudes entre palabras de una forma más significativa.

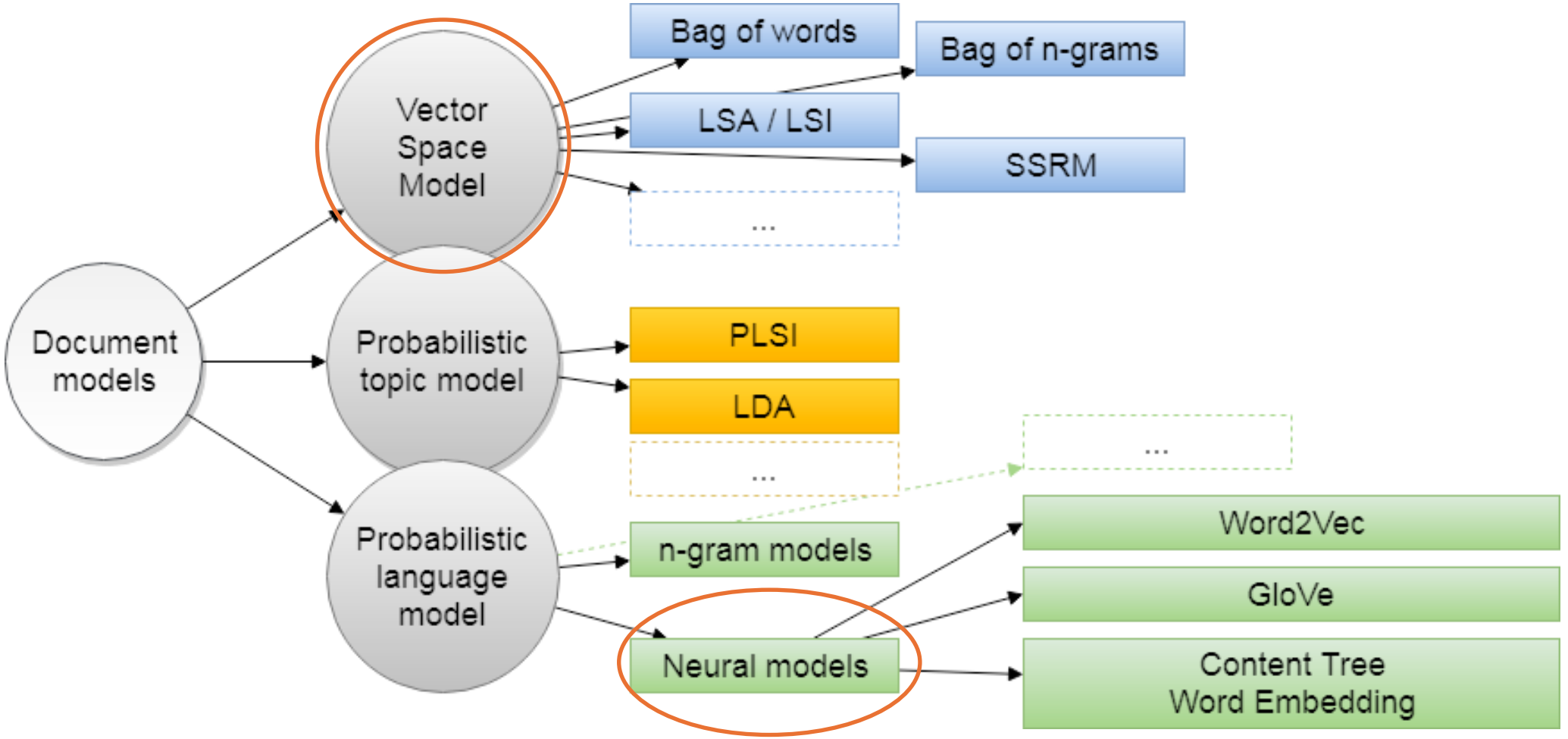
Con una representación simbólica
De la palabras no podríamos
determinar la relación entre
“perro”, “banco” y ”ahorro”

Representaciones distribuidas de las palabras

- En lugar de usar representaciones simbólicas para las palabras, podemos representar el significado de una palabra o concepto mediante un vector de números, donde cada elemento del vector contribuye a la codificación del significado.
- La representación de texto es un campo clave en NLP. Se refiere a la transformación de textos en estructuras matemáticas que permiten a las computadoras entenderlos y procesarlos de forma efectiva.

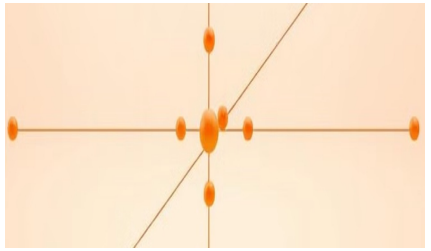


Modelos para la representación de documentos



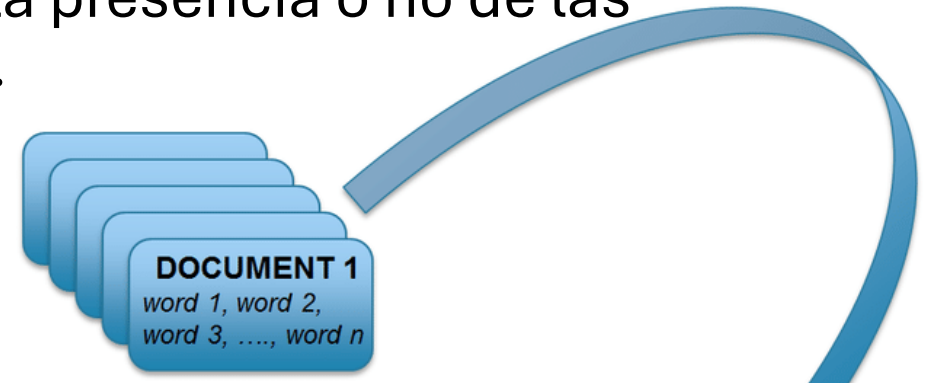
Source:link

Modelos espacio-vectoriales



- **Representan, de forma general, los documentos como vectores en un espacio multidimensional**. Cada dimensión corresponde a una palabra del vocabulario.
- Las representaciones que generan se basan en la presencia o no de las palabras en los documento (o en su frecuencia).

La magnitud de cada dimensión refleja la frecuencia o importancia de la palabra en el documento.
Similitud: La similitud entre documentos se calcula como la distancia entre sus vectores, lo que permite clasificar y buscar documentos relacionados.



	document 1	document 2	…	document N
word 1	1	0	…	1
word 2	0	1	…	0
…	…	…	…	0
word n	0	0	…	1

Bolsa de Palabras (BoW)

BoW es una representación que describe la presencia de palabras dentro de un documento sin considerar el orden o el contexto de las palabras.

Involucra dos pasos:

- Creación de vocabulario
- Vectorización: cuenta la frecuencia de cada palabra en un texto y se coloca en un vector de tamaño fijo.

	the	red	dog	cat	eats	food
1. the red dog →	1	1	1	0	0	0
2. cat eats dog →	0	0	1	1	1	0
3. dog eats food →	0	0	1	0	1	1
4. red cat eats →	0	1	0	1	1	0

BoW: representación basada en ponderación TF (pesos)

d0
d1
d2
d3

```
['El pescado al horno es el mejor.',
 'El servicio es insuperable',
 'El pescado al horno es insuperable, pero el servicio es malo.',
 'El mejor servicio es el pescado al horno.']
```

	al	el	es	horno	insuperable	malo	mejor	pero	pescado	servicio
d0	1	2	1	1	0	0	1	0	1	0
d1	0	1	1	0	1	0	0	0	0	1
d2	1	1	1	1	1	0	0	0	1	0
d3	1	2	1	1	0	0	1	0	1	1

Pros y contras de BoW

Pros:

- Es fácil de implementar y útil en problemas de clasificación y modelado de temas (eficiencia).

Contras:

- Pierde toda la información sobre el orden y estructura de las palabras.
- BoW tiende a producir representaciones muy dispersas (“sparse”)
- BoW sólo mira a la forma superficial de las palabras, ignorando toda información semántica de las mismas, por tanto, puede haber problemas “semánticos” con la polisemia y la sinonimia.
    - **Por ejemplo**: La  de la  se quebró.

TF – IDF: BoW ponderado (pesos normalizados)

- TF-IDF es una extensión de BoW que intenta corregir algunas de sus limitaciones, asignando un **peso** a cada palabra en función de su frecuencia en el documento (TF) y su frecuencia inversa en el conjunto de documentos (IDF).
- **Funcionamiento**: TF-IDF combina dos componentes:
  - **TF (Term Frequency):** Es la frecuencia de una palabra en un documento específico.
  - **IDF (Inverse Document Frequency):** Calcula la importancia de una palabra en todo el conjunto de documentos, de modo que las palabras que aparecen en muchos documentos (como "y", "el", "de") tengan un peso menor.
- Ventaja sobre BoW: Al reducir el peso de palabras comunes y aumentar el peso de palabras que son específicas de ciertos documentos, TF-IDF ofrece una representación más informativa dentro del modelo espacio-vectorial.

d0 ['El pescado al horno es el mejor.',

d1 'El servicio es insuperable',

d2 'El pescado al horno es insuperable, pero el servicio es malo.',

d3 'El mejor servicio es el pescado al horno.']

TF

	al	el	es	horno	insuperable	malo	mejor	pero	pescado	servicio
d0	1	2	1	1	0	0	1	0	1	0
d1	0	1	1	0	1	0	0	0	0	1
d2	1	1	1	1	1	0	0	0	1	0
d3	1	2	1	1	0	0	1	0	1	1

w

	al	el	es	horno	insuperable	malo	mejor	pero	pescado	servicio
d0	0,35651	0,58294	0,29147	0,35651	0	0	0,44036	0	0,35651	0
d1	0	0,41599	0,41599	0	0,62849	0	0	0	0	0,50882
d2	0,25172	0,41160	0,41160	0,25172	0,31092	0,39437	0	0,39437	0,25172	0,25172
d3	0,33581	0,54909	0,27455	0,33581	0	, 0	0,41479	0	0,33581	0,33581

IDF

al	el	es	horno	insuperabl	malo	mejor	pero	pescado	servicio
1,223144	1,000000	1,000000	1,223144	1,510826	1,916291	1,510826	1,916291	1,223144	1,223144

¿Qué pasa si preprocesamos el texto?

review	no_stopwords
El pescado al horno es el mejor.	[pescado, horno, mejor]
El servicio es insuperable	[servicio, insuperable]
El pescado al horno es insuperable, pero el servicio es malo.	[pescado, horno, insuperable, servicio, malo]
El mejor servicio es el pescado al horno.	[mejor, servicio, pescado, horno]

IDF

horno	insuperable	malo	mejor	pescado	servicio
1,2231	1,5108	1,9163	1,5108	1,2231	1,2231

w

	horno	insuperable	malo	mejor	pescado	servicio
d0	0,5326	0	0	0,6578	0,5326	0
d1	0	0,7772	0	0	0	0,6292
d2	0,3785	0,4675	0,5930	0	0,3785	0,3785
d3	0,4701	0	0	0,5806	0,4701	0,4701

Modelos probabilísticos de tópicos

Tópicos

El modelo probabilístico de tópicos supone que los documentos están compuestos por una mezcla de tópicos.

Probabilidades

Cada tópico está asociado con una distribución de probabilidad sobre las palabras del vocabulario.

Inferencia

El modelo infiere la probabilidad de que un documento pertenezca a un tópico dado.

1

Método: LDA

LDA permite una representación más profunda de los documentos.

2

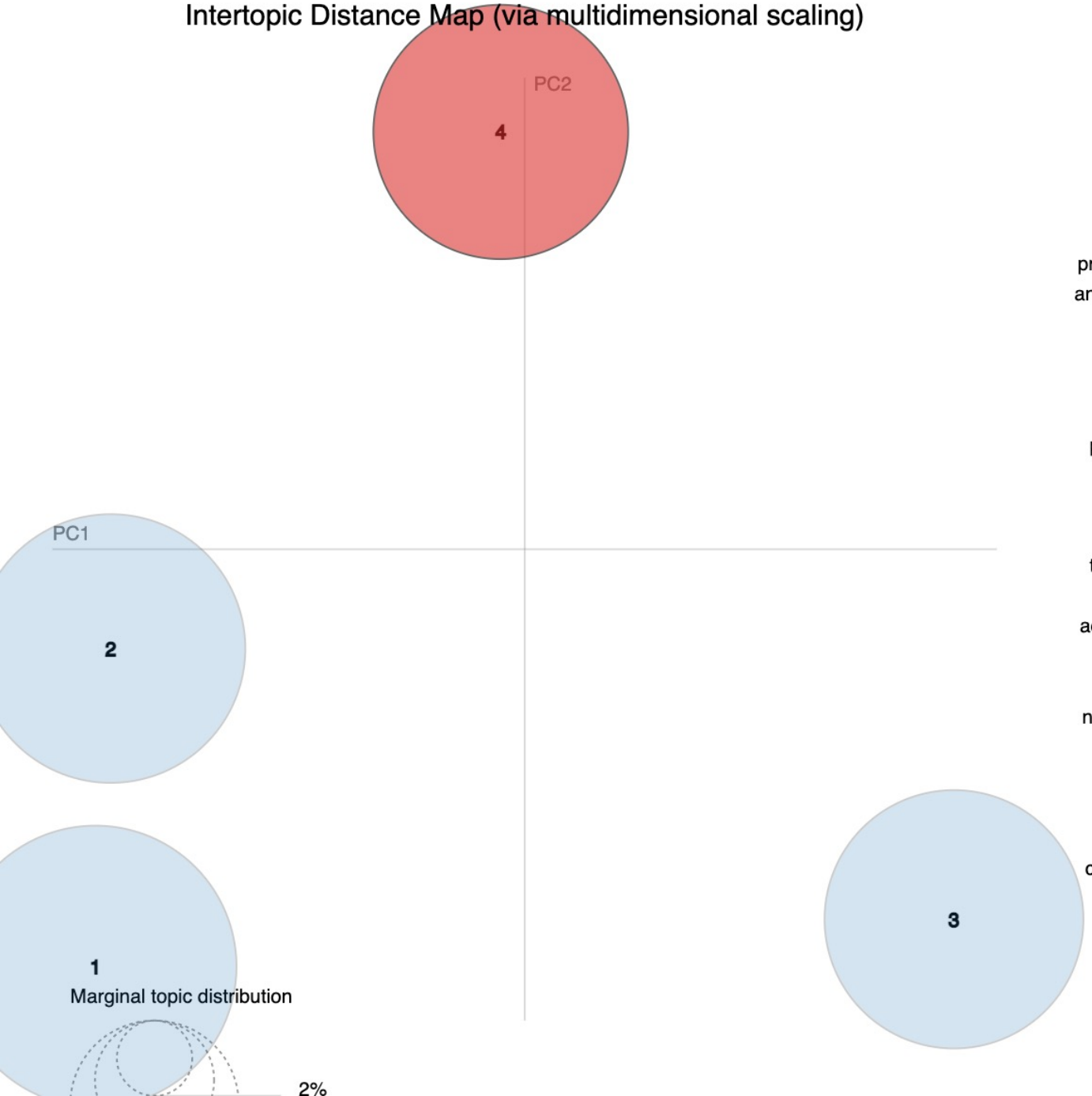
Ventajas

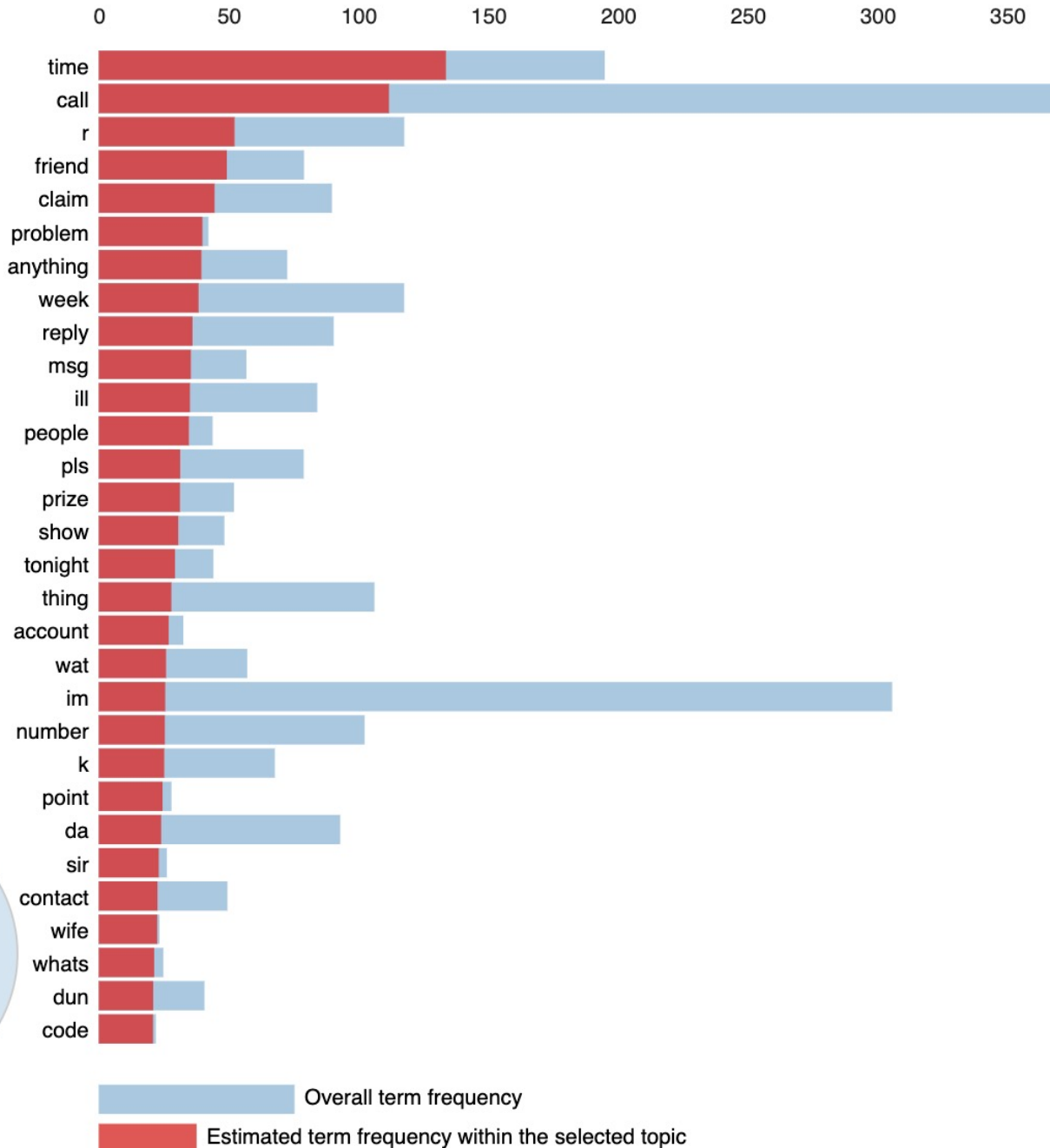
Captura la estructura latente de los documentos y permite la extracción de tópicos relevantes.

3

Usos

Se utiliza en tareas como la clasificación de documentos, y la recomendación de contenido.

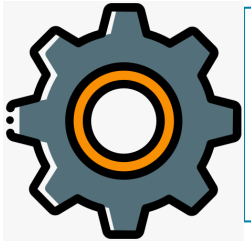

Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
4
2
3
1
Marginal topic distribution
2%


Top-30 Most Relevant Terms for Topic 4 (22.9% of tokens)
0
50
100
150
200
250
300
350
time
call
r
friend
claim
problem
anything
week
reply
msg
ill
people
pls
prize
show
tonight
thing
account
wat
im
number
k
point
da
sir
contact
wife
whats
dun
code
Overall term frequency
Estimated term frequency within the selected topic

Bloques temáticos y actividades

1. Fundamentos y pre-procesamiento del lenguaje

- ✓ **Test de la Unidad 1** A
- ✓ Caso práctico 1: pre-procesamiento de texto



2. Modelado de lenguaje y extracción de características

- ✓ Test de la Unidad 2
- ✓ **Caso práctico 2: representación de texto** A

Gracias

uhemisferios uhe.oficial uhe_oficial Universidad Hemisferios

uhemisferios.edu.ec