Programa de formación
MACHINE LEARNING AND DATA SCIENCE MLDS

Facultad de
INGENIERÍA

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Módulo 4

Procesamiento y Entendimiento del Lenguaje Natural

Unidad 3

Embeddings

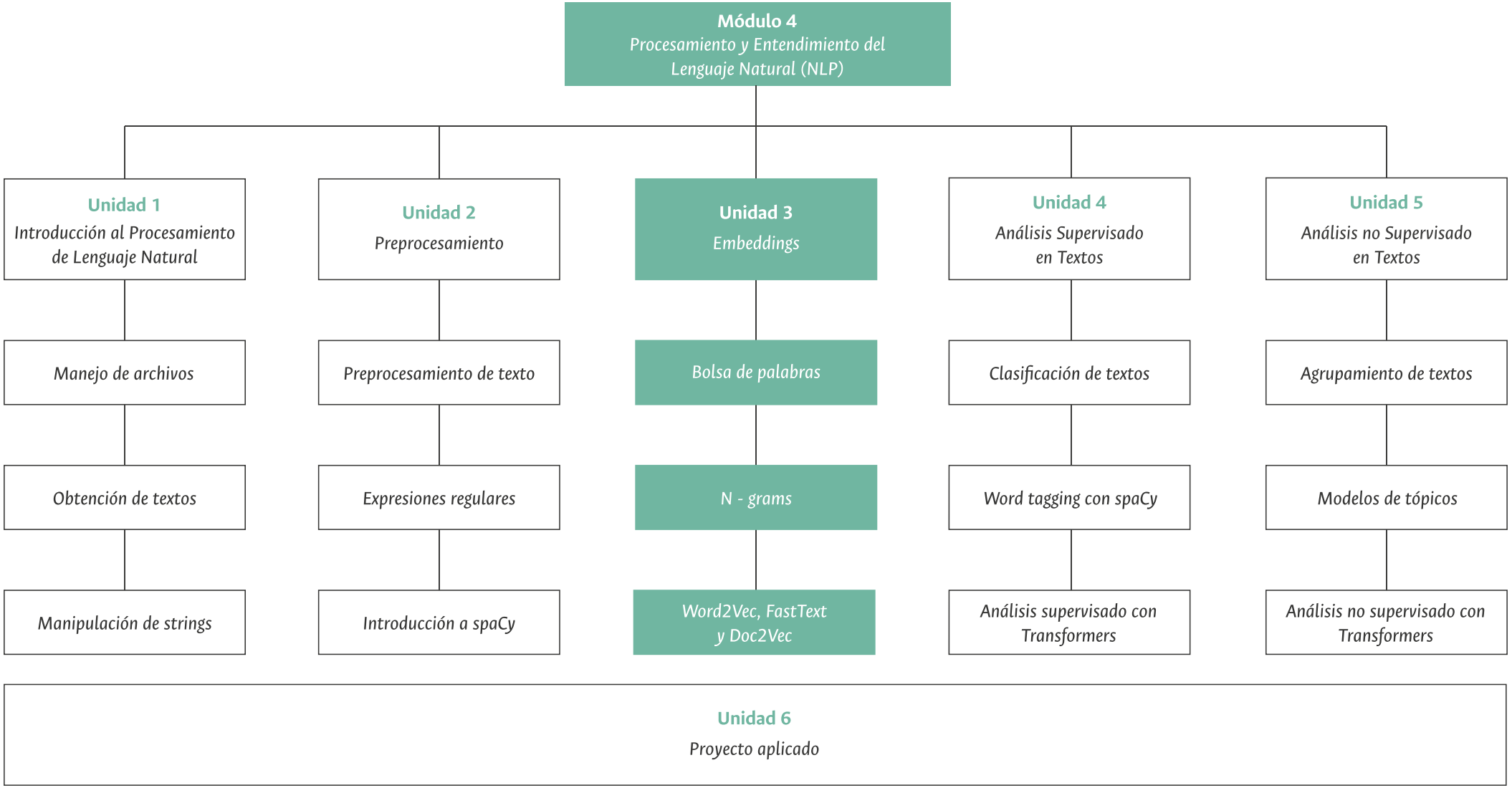
Clase sincrónica

Felipe Restrepo Calle, PhD.

Facultad de INGENIERÍA

UNIVERSIDAD NACIONAL DE COLOMBIA

Mapa de contenidos



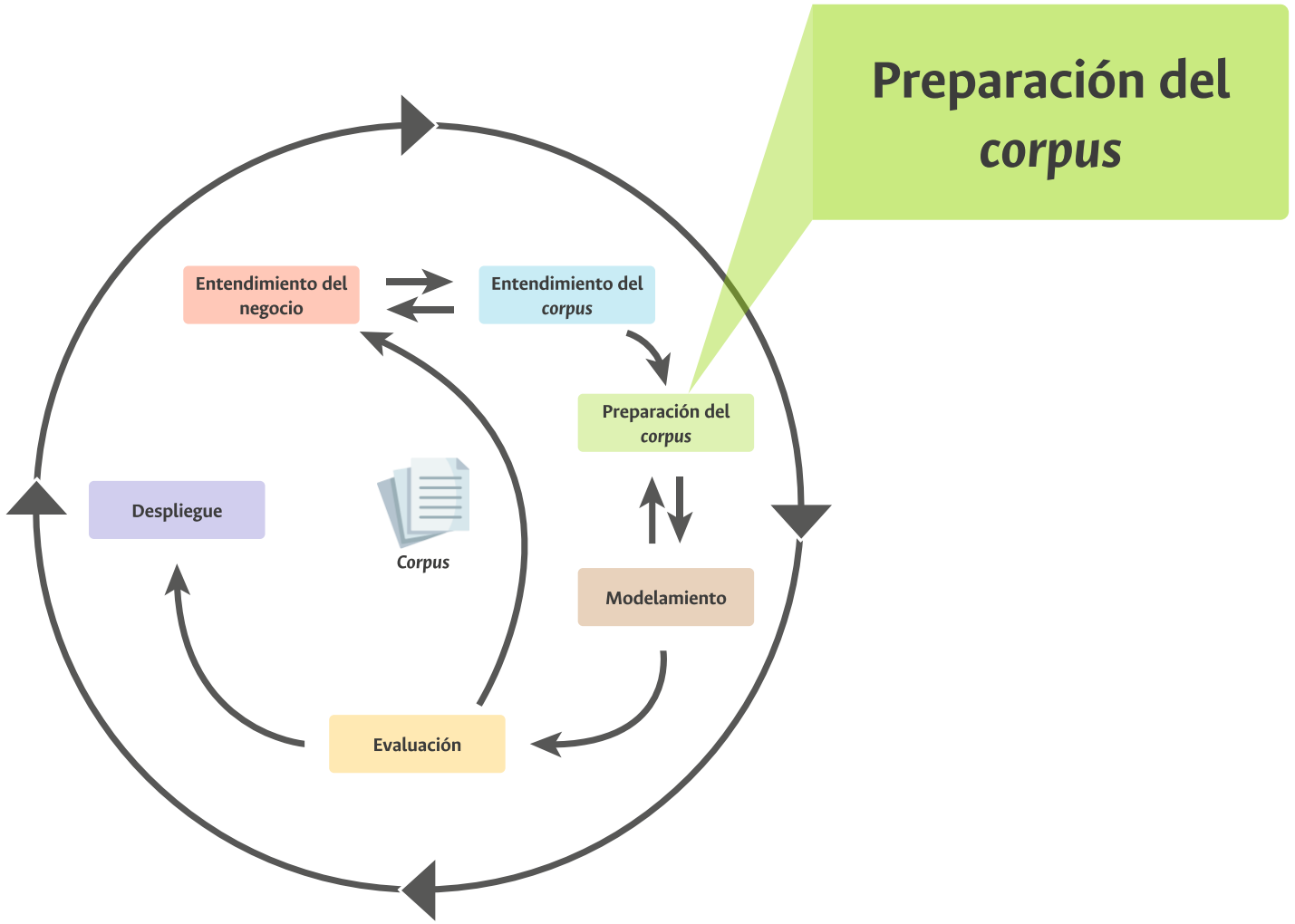
Agenda

1. Definición
2. Representaciones basadas en conteos
   - 2.1 Bolsas de palabras
   - 2.2 Bolsas de N-gramas
   - 2.3 TF-IDF
3. Modelos semánticos
   - 3.1 Word2Vec
   - 3.2 FastText
   - 3.3 Doc2Vec
4. Medidas de similitud

Etapa actual del ciclo de vida en NLP

Además del **preprocesamiento**, la representación de textos o **extracción de características** es un proceso importante dentro del ciclo de vida de NLP.

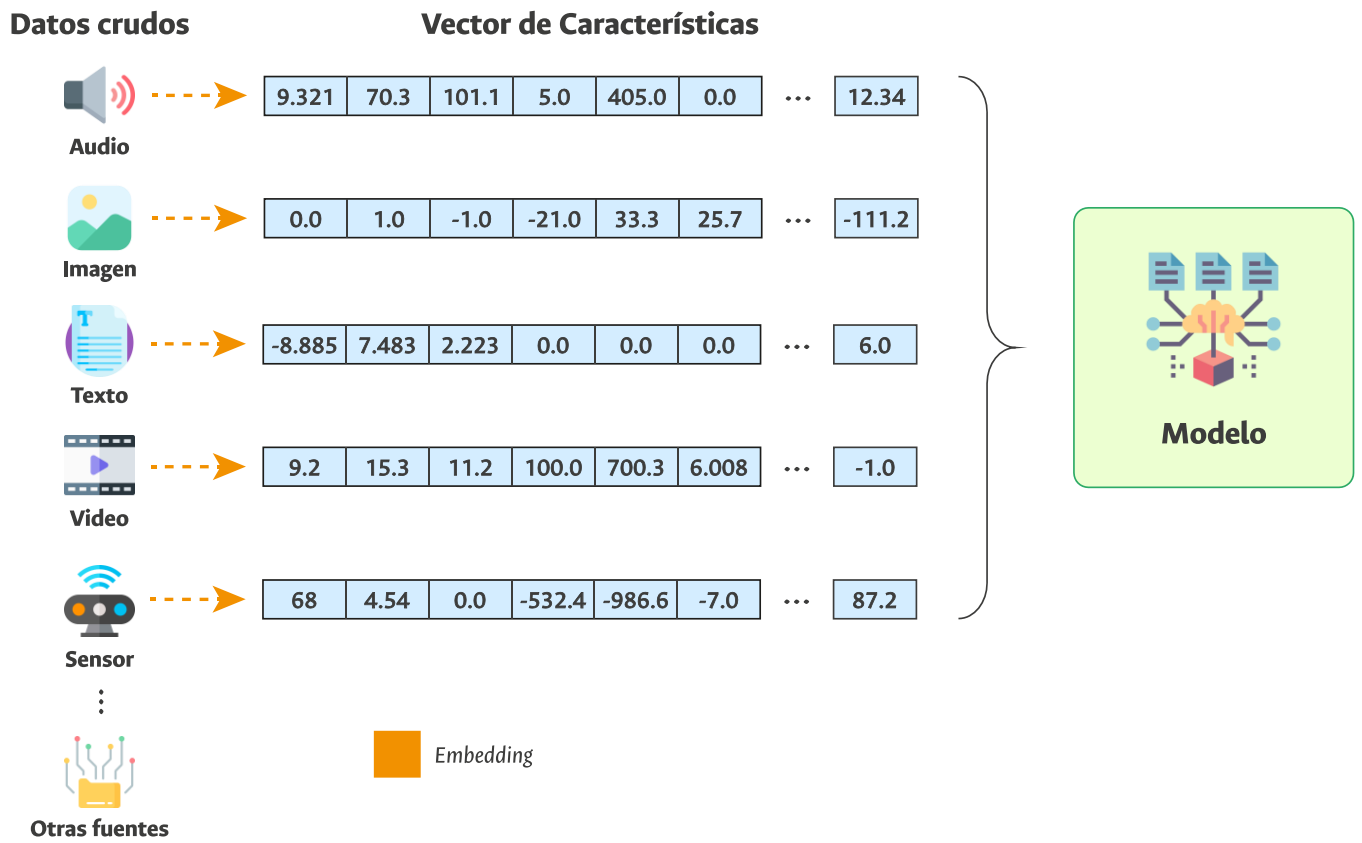
Hace parte de la etapa de **preparación del *corpus*** y es la base para un posterior **modelamiento**.



Embeddings

Los ***embeddings*** son representaciones vectoriales de distintos tipos de objetos.

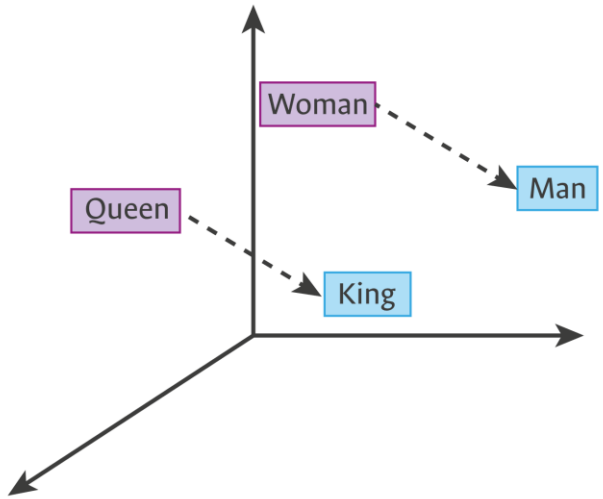
Se utilizan para capturar relaciones entre objetos y permitir a los modelos de aprendizaje automático trabajar con ellos.



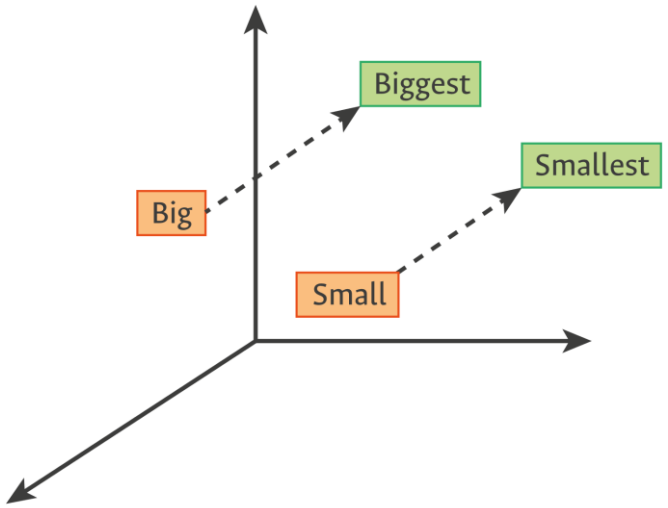
Definición

Embedding de texto

Relación semántica



Relación sintáctica



El *embedding* de texto es una representación vectorial de una secuencia de palabras (como un documento o una oración) en un espacio de características de baja dimensión.

Su propósito es capturar la semántica y el significado de las palabras y permitir a los modelos de aprendizaje automático trabajar con una representación numérica en lugar de texto.

Definición

Técnicas

Existen 2 tipos de ***embeddings*** de texto:

Modelos basados en conteos

Son fáciles de implementar y entender.

Modelos semánticos

Codifican mayor información de los lenguajes naturales.

Característica	Basadas en Conteos	Modelos Semánticos
Interpretabilidad	✓	✗
Escalabilidad computacional	✓	✗
Contexto	✗	✓
Sinonimia	✗	✓

Agenda

1. Definición
2. **Representaciones basadas en conteos**
   - 2.1 Bolsas de palabras
   - 2.2 Bolsas de N-gramas
   - 2.3 TF-IDF
3. Modelos semánticos
   - 3.1 Word2Vec
   - 3.2 FastText
   - 3.3 Doc2Vec
4. Medidas de similitud

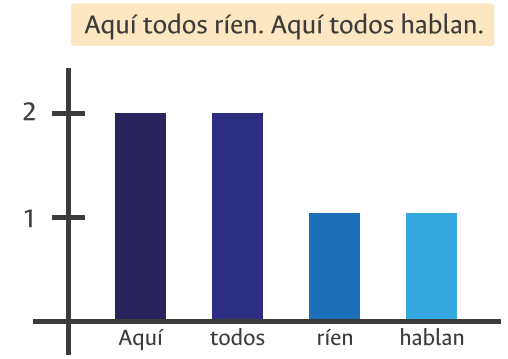
Modelos basados en Conteos

Los **modelos basados en conteos** representan palabras como vectores basados en su frecuencia de aparición en el *corpus* de texto.

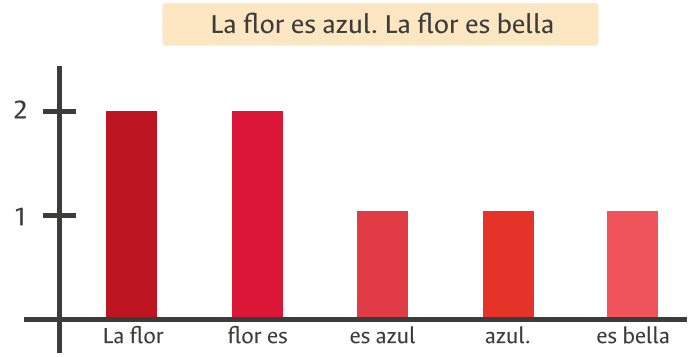
Son utilizados en tareas NLP como la clasificación de documentos y la agrupación de palabras similares.

Normalmente los usamos como primera aproximación en aplicaciones de alto desempeño y con *corpus* pequeños.

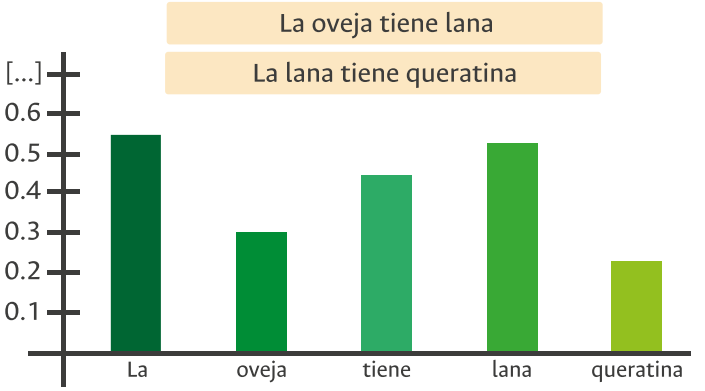
Bolsa de palabras



N – gramas (Bigramas)



TF – IDF



Modelos basados en Conteos

Bolsa de Palabras

Una **bolsa de palabras** es una representación numérica de texto que muestra cada documento como un vector de conteos de palabras.

Se construye un vocabulario de todas las palabras en un *corpus* de texto y luego se cuentan las apariciones de cada palabra en cada documento.

el perro es marrón

el gato está con el perro

	el	perro	es	marrón	gato	está	con
	1	1	1	1	0	0	0
	2	1	0	0	1	1	1

Modelos basados en Conteos

Bolsa de Palabras

- Se utiliza ampliamente en tareas NLP como: clasificación de documentos y agrupación de palabras similares.
- Son la base de modelos de tópicos y sistemas de recuperación de información.

Modelos basados en Conteos

Bolsa de N-gramas

Una **bolsa de N-gramas** es una representación numérica de texto que considera secuencias consecutivas de *tokens* en lugar de simplemente palabras individuales.

Los **N-gramas** o secuencias se pueden construir a nivel de caracter o a nivel de palabra.

Se construye un vocabulario de N-gramas a partir de un *corpus* de texto y se cuentan las apariciones de cada una en cada documento.

Oración: "*Albert Einstein era un científico. Albert Einstein era alemán*".

Unigrama	
t_1	Conteo
Albert	2
Einstein	2
era	2
un	1
científico	1
.	1
alemán	1

Bigrama		
t_1	t_2	Conteo
Albert	Einstein	2
Einstein	era	2
era	un	1
un	científico	1
científico	.	1
.	Albert	1
era	alemán	1

Modelos basados en Conteos

N-gramas a nivel de caracter

Oración: "*universidad universo*"

Unigrama	
c_1	Conteo
u	2
n	2
i	3
v	2
e	2
r	2
s	2
d	2
a	1
_	1
o	1

_ = *Espacio en blanco*

Bigrama	
c_1c_2	Conteo
un	2
ni	2
iv	2
ve	2
er	2
rs	2
si	1
id	1
da	1
ad	1
d_	1
_u	1
so	1

Modelos basados en Conteos

Bolsa de N-gramas

- Se utiliza para mejorar la representación del texto en tareas NLP, especialmente en tareas que requieren una comprensión más profunda del contexto.
- Son la base de muchos modelos de autocompletado, correctores de ortografía, modelos de identificación del lenguaje, entre otros.

Corrector de ortografía

encontré

Estoy felis porque encontre la libreria.

feliz librería

Autocompletado

Modelos basados en Conteos

TF–IDF

Term Frequency - Inverse Document Frequency (**TF–IDF**) es una representación que combina la frecuencia de una palabra en un documento (**TF**) con su rareza a nivel de corpus (**IDF**).

Para ello se calcula la frecuencia de una palabra en un documento y se multiplica por el logaritmo inverso de su frecuencia en el corpus de texto.

$$\mathrm{TFIDF}(t_j, d_i) = \underbrace{\mathrm{TF}(t_j, d_i)}_{\text{Frecuencia de aparición de una palabra } \boldsymbol{t} \text{ en un documento } \boldsymbol{d}.} \times w_j$$

Frecuencia de la palabra en un conjunto de documentos (IDF).

$$w_j = 1 + \log\left(\frac{n}{1 + df\,(t_j)}\right)$$

n = cantidad de documentos.

$df\,(t_j)$ = frecuencia de documentos de la palabra t.

Esta representación es especialmente útil en las siguientes tareas:

Clasificación

Ayuda a encontrar representaciones más acertadas de los textos, lo que simplifica el entrenamiento de muchos modelos.

Búsqueda de información

Se utiliza para mejorar la relevancia de los resultados por ponderación de términos.

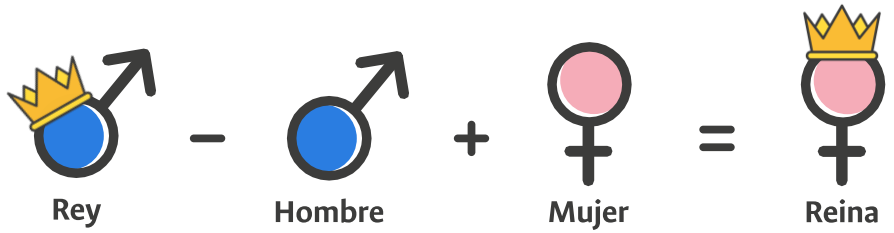
Análisis de textos

Permite obtener las palabras más relevantes de un *corpus*, simplificando su análisis.

Agenda

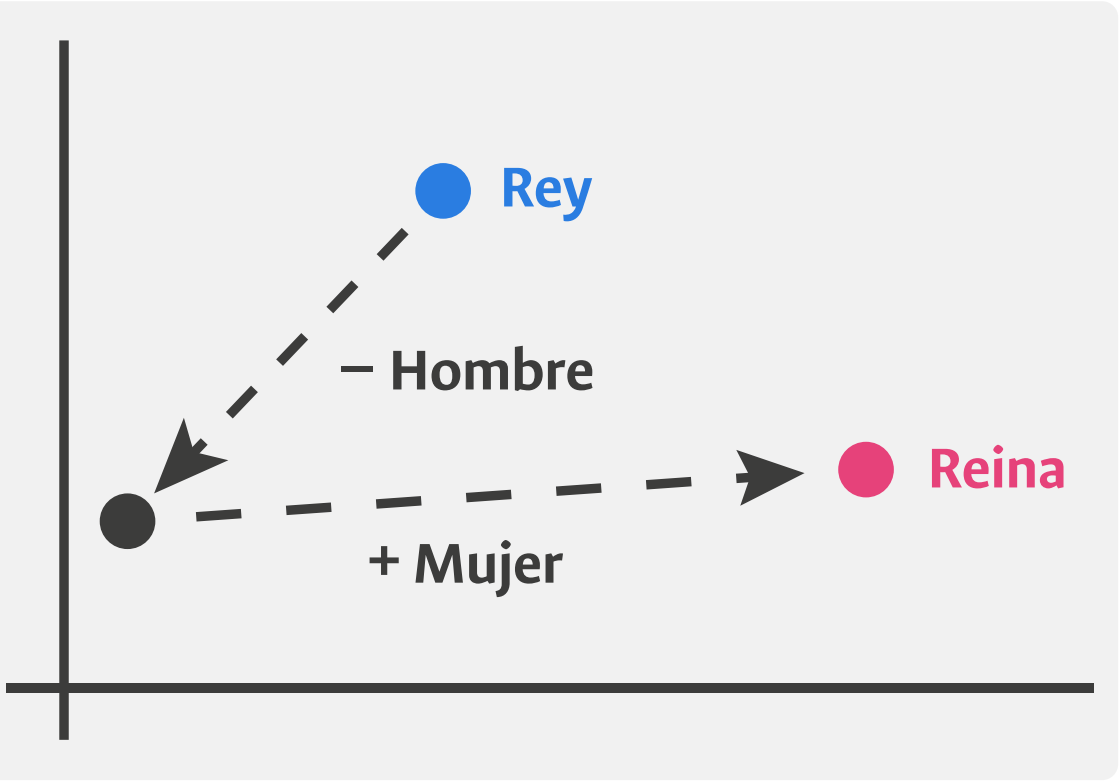
1. Definición
2. Representaciones basadas en conteos
   - 2.1 Bolsas de palabras
   - 2.2 Bolsas de N-gramas
   - 2.3 TF-IDF
3. **Modelos semánticos**
   - 3.1 Word2Vec
   - 3.2 FastText
   - 3.3 Doc2Vec
4. Medidas de similitud

Modelos Semánticos



Los **modelos semánticos** representan palabras como vectores en un espacio donde las palabras con significado similar están cerca y las palabras disímiles están lejos.

Estos aprenden las relaciones entre palabras que pueden ser codificadas por medio de una medida de similitud numérica.



Modelos Semánticos

Word2Vec

Word2Vec es un modelo basado en redes neuronales que busca codificar el contexto de una palabra.

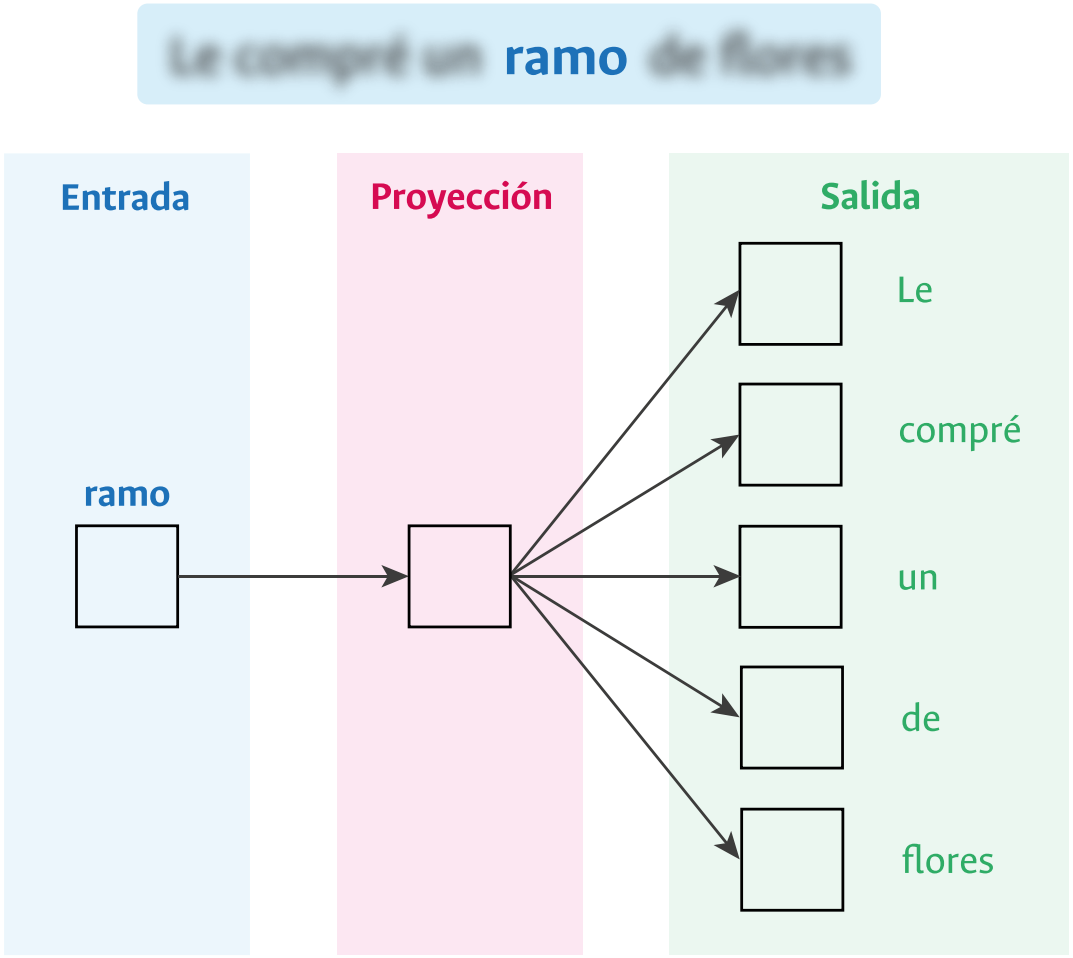
Existen dos variaciones de Word2Vec: *skip-gram* y *continuous bag-of-words* (CBOW).

Le compré un ramo de flores

- Contexto
- Palabra a representar

Modelos Semánticos – Word2Vec

Skip-gram



El modelo ***Skip-gram*** es un modelo que busca predecir el contexto de una palabra.

Toma como entrada una codificación *dummy* de una palabra y trata de predecir la secuencia de palabras que está antes y después de esta dentro del texto.

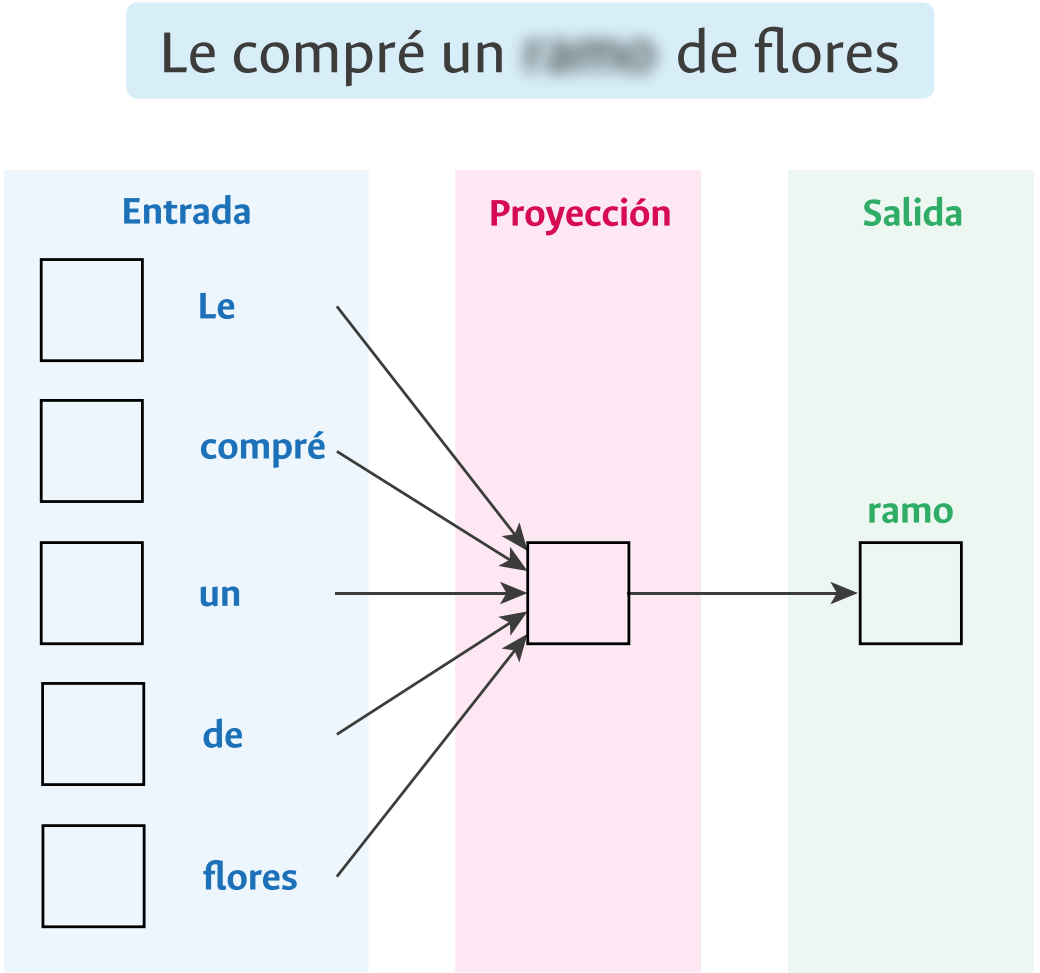
Usa una representación intermedia de red neuronal como *embedding*.

Modelos Semánticos – Word2Vec

CBOW

El modelo **CBOW** es un modelo que busca predecir una palabra a partir de su contexto.

- Tiene como entrada varias palabras del contexto y trata de predecir la palabra correspondiente.
- No es tan común para extraer *embeddings* de una palabra, ya que la entrada se conforma por varias de estas. Sin embargo, es útil para representar oraciones.

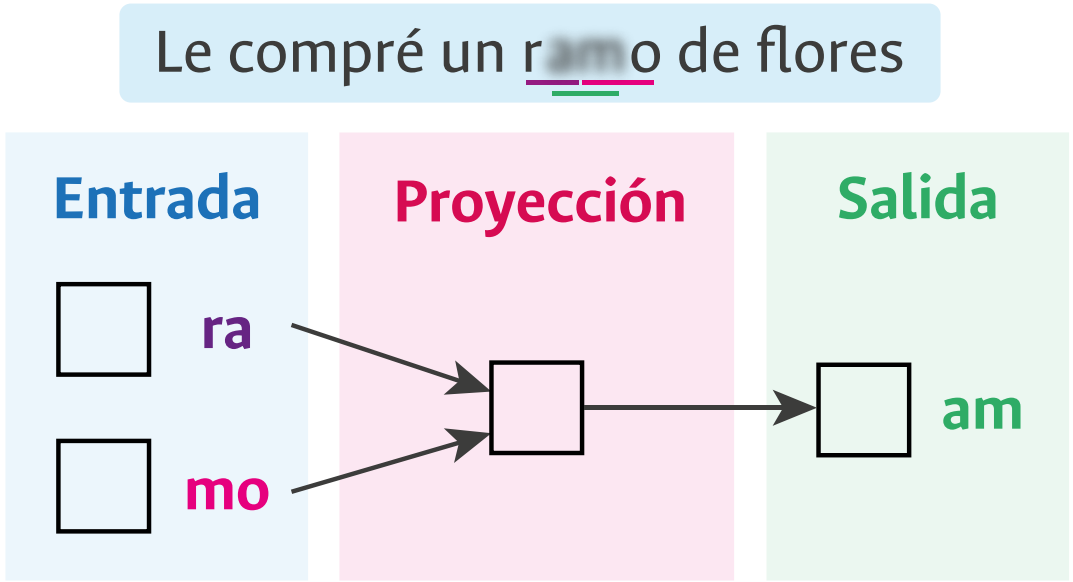


Modelos Semánticos

FastText

FastText es un modelo cuya estructura es igual que *Word2Vec*, la única diferencia es que utiliza *N-Grams* a nivel de caracter en lugar de palabras.

Este modelo fue propuesto para que funcione con **palabras que no estén en el vocabulario**, lo cual es uno de los principales problemas de *Word2Vec*.

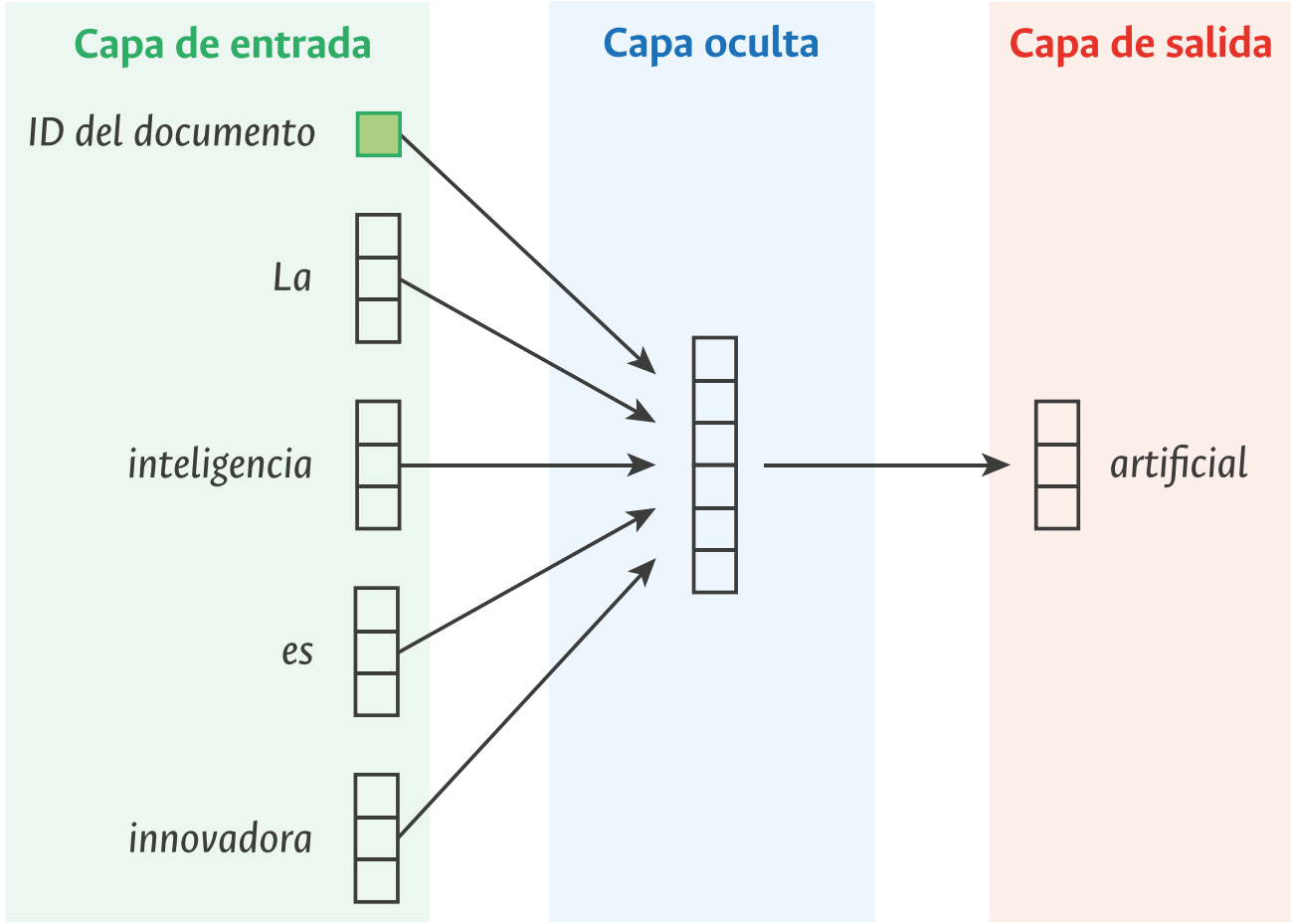


Modelos Semánticos

Doc2Vec

Doc2Vec es un modelo que es una modificación de *Word2Vec*, con la diferencia de que busca codificar **documentos completos** en lugar de palabras individuales.

Esto se consigue al modificar la estructura del modelo e incluir como entrada un identificador de documento.



Los modelos semánticos comúnmente los podemos encontrar en las siguientes tareas:

Análisis de sentimientos

Suele ser usado como mecanismo de representación para identificar si un texto se relaciona con conceptos positivos, negativos o neutrales.

Búsqueda semántica

Las representaciones de *Word2Vec* contienen mucha información semántica, lo que nos permite realizar búsquedas por conceptos y no por términos.

Llenado de máscaras

Permite interpolar palabras desconocidas que incluso pueden llegar a ser usadas como corrección de gramática.

Aplicaciones

Los modelos semánticos comúnmente los podemos encontrar en las siguientes tareas:

Clasificación de textos

Las representaciones obtenidas con los modelos semánticos en muchas oportunidades codifican información clave para discriminar entre categorías.

Agrupamiento de textos

La información semántica resulta ser muy útil para agrupar textos por contenido y similitud semántica en lugar de términos comunes.

Agenda

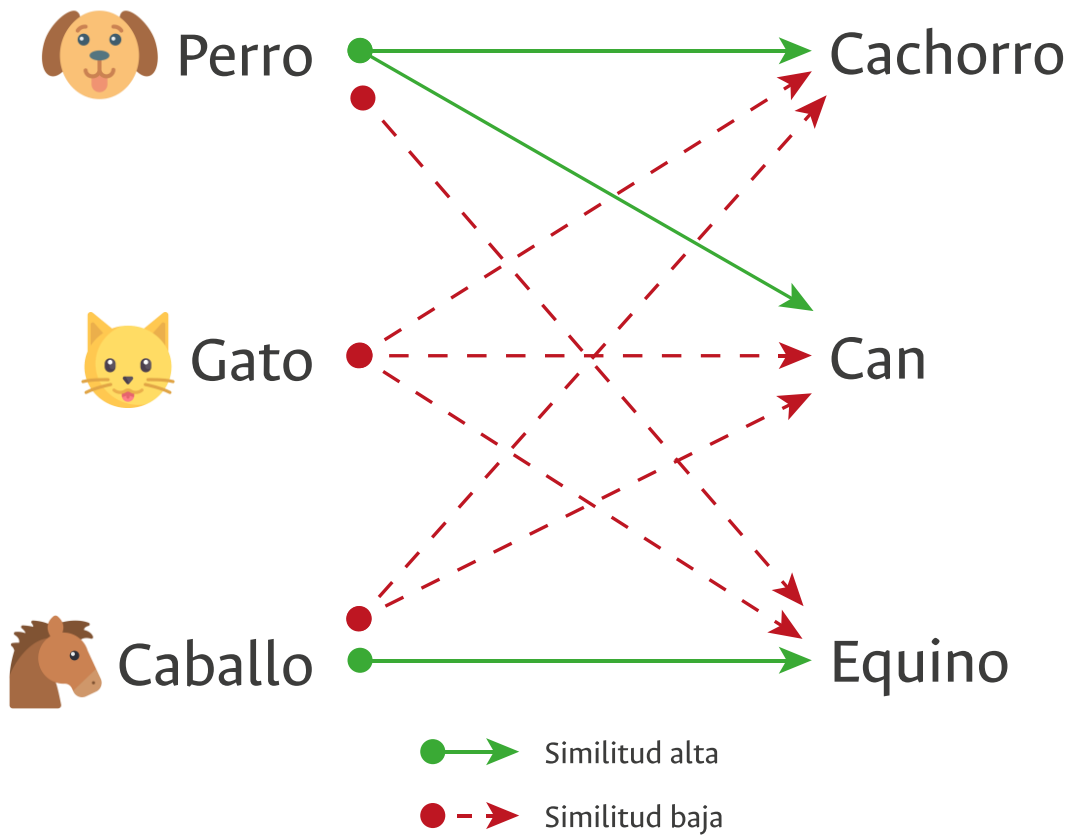
1. Definición
2. Representaciones basadas en conteos
   - 2.1 Bolsas de palabras
   - 2.2 Bolsas de N-gramas
   - 2.3 TF-IDF
3. Modelos semánticos
   - 3.1 Word2Vec
   - 3.2 FastText
   - 3.3 Doc2Vec
4. **Medidas de similitud**

Medidas de Similitud

La **similitud** entre palabras y documentos es una métrica clave en muchas tareas de NLP, como la clasificación de documentos y la búsqueda de información.

- Las medidas de similitud se utilizan para comparar y evaluar la relación entre dos elementos de texto y así mejorar la precisión y eficacia de los modelos y algoritmos NLP.

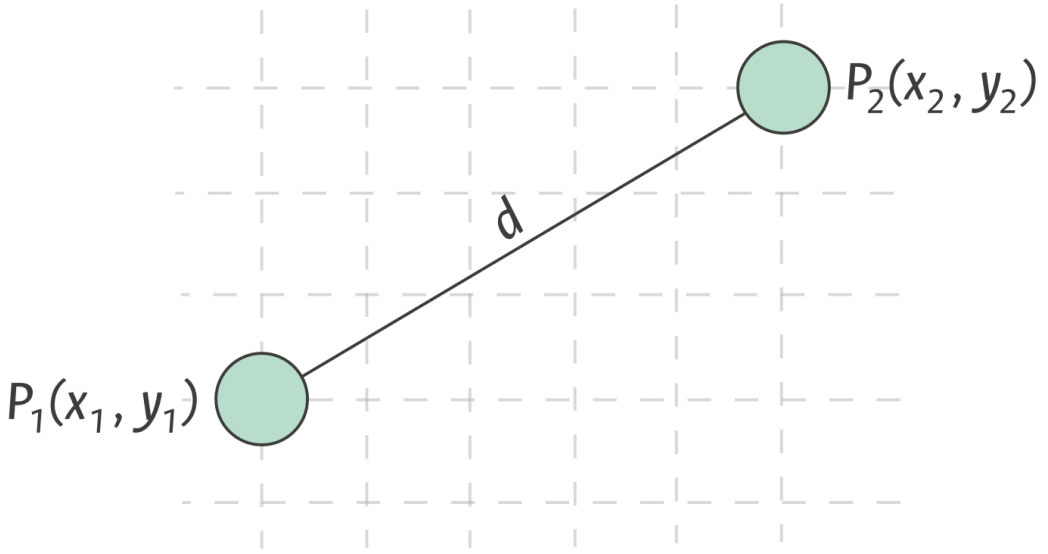
- Normalmente se utilizan la distancia Euclidiana y la similitud coseno.



Medidas de Similitud

Distancia Euclidiana

- Representa la distancia espacial entre dos *embeddings*.
- Su uso es muy común en aplicaciones con *embeddings* de redes neuronales profundas.
- No es muy recomendable con representaciones basadas en conteos, ya que es sensible a la magnitud de los vectores fruto de longitudes de documentos variables.



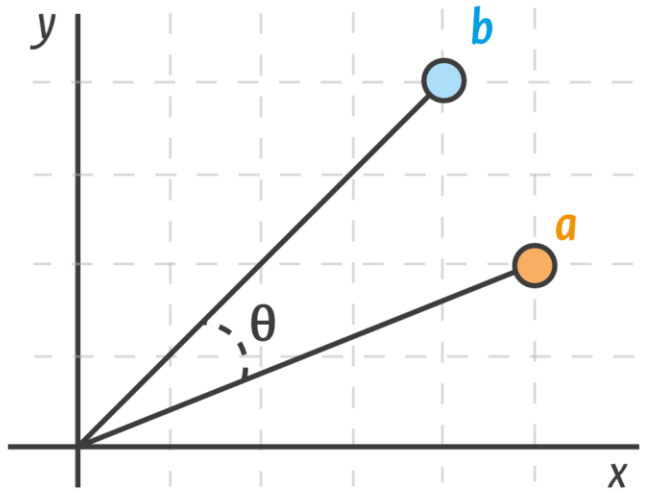
$$\text{Distancia Euclidiana } (d) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Medidas de Similitud

Distancia Coseno

Es la medida de similitud más recomendada en aplicaciones de NLP.

- Representa la alineación o ángulo entre dos *embeddings*.
- Su uso es muy común en aplicaciones con *embeddings* de conteos y *embeddings* semánticos.



$$\cos(\theta) = \frac{a \cdot b}{||a||\,||b||}$$

Referencias

- Manning, C., Raghavan, P., Schütze, H. (2009). *Introduction to Information Retrieval.* (Edición online). *Cambridge University Press.* Recuperado de https://nlp.stanford.edu/IR-book/
- Taher, M., Camacho, J. (2021). *Embeddings in Natural Language Processing: Theory and Advances in Vector Representations of Meaning.* (Primera edición). *Morgan & Claypool Publishers.* Recuperado de http://josecamachocollados.com/book_embNLP_draft.pdf
- Jurafsky, D., Martin, J. (2020). Speech and Language Processing: An Introduction to Natural Language Processing. (Tercera edición). *Stanford University.* Recuperado de https://plato.stanford.edu/entries/computational-linguistics/

- Word2Vec Model. (21 de diciembre de 2022). Gensim. https://radimrehurek.com/gensim/auto_examples/tutorials/run_word2vec.html#sphx-glr-auto-examples-tutorials-run-word2vec-py
- FastText Model. (21 de diciembre de 2022). Gensim. https://radimrehurek.com/gensim/auto_examples/tutorials/run_fasttext.html#sphx-glr-auto-examples-tutorials-run-fasttext-py
- Doc2Vec Model. (21 de diciembre de 2022). Gensim https://radimrehurek.com/gensim/auto_examples/tutorials/run_doc2vec_lee.html#sphx-glr-auto-examples-tutorials-run-doc2vec-lee-py

Derechos de imágenes

- Flaticon. (s.f.). Countdown free icon. [Icono]. https://www.flaticon.com/free-icon/countdown_2346286
- Flaticon. (s.f.). Search free icon. [Icono]. https://www.flaticon.com/free-icon/search_2810495
- Flaticon. (s.f.). Data recovery free icon. [Icono]. https://www.flaticon.com/free-icon/data-recovery_3786073
- Flaticon. (s.f.). Discussion free icon. [Icono]. https://www.flaticon.com/free-icon/discussion_2779760
- Flaticon. (s.f.). Rename free icon. [Icono]. https://www.flaticon.com/free-icon/rename_5432801
- Flaticon. (s.f.). Hierarchy free icon. [Icono]. https://www.flaticon.com/free-icon/hierarchy_6261577
- Flaticon. (s.f.). Search free icon. [Icono]. https://www.flaticon.com/free-icon/search_3093773
- Flaticon. (s.f.). Predictive models free icon. [Icono]. https://www.flaticon.com/free-icon/predictive-models_2103652
- Flaticon. (s.f.). Sound free icon. [Icono]. https://www.flaticon.com/free-icon/sound_3208765
- Flaticon. (s.f.). Photo gallery free icon. [Icono]. https://www.flaticon.com/free-icon/photo-gallery_8344913
- Flaticon. (s.f.). Text free icon. [Icono]. https://www.flaticon.com/free-icon/text_8356088
- Flaticon. (s.f.). Film free icon. [Icono]. https://www.flaticon.com/free-icon/film_1146101
- Flaticon. (s.f.). Motion sensor free icon. [Icono]. https://www.flaticon.com/free-icon/motion-sensor_1003367
- Flaticon. (s.f.). Datta collection free icon. [Icono]. https://www.flaticon.com/free-icon/data-collection_8637090
- Flaticon. (s.f.). Dog free icon. [Icono]. https://www.flaticon.com/free-icon/dog_5511726
- Flaticon. (s.f.). Horse free icon. [Icono]. https://www.flaticon.com/free-icon/horse_5511666
- Flaticon. (s.f.). Cat free icon. [Icono]. https://www.flaticon.com/free-icon/cat_616430

Créditos

Facultad de
INGENIERÍA

Profesor
Felipe Restrepo-Calle, PhD

Asistente docente
Juan Sebastián Lara Ramírez

Coordinador de virtualización
Edder Hernández Forero

Diagramadora PPT
Rosa Alejandra Superlano Esquibel

Diseño Gráfico
Clara Valeria Suárez Caballero
Milton R. Pachón Pinzón

2024