



# Redes Neuronales para Lenguaje Natural

2024

Grupo de Procesamiento de Lenguaje Natural  
Instituto de Computación



# Word Embeddings

# Representación de palabras

Queremos una representación que sea computacionalmente eficiente

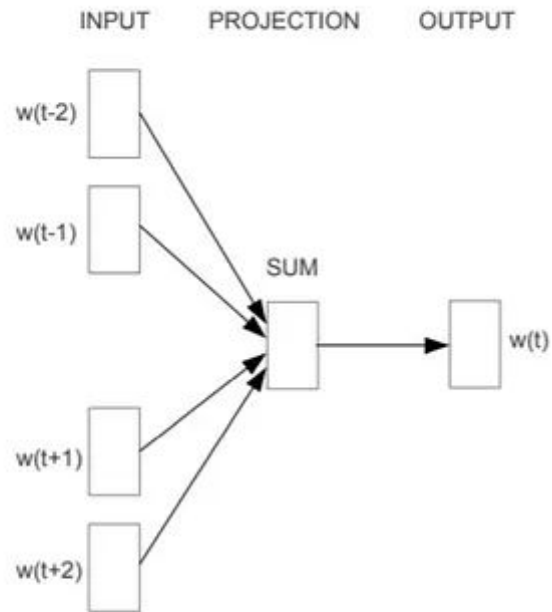
- vectores densos de baja dimensionalidad

Idealmente, palabras similares deberían estar más cerca en el espacio, y palabras diferentes deberían estar más lejos

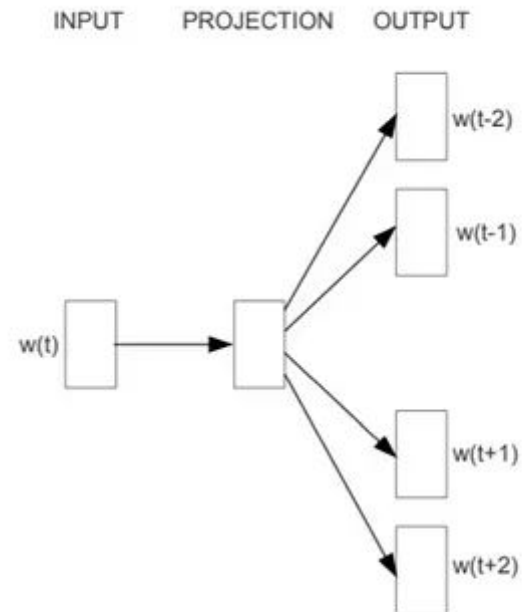
# word2vec

Dos algoritmos de creación de embeddings propuestos por Mikolov, 2013


Auto-supervisados (Representation learning)



**CBOW**



**Skip-gram**



# Embeddings - Propiedades

# Similitud de palabras

Usando similitud coseno

¿Qué está más cerca de “perro”: “gato” o “democracia”?

¿Qué está más cerca de “monarquía”: “perro” o “democracia”?

<b>cos(x,y)</b>	<b>perro</b>	<b>gato</b>	<b>democracia</b>	<b>monarquía</b>
<b>perro</b>	1.0	0.7445	0.1737	0.1647
<b>gato</b>	0.7445	1.0	0.1455	0.1235
<b>democracia</b>	0.1737	0.1455	1.0	0.5584
<b>monarquía</b>	0.1647	0.1235	0.5584	1.0

# Similitud de palabras

Usando similitud coseno

¿Qué palabras están más cerca de “perro”?

<b>perros</b>	0.7533	<b>caniche</b>	0.7329
<b>cachorro</b>	0.753	<b>teckel</b>	0.7326
<b>gato</b>	0.7445	<b>pinscher</b>	0.7305
<b>schnauzer</b>	0.737	<b>collie</b>	0.7226
<b>mastín</b>	0.7347	<b>bulldog</b>	0.7155

# Similitud de palabras

Los vecinos encontrados dependen del tamaño de ventana

**Ventanas pequeñas** ( $C = +/- 2$ ) : las palabras más cercanas son palabras sintácticamente similares en la taxonomía

- *Hogwarts* tiene como vecinos otras escuelas o instituciones ficticias
- *Sunnydale, Evernight, Blandings*

**Ventanas grandes** ( $C = +/- 5$ ) : las palabras más cercanas son palabras relacionadas en el campo semántico

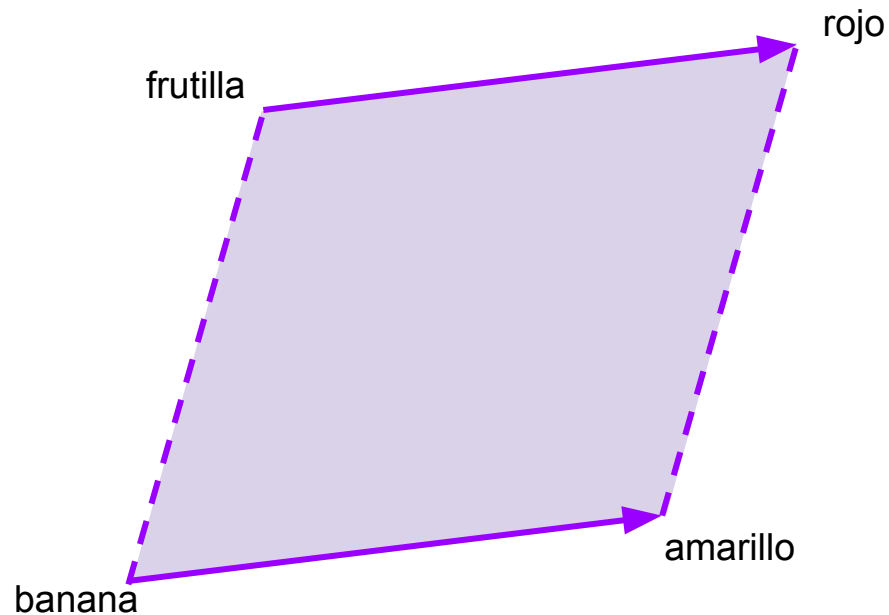
- *Hogwarts* tiene como vecinos conceptos del mundo de Harry Potter
- *Dumbledore, half-blood, Malfoy*



# Relaciones analógicas

Modelo del paralelogramo para razonamientos sobre analogías (Rumelhart and Abrahamson 1973)

Resolver: “banana es a amarillo como frutilla es a \_\_\_\_\_”



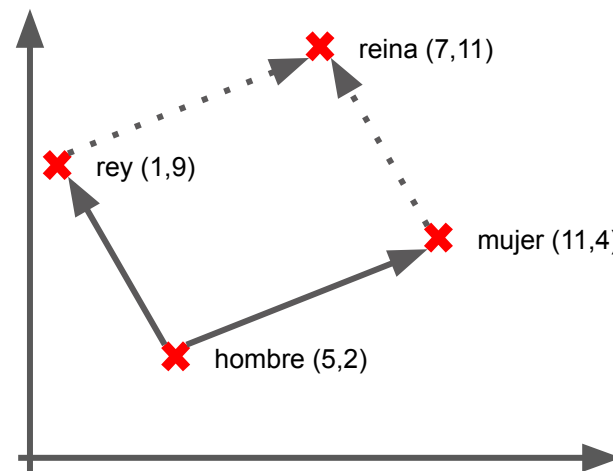
Se trasladan muy bien a operaciones con embeddings!

# Relaciones analógicas

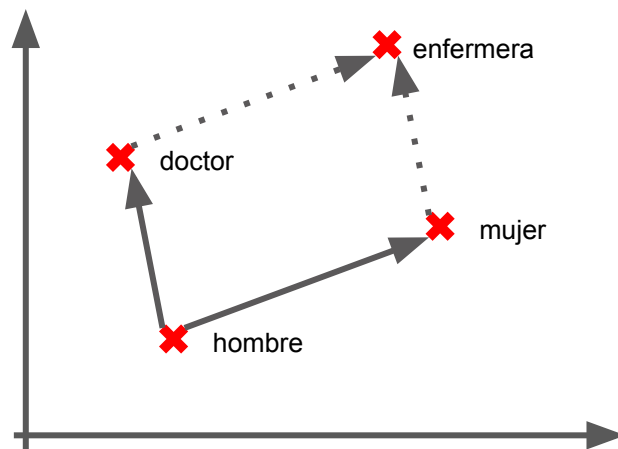
Podemos operar con aritmética de vectores para descubrir estas relaciones entre palabras:

rey - hombre + mujer  $\cong$  reina

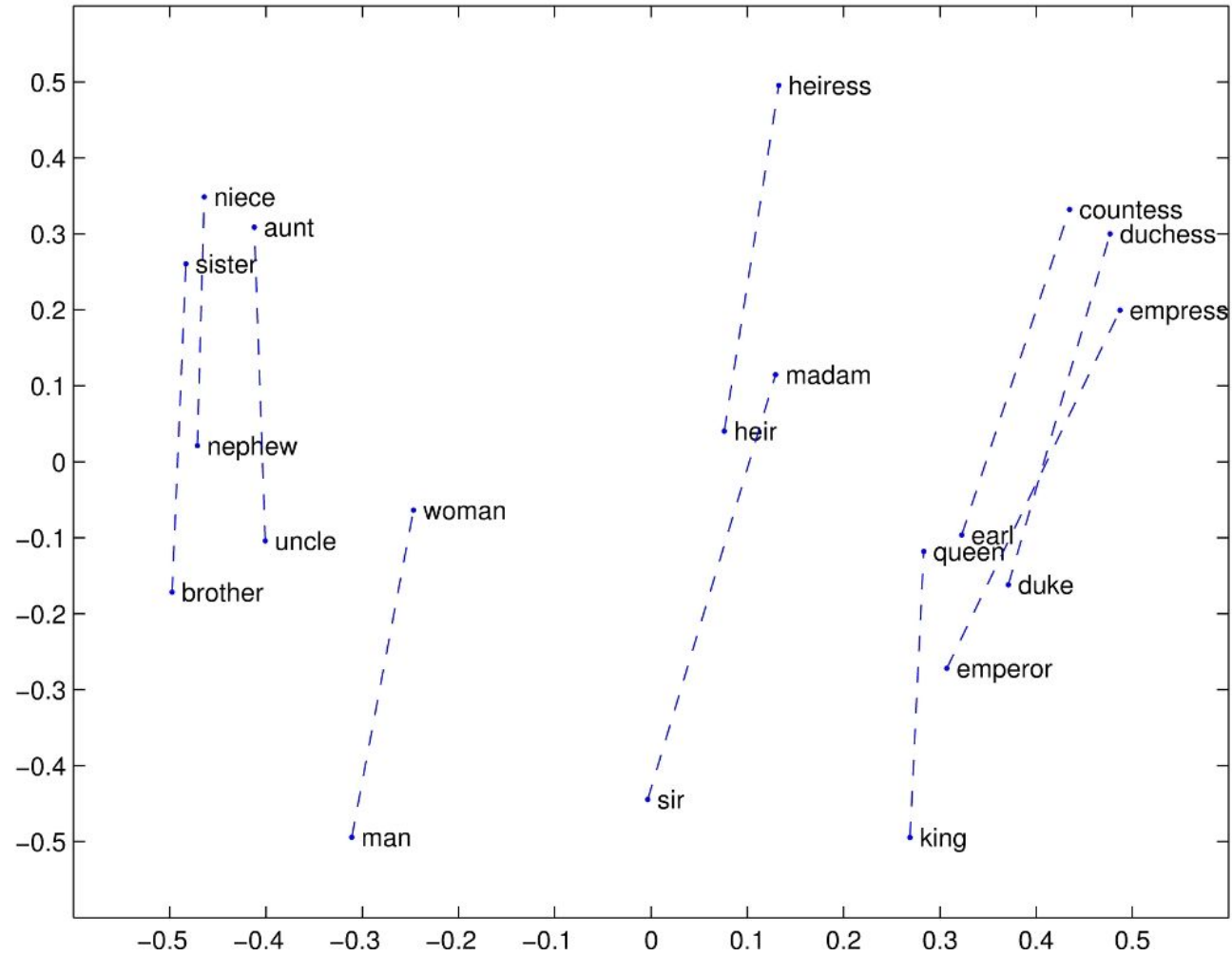
uruguay - montevideo + francia  $\cong$  parís



Pero también puede amplificar sesgos incorrectos encontrados en los datos

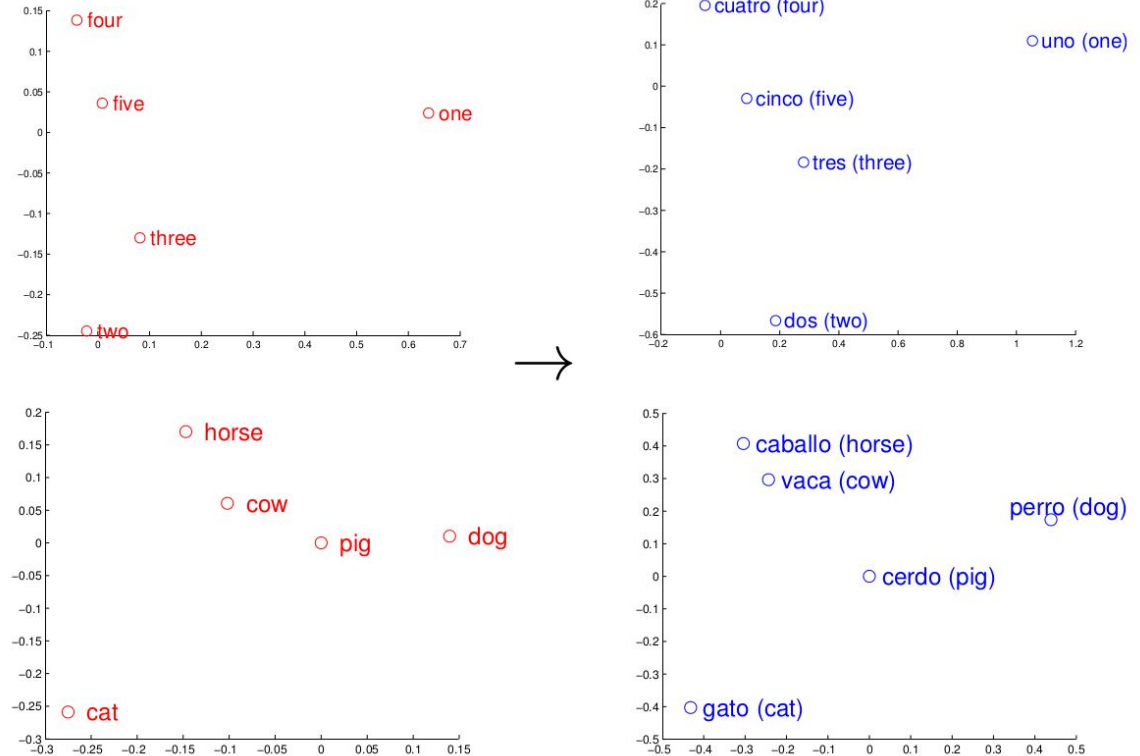


# Relaciones analógicas (GloVe)



# Transferencia entre idiomas

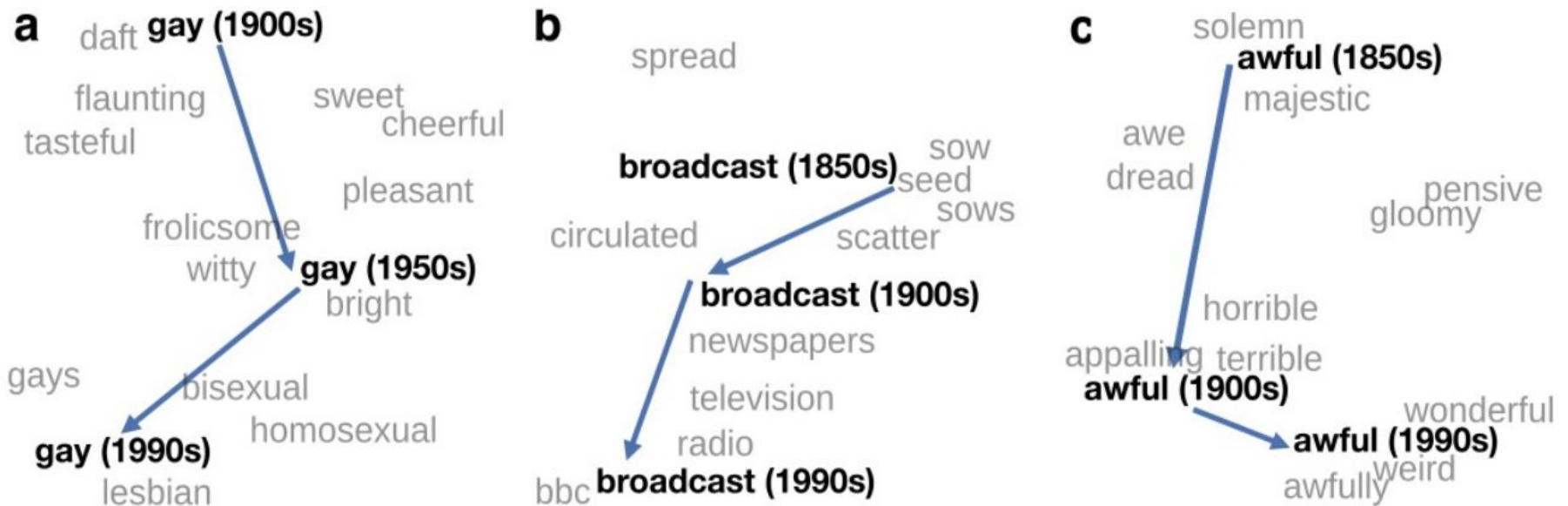
- Dadas dos colecciones de embeddings entrenadas para idiomas diferentes
- Y un conjunto de traducciones inicial
- Encontramos una transformación lineal entre los espacios de palabras en un idioma y el otro
- Las palabras se mapean a palabras similares en el espacio destino
- Puede servir para construir léxicos multilingüaje



# Análisis Histórico

Entrenar embeddings en textos de diferentes décadas para ver los cambios en su significado

~30 millones de libros, 1850-1990, datos de Google Books



# Problemas - Vectores estáticos

Representación única de palabras

A qué queda asociado el vector de “*banco*”? financiero, mueble, de arena?

*bancos, HSBC, bancaria, Citigroup*

Y el vector de “*sobre*”?

*acerca, sobe, torno, que, también*

Esto puede solucionarse usando métodos de embeddings no estáticos, como ELMO o BERT

# Problemas - Sesgos

Los embeddings reflejan y amplifican sesgos en los datos

Consulta: “Paris : France :: Tokyo : x”

x = Japan

Consulta: “father : doctor :: mother : x”

x = nurse

Consulta: “man : computer programmer :: woman : x”

x = homemaker

Algoritmos que usen embeddings por ejemplo para buscar candidatos a trabajar como programadores pueden incurrir en sesgos!


# Problemas - Sesgos

Se puede utilizar embeddings históricos para estudiar sesgos culturales

- **Sesgos de género o étnicos** para adjetivos: por ejemplo, si un adjetivo está más cerca de sinónimos de “*mujer*” que de sinónimos de “*hombre*”, o de nombres con fuerte asociación étnica
- Embeddings de adjetivos asociados a **competente** (*smart, wise, brilliant, resourceful, thoughtful, logical*) tienen sesgos hacia “*hombre*”, pero el sesgo ha ido decreciendo entre 1960-1990
- Embeddings de adjetivos **deshumanizantes** (*barbaric, monstrous, bizarre*) tenían sesgo hacia nombres asiáticos en los 1930s, y va disminuyendo durante el siglo XX







# Embeddings - Evaluación

# Evaluación Intrínseca vs Extrínseca

- Extrínseca

Entrenar una tarea con y sin embeddings (o con distintos conjuntos) y ver en qué caso funciona mejor

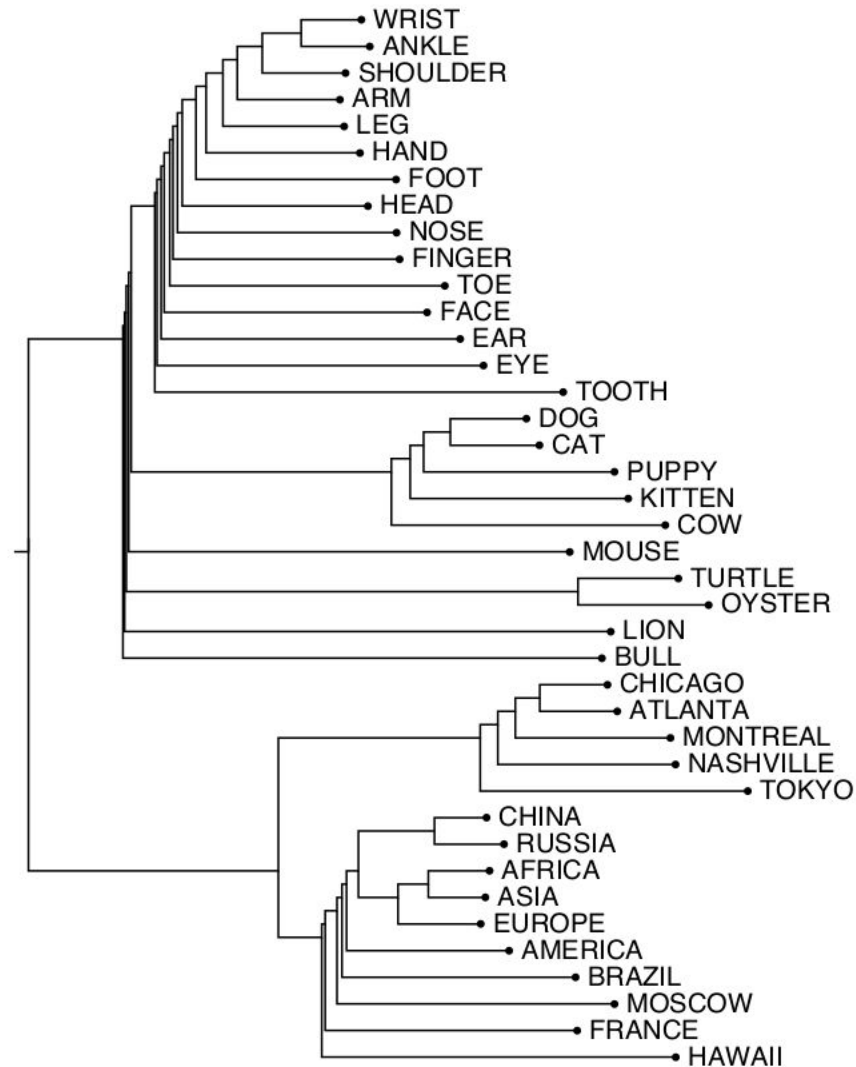
- Intrínseca

En vez de entrenar una tarea, tratar de analizar las propiedades internas del conjunto de embeddings

# Visualización

- Clustering jerárquico
- Reducción de dimensionalidad
  - PCA
  - t-SNE

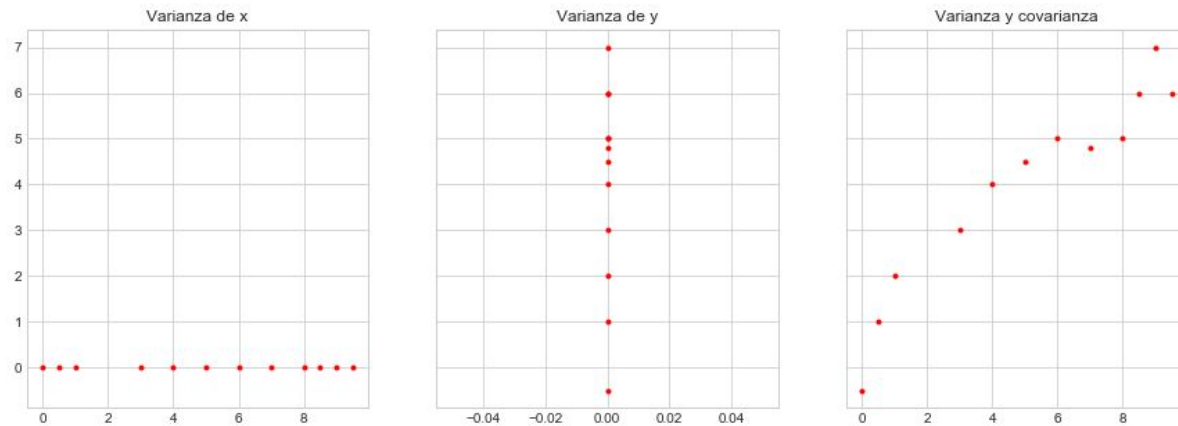
# Clustering jerárquico



# PCA

## Principal Component Analysis

- El objetivo es encontrar un conjunto de ejes sobre los cuales los datos tengan la mayor varianza
- De esta forma al proyectar los datos quedan “separados” lo más posible



- Calcula la matriz de covarianza, obtiene los valores propios y se queda con el conjunto de vectores propios donde los valores propios sean mayores

# t-SNE

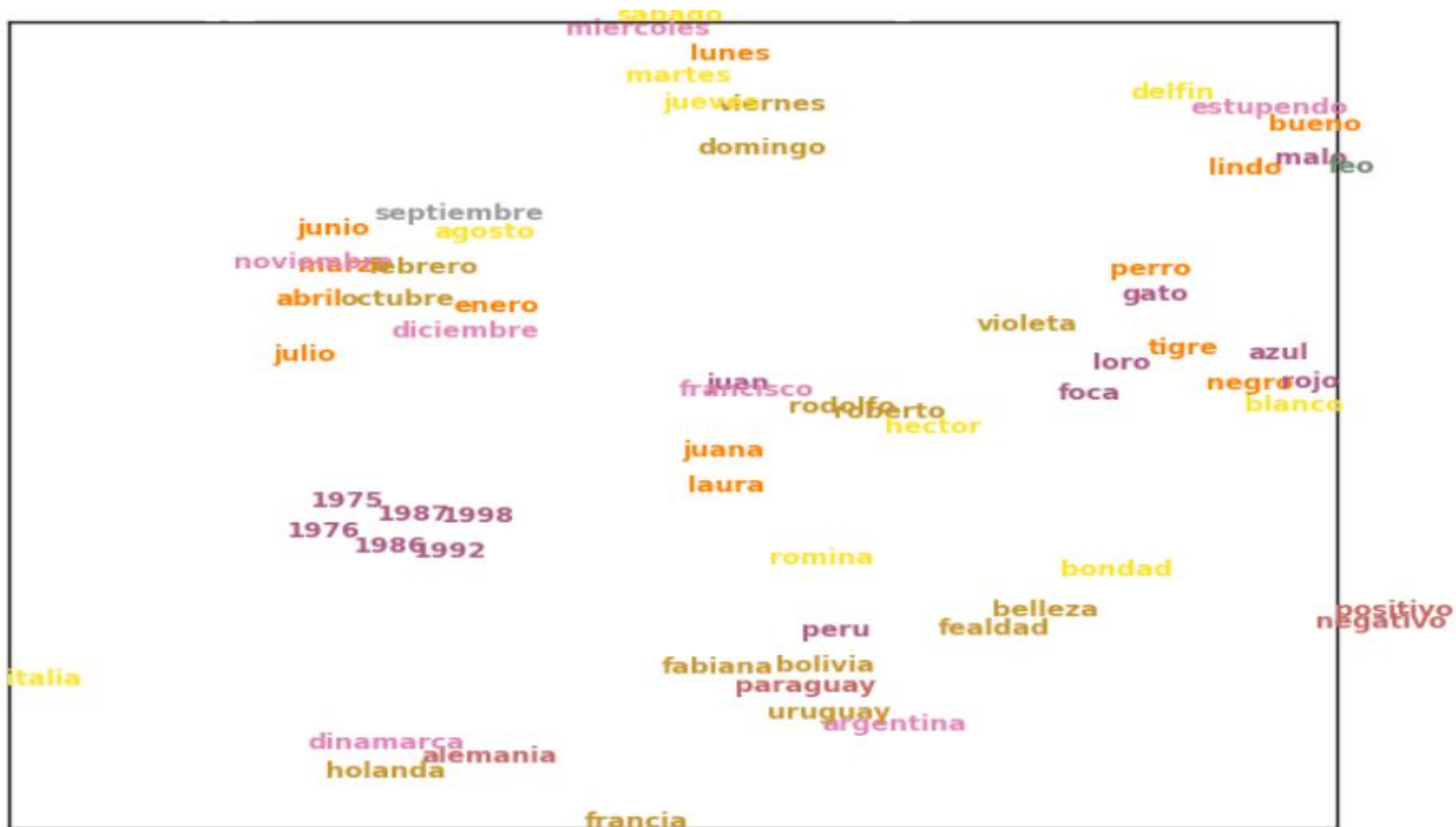
t-distributed Stochastic Neighbor Embedding

En vez de considerar distancias euclídeas, considera la probabilidad condicional de que un punto elija a otro como su vecino

Método no lineal, intenta preservar la estructura local entre puntos, además de su distancia global relativa

# t-SNE

Ejemplo para el español







# Test de analogías

- Sabiendo la propiedad del paralelogramo para analogías
- Podemos construir colecciones de cuatro-tuplas en las que, dadas tres palabras, al calcular el paralelogramo debería encontrarse la cuarta palabra
- Calculamos accuracy con top-n
- Normalmente se elimina la palabra misma y otras flexiones comunes

# Test de analogías

Colección de analogías de Pennington, (2014), Mikolov (2013)

- Sintácticas:
  - infinitivo - gerundio: *comer : comiendo :: saltar : saltando*
  - adj - adv: *rápido : rápidamente :: sabio : sabiamente*
- Semánticas:
  - países - capitales: *Francia : París :: Uruguay : Montevideo*
  - relaciones familiares: *madre : padre :: tía : tío*

# Test de analogías

Inglés

Model	Dim.	Size	Sem.	Syn.	Tot.
ivLBL	100	1.5B	55.9	50.1	53.2
HPCA	100	1.6B	4.2	16.4	10.8
GloVe	100	1.6B	<u>67.5</u>	<u>54.3</u>	<u>60.3</u>
SG	300	1B	61	61	61
CBOW	300	1.6B	16.1	52.6	36.1
vLBL	300	1.5B	54.2	<u>64.8</u>	60.0
ivLBL	300	1.5B	65.2	63.0	64.0
GloVe	300	1.6B	<u>80.8</u>	61.5	<u>70.3</u>
SVD	300	6B	6.3	8.1	7.3
SVD-S	300	6B	36.7	46.6	42.1
SVD-L	300	6B	56.6	63.0	60.1
CBOW <sup>†</sup>	300	6B	63.6	<u>67.4</u>	65.7
SG <sup>†</sup>	300	6B	73.0	66.0	69.1
GloVe	300	6B	<u>77.4</u>	67.0	<u>71.7</u>
CBOW	1000	6B	57.3	68.9	63.7
SG	1000	6B	66.1	65.1	65.6
SVD-L	300	42B	38.4	58.2	49.2
GloVe	300	42B	<b><u>81.9</u></b>	<b><u>69.3</u></b>	<b><u>75.0</u></b>

GloVe: Global Vectors for Word Representation  
Pennington, Socher, Manning (2014)

Español

Dataset	Dimension				
	25	50	100	150	200
<b>semantic</b>					
capital-comm	40.4	65.1	72.5	74.4	75.4
capital-world	21.3	40.3	51.3	53.2	51.8
city-in-state	25.6	42.8	52.6	57.1	59.0
currency	0.3	0.7	0.7	0.7	0.6
family	62.6	78.0	79.6	81.8	80.1
<b>syntactic</b>	<b>25</b>	<b>50</b>	<b>100</b>	<b>150</b>	<b>200</b>
adj-to-adv	4.5	6.0	8.9	9.7	8.3
opposite	4.0	7.6	8.5	10.1	11.7
comparative	-	-	-	-	-
superlative	-	-	-	-	-
present-part	21.9	29.0	37.1	35.7	32.9
nation-adj	44.0	68.3	81.8	86.0	86.6
past-tense	12.3	21.4	26.9	27.5	27.7
plural	13.5	22.7	30.9	33.0	36.5
plural-verbs	26.9	39.8	47.5	45.7	43.1

Spanish word vectors from Wikipedia  
Etcheverry, Wonsever (2016)

# Test de analogías

Guaraní

Wiki	Chiruzzo et al. 2020	Parallel News Set	Reliable Tweets	Unreliable Tweets	family		ccc	
					Exact	Top 5	Exact	Top 5
X					29.97%	38.89%	4.41%	10.01%
X	X				<b>41.27%</b>	<b>48.41%</b>	5.27%	11.53%
X	X	X			32.54%	34.92%	<b>5.53%</b>	<b>13.37%</b>
X	X	X	X		28.57%	36.51%	5.27%	13.04%
X	X	X	X	X	26.98%	35.71%	4.55%	12.25%

Experiments on a Guaraní Corpus of News and Social Media  
Góngora, Giossa, Chiruzzo (2021)

# Test de similitud

- Sabiendo que palabras similares están más cerca, y palabras distintas están más lejos
- Teniendo una colección de palabras cuya similitud está evaluada por humanos
- Ordenamos las palabras según nuestra colección y evaluamos qué tan parecido es el orden a lo indicado por los humanos

# Test de similitud

- Colecciones de tests
  - WordSim-353: 353 pares de sustantivos rankeados por humanos entre 0 y 10
  - SimLex-999: Similar al anterior, pero incluye también adjetivos y verbos, y palabras abstractas y concretas
  - Otros...
- Cómo evaluamos si el orden asignado por humanos es similar al de los embeddings?
  - Coeficiente de correlación de Spearman

# Test de similitud

word1	word2	similarity
vanish	disappear	9.8
behave	obey	7.3
belief	impression	5.95
muscle	bone	3.65
modest	flexible	0.98
hole	agreement	0.3

SimLex-999 dataset (Hill et al., 2015)



# Test de similitud

Inglés

Model	Size	WS353	MC	RG	SCWS	RW
SVD	6B	35.3	35.1	42.5	38.3	25.6
SVD-S	6B	56.5	71.5	71.0	53.6	34.7
SVD-L	6B	65.7	<u>72.7</u>	75.1	56.5	37.0
CBOW <sup>†</sup>	6B	57.2	65.6	68.2	57.0	32.5
SG <sup>†</sup>	6B	62.8	65.2	69.7	<u>58.1</u>	37.2
GloVe	6B	<u>65.8</u>	<u>72.7</u>	<u>77.8</u>	53.9	<u>38.1</u>
SVD-L	42B	74.0	76.4	74.1	58.3	39.9
GloVe	42B	<b><u>75.9</u></b>	<b><u>83.6</u></b>	<b><u>82.9</u></b>	<b><u>59.6</u></b>	<b><u>47.8</u></b>
CBOW*	100B	68.4	79.6	75.4	59.4	45.5

Español

Dim	WS353	MC30	SL999a	SL999
25	19.9	64.6	14.7	11.7
50	26.7	67.6	18.8	16.0
100	28.8	67.0	23.7	19.3
150	30.5	65.5	25.5	20.0
200	30.5	64.2	26.0	20.7
250	30.5	61.6	27.2	21.3

Spanish word vectors from Wikipedia  
Etcheverry, Wonsever (2016)

GloVe: Global Vectors for Word Representation  
Pennington, Socher, Manning (2014)



# Embeddings - Usos

# Input para redes neuronales

Los embeddings se pueden usar como vectores de entrada

Arquitecturas secuenciales: palabra a palabra

- CNN, LSTM, Transformers próxima unidad!

Arquitecturas fijas: cómo pasamos de una palabra a un texto?

- Centroide próxima clase!

Otros métodos de aprendizaje automático

# Aprendizaje profundo

Redes de “varias capas”

En redes para lenguaje natural, el término viene asociado al uso de una representación entrenada a partir de mucho texto

- Autosupervisión

Los word embeddings preentrenados fueron el primer método para hacer esto de forma efectiva

# MLP con Centroide de Embeddings

