Redes Neuronales para Lenguaje Natural

2023

Grupo de Procesamiento de Lenguaje Natural
Instituto de Computación


Word Embeddings

Representación de palabras

Queremos una representación que sea computacionalmente eficiente

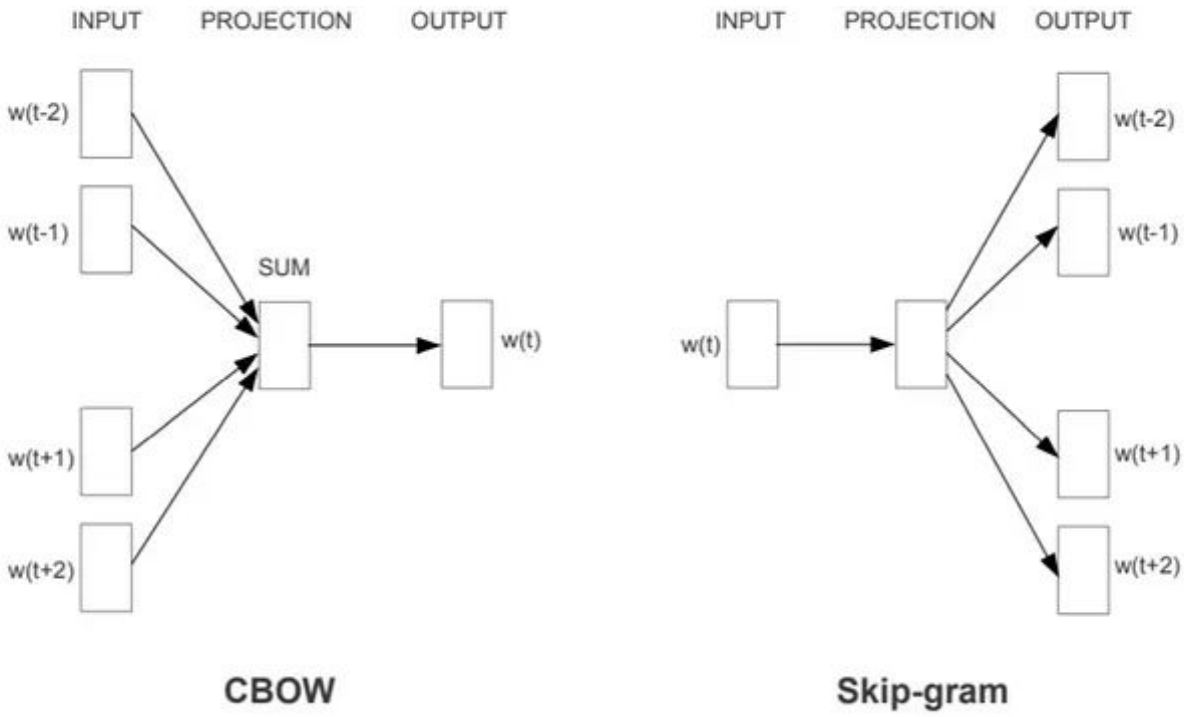
- vectores densos de baja dimensionalidad

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)



GloVe (Pennington et al., 2014)

- Global Vectors es un método para construir representaciones de las palabras formulando un problema de mínimos cuadrados a partir de las coocurrencias de palabras en un corpus.
- El método considera una ventana de tamaño fijo de palabras para construir una matriz de coocurrencias (X) usada para estimar las probabilidades condicionales de las palabras.

$$P_{ij} = P(j|i) = \frac{X_{ij}}{\sum_k X_{ik}}$$

- Se vincula el cociente de probabilidades condicionales (P(k|i)/P(k|j)) con la diferencia de representaciones.
- Se resuelve por mínimos cuadrados las matrices de embeddings W y W̃, tales que:

$$w_i^T \tilde{w}_k = \log P_{ik}$$

FastText (Bojanowski et al., 2017)

- Similar a word2vec, pero considera información sub-palabra
- Divide una palabra en N-gramas de caracteres
- Por ejemplo con N=3, “where” se separa en los subtokens

  <wh, whe, her, ere, re>

- Luego se suman los embeddings de cada subtoken
- Puede obtener representaciones para palabras no vistas en el corpus de training

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”*?

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

Similitud de palabras

Usando similitud coseno

¿Qué palabras están más cerca de *“perro”*?

perros	0.7533	**caniche**	0.7329
cachorro	0.753	**teckel**	0.7326
gato	0.7445	**pinscher**	0.7305
schnauzer	0.737	**collie**	0.7226
mastín	0.7347	**bulldog**	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*

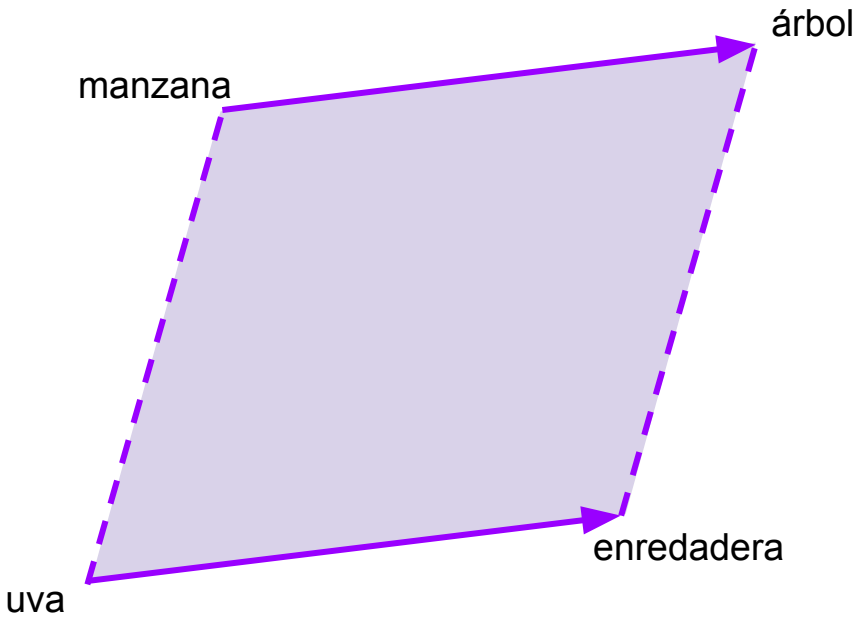
Ventanas grandes (C= +/- 5) : las palabras más cercanas son palabras relacionadas en el campo semántico

- *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: “manzana es a árbol como uva es a _____”



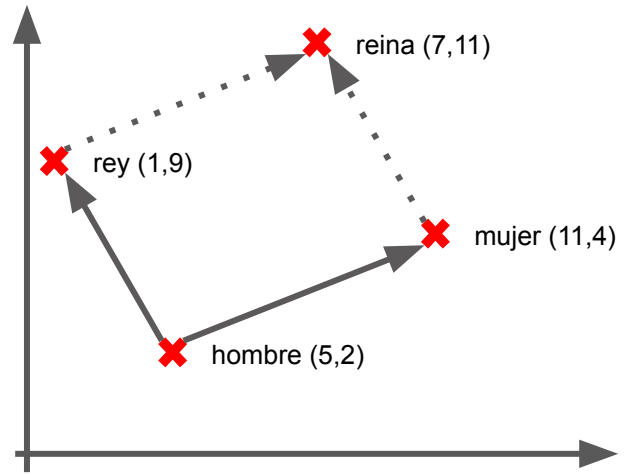
Se trasladan muy bien a operaciones con embeddings!

Relaciones analógicas

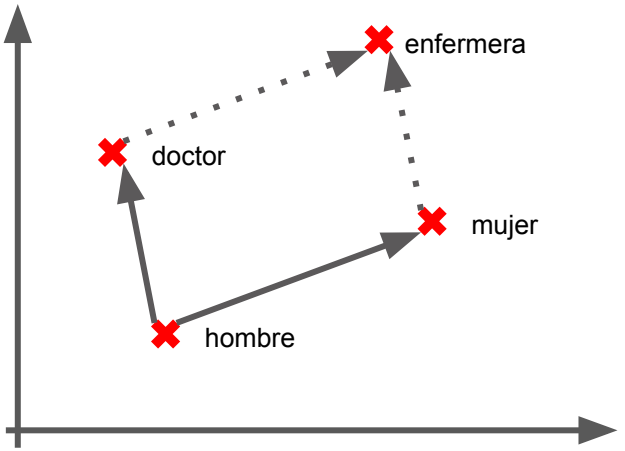
Podemos operar con aritmética de vectores para descubrir estas relaciones entre palabras:

rey - hombre + mujer ≅ reina

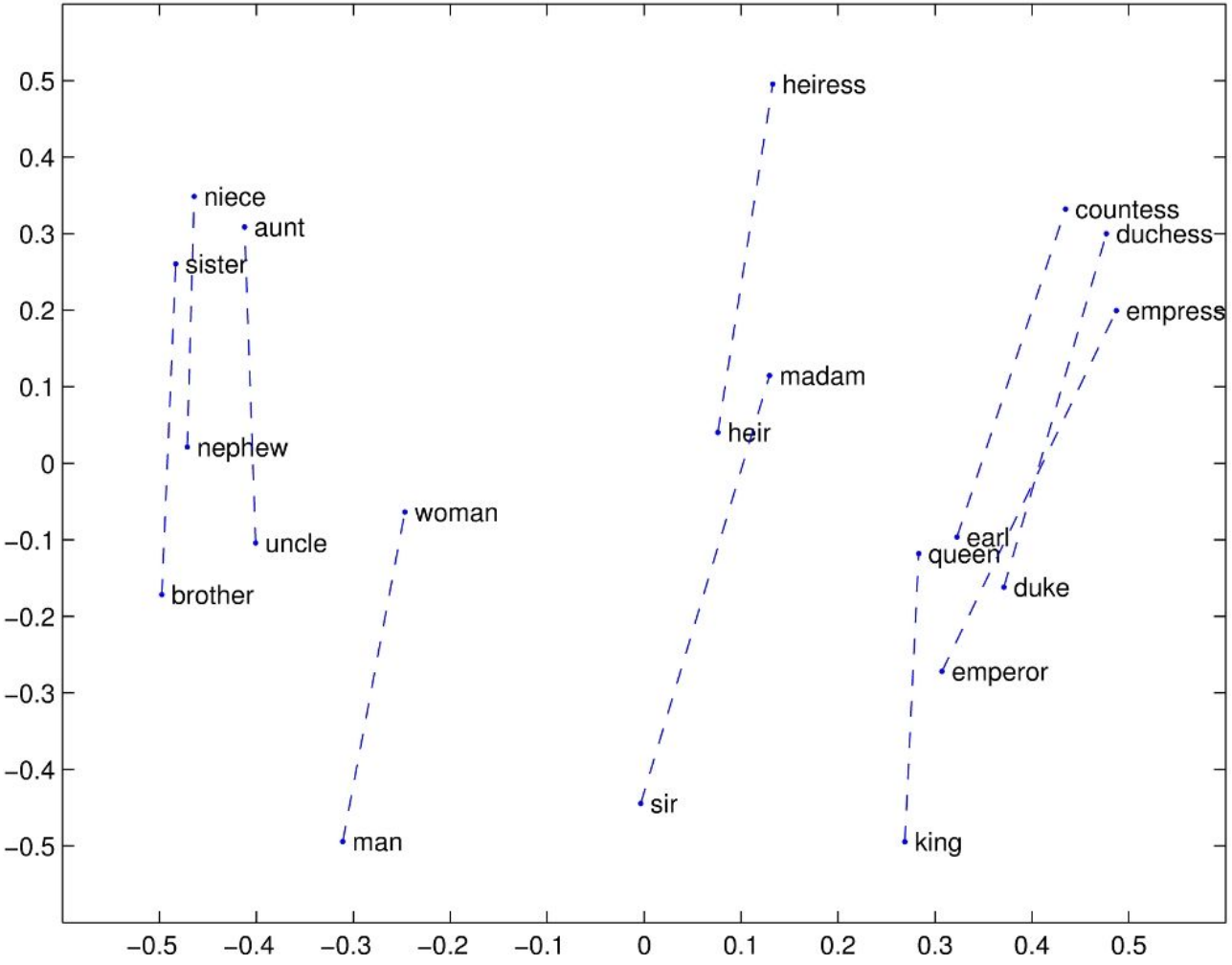
uruguay - montevideo + francia ≅ París



Pero también puede amplificar sesgos incorrectos encontrados en los datos

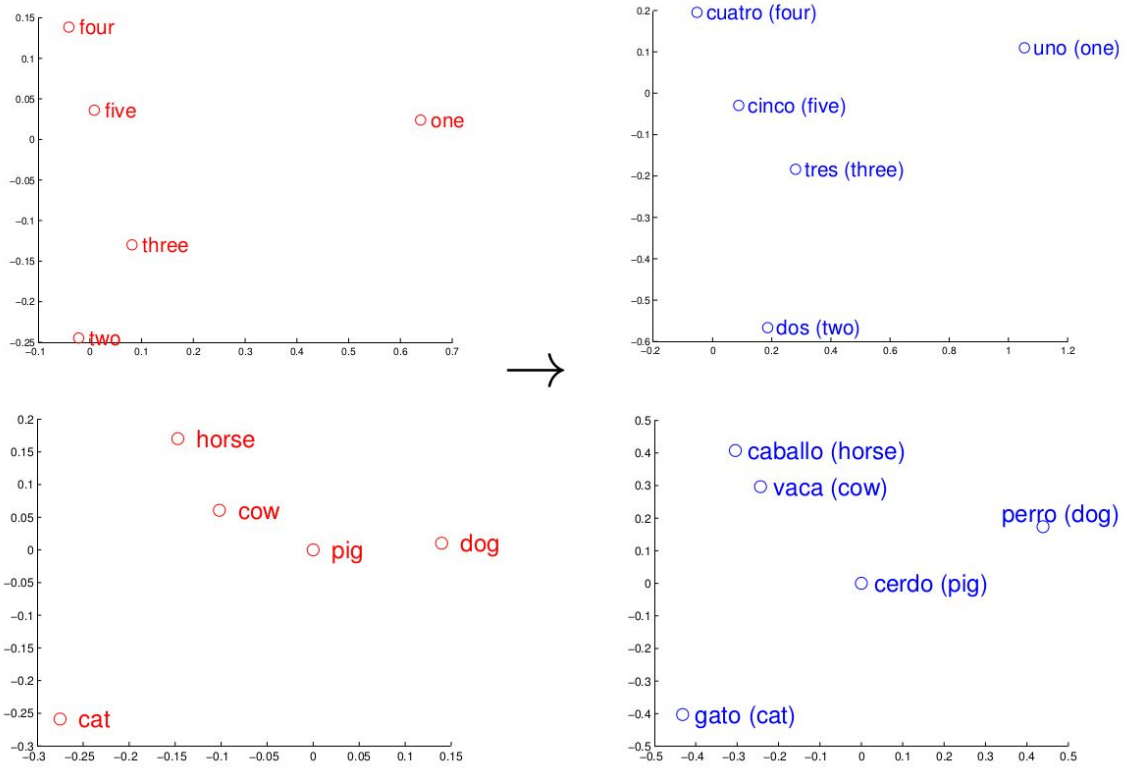


Relaciones analógicas (GloVe)



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 multilenguaje



Exploiting Similarities among Languages for Machine Translation
Mikolov, Le, Sutskever, 2013

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!

Man is to computer programmer as woman is to homemaker? debiasing word embeddings
Bolukbasi et al. (2016)

Problemas - Sesgos

Se puede utilizar embeddings históricos para estudiar sesgos culturales

- Calcular **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 a 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

Word embeddings quantify 100 years of gender and ethnic stereotypes
Garg et al. (2018)

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

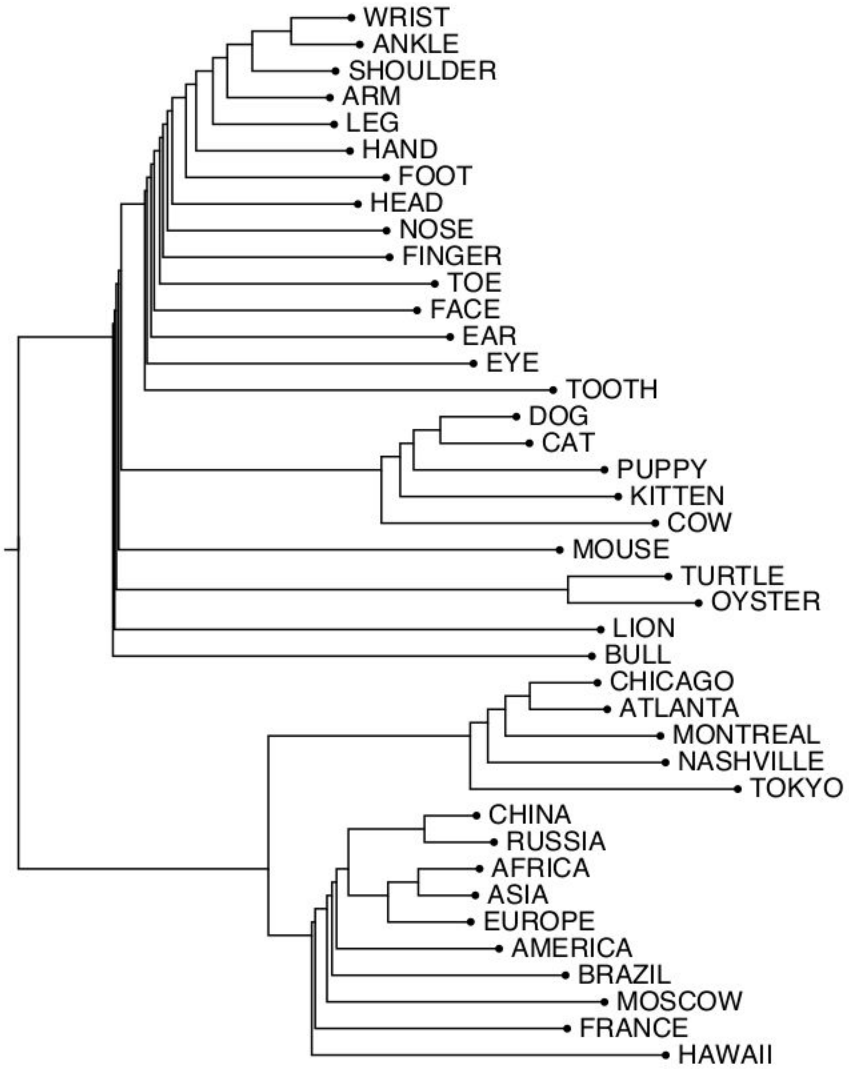
- Intrínseca

  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

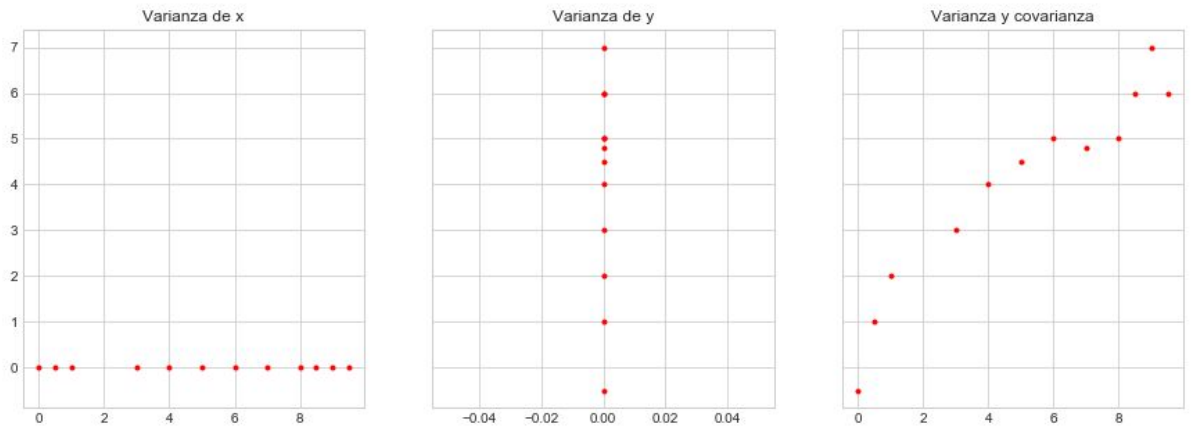
Clustering jerárquico



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

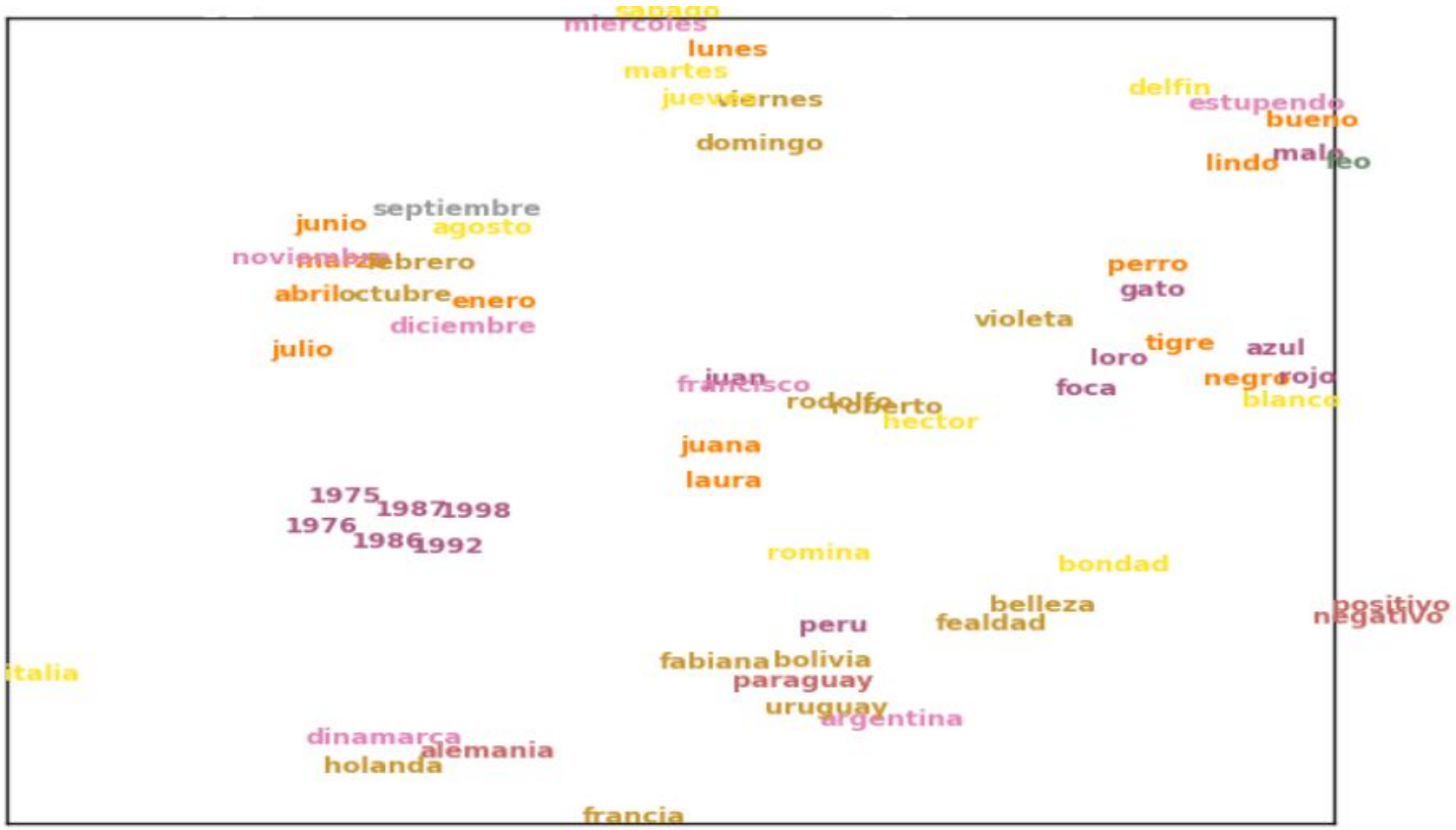
t-distributed Stochastic Neighbor Embedding

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

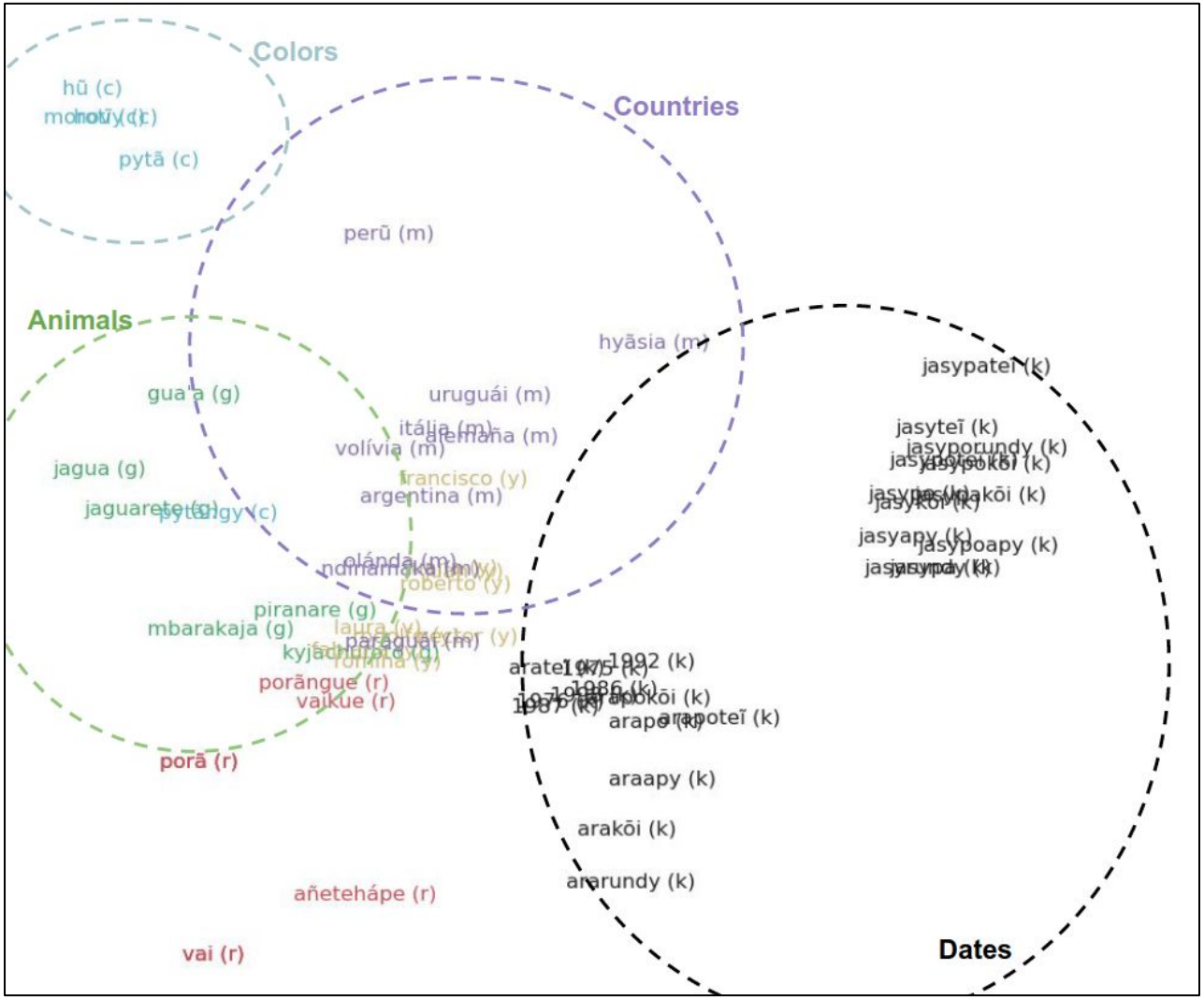
Ejemplo para el español



Spanish word vectors from Wikipedia
Etcheverry, Wonsever (2016)

PCA

Ejemplo para el guaraní



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

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	67.5	54.3	60.3
SG	300	1B	61	61	61
CBOW	300	1.6B	16.1	52.6	36.1
vLBL	300	1.5B	54.2	64.8	60.0
ivLBL	300	1.5B	65.2	63.0	64.0
GloVe	300	1.6B	80.8	61.5	70.3
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†	300	6B	63.6	67.4	65.7
SG†	300	6B	73.0	66.0	69.1
GloVe	300	6B	77.4	67.0	71.7
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	**81.9**	**69.3**	**75.0**

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

Español

Dataset	**Dimension**				
semantic	**25**	**50**	**100**	**150**	**200**
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
syntactic	**25**	**50**	**100**	**150**	**200**
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	familiy		ccc	
					Exact	Top 5	Exact	Top 5
X					29.97%	38.89%	4.41%	10.01%
X	X				**41.27%**	**48.41%**	5.27%	11.53%
X	X	X			32.54%	34.92%	**5.53%**	**13.37%**
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 Guarani 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	72.7	75.1	56.5	37.0
CBOW†	6B	57.2	65.6	68.2	57.0	32.5
SG†	6B	62.8	65.2	69.7	58.1	37.2
GloVe	6B	65.8	72.7	77.8	53.9	38.1
SVD-L	42B	74.0	76.4	74.1	58.3	39.9
GloVe	42B	**75.9**	**83.6**	**82.9**	**59.6**	**47.8**
CBOW*	100B	68.4	79.6	75.4	59.4	45.5

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

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)


Embeddings - Usos

Input para redes neuronales

Los embeddings se pueden usar como vectores de activación

Arquitecturas secuenciales: palabra a palabra

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

Otros métodos de aprendizaje automático

Aprendizaje profundo

Con word embeddings podemos…

- Usar arquitecturas de redes neuronales más complejas
  - CNN

    próxima unidad!
  - LSTM
  - Transformers
- Usarlos como features en métodos de aprendizaje clásicos
  - Centroide

    próxima clase!


Word Embeddings