Representaciones vectoriales de palabras basadas en grafos de asociación libre

Procesamiento de Lenguaje Natural

CICADA
Centro Interdisciplinario en Ciencia de Datos y Aprendizaje Automático

¿Qué son los *word embeddings*?

La hipótesis distribucional (Harris, 1954) supone que las palabras que ocurren en contextos similares tienen semántica similares.

Los Word Embeddings son representaciones vectoriales de palabras construidos a partir de su distribución en los textos, y de los contextos en los que aparecen (Jurafsky, 2021)

Este vector guarda información semántica, lo que permite que pueda ser asociado o disociado a otros vectores según distintos contextos gramaticales.

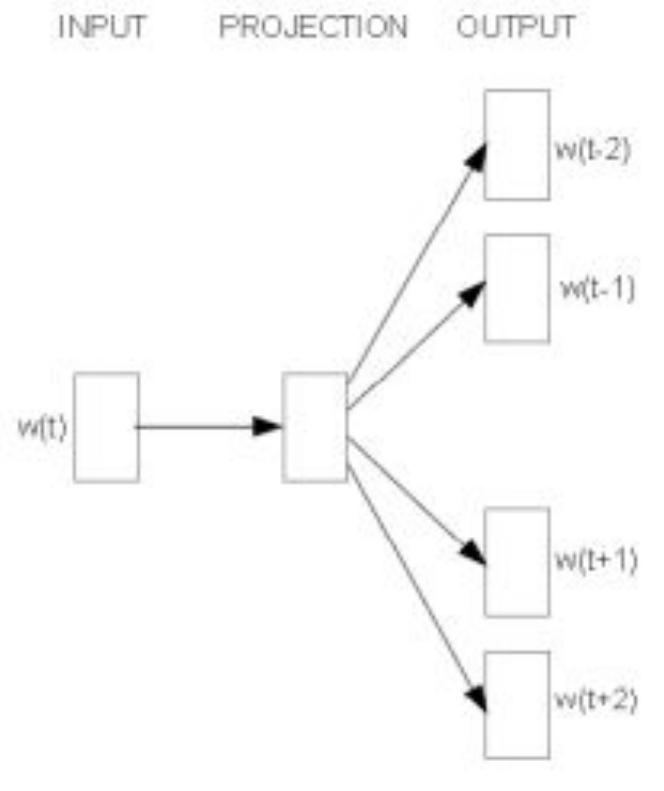
Cómo entrenar algoritmos de word embedding

Los algoritmos de word embedding suelen entrenarse en base a grandes colecciones de texto (páginas web, subtítulos, libros).

Proyecto Lexicón: Datos de asociación libre de palabras (pares estímulo--respuesta) para español rioplatense. Permiten obtener word embeddings a partir de la factorización de la matriz de adyacencia del grafo obtenido.

¿Será que pueden entrenarse algoritmos típicos de word embeddings a partir de datos de asociación libre?

Modelo SkipGram

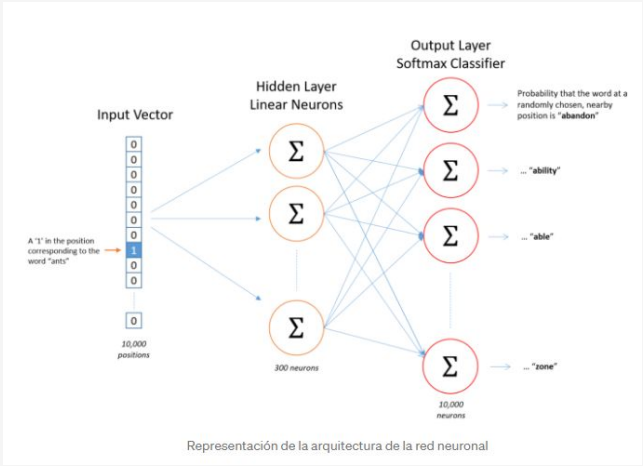


Dado un conjunto de frases (también llamado **corpus**) el modelo analiza las palabras de cada oración y trata de usar cada palabra para predecir qué palabras serán vecinas.

Se toma una palabra y se intenta calcular las mejores probabilidades de las palabras vecinas, por ejemplo: a la palabra “Caperucita” le seguirá “Roja” con más probabilidad que cualquier otra palabra.

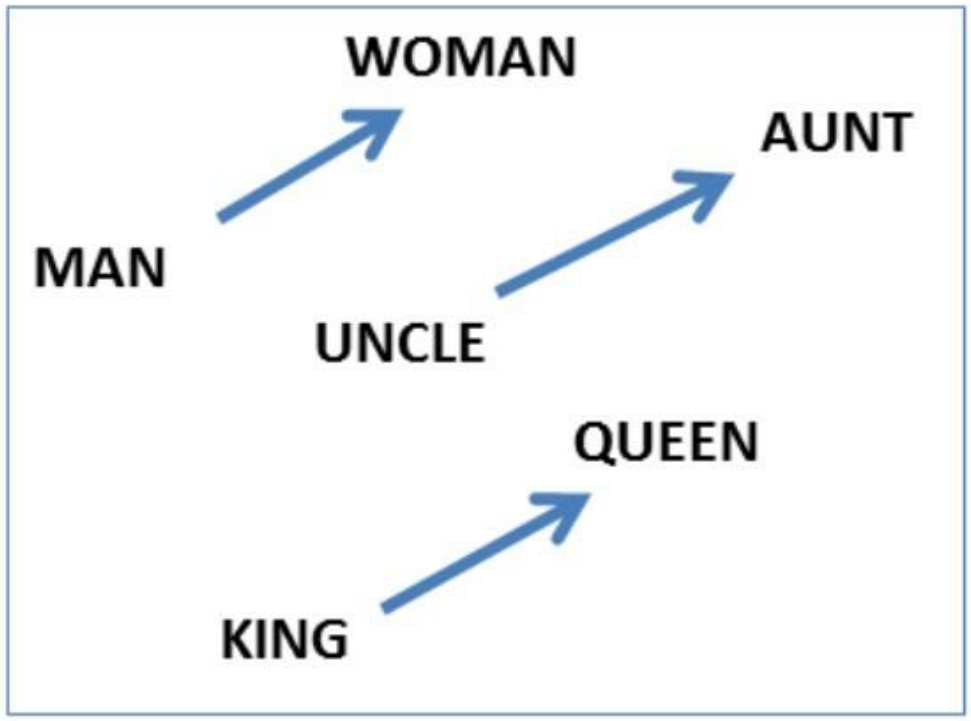
(...) el Lobo persiguió a caperucita roja por el bosque camino a la casa (...)

Se calculan los pesos de la red, con los cuales se transforma el vector de la palabra objetivo en un vector de probabilidades.



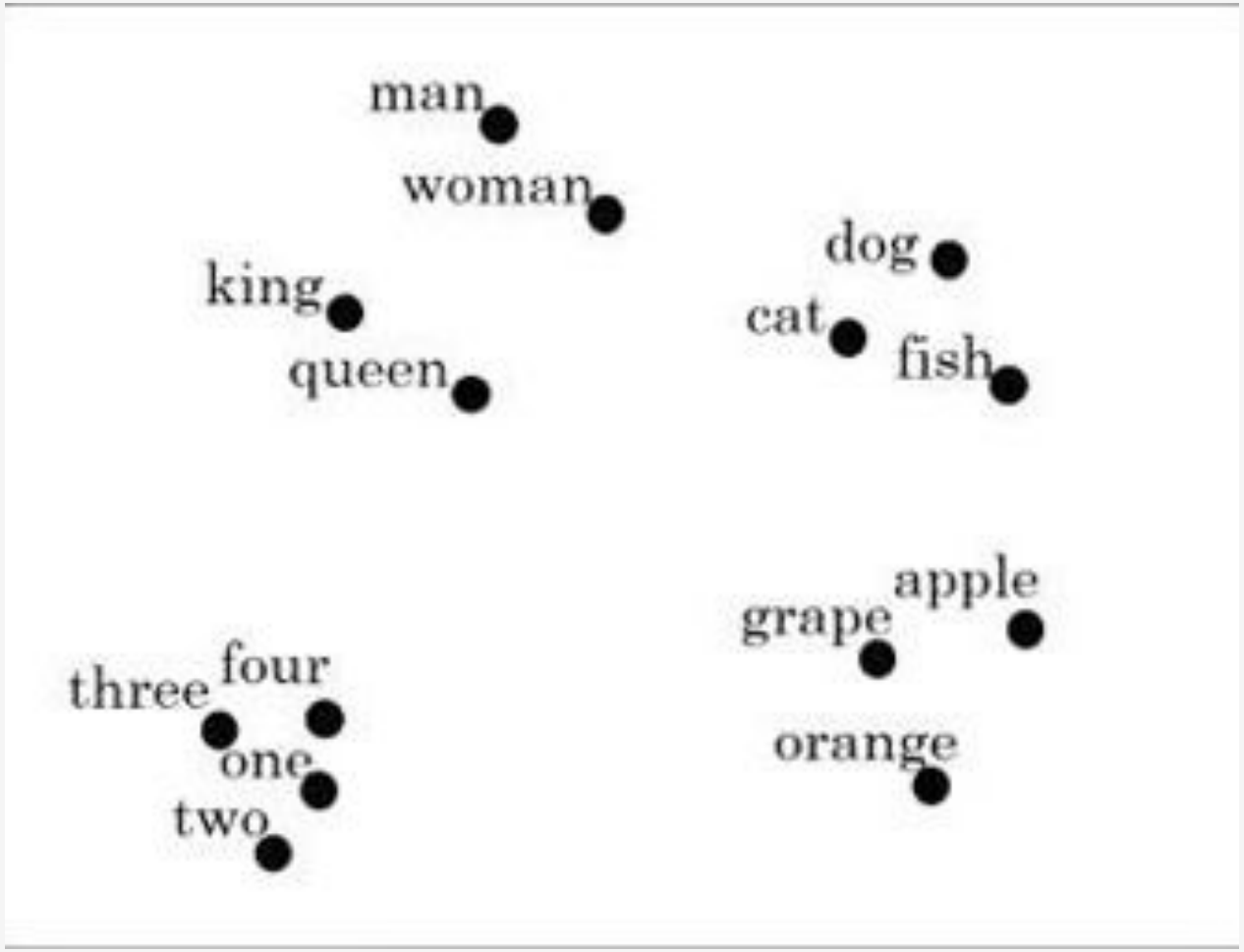
Representación de la arquitectura de la red neuronal

Modelo SkipGram



Para una palabra se tendrá como salida:

- Un vector de reales de varias dimensiones.
- Las palabras más similares son las más “cercanas” semánticamente hablando.



Objetivo

Adaptar un algoritmo para aprender una representación vectorial de palabras a partir de datos de asociación libre (pares de palabras estímulo-respuesta).

Datos de entrenamiento

Un archivo de texto que contiene para una palabra una o más palabras asociadas.

cue	R1	R2	R3
bar	abierto	cerveza	noche
tren	expreso	nocturno	bala
mano	libre	derecha	hermano
sopa	fría	Mafalda	verde
especie	ave	Darwin	extinción
mina	linda	minero	carbón
asco	puaj	freud	feo
gana	pierde	partido	festeja
venta	compra	garage	mercado
iglesia	fuego	cruz	cura
cantar	silbar	alto	canción
papa	puré	pisada	rosada
mercado	libre	fruta	demanda
pescado	frito	barco	apestoso
mientras	tanto	dure	durante
escenario	tabla	madera	teatro
llegado	llegar	fin	viaje
hablar	decir	lengua	comunicar
madre	padre	hermano	hijo
marca	registrada	vender	comprar
comercial	economía	intercambio	publicidad

Método

wordA	wordB	rating	Coseno	Rank Avg Lex	Rank Avg Skipgram
brazo	músculo	1.4	0.300231	344	315
democracia	monarquía	1.3	0.247298	362	450
amigo	profesor	0.4	0.13256	642	742
mano	pie	1.1	0.416457	414	106
disco	ordenador	1.3	0.153859	362	703
banco	asiento	5.1	0.354957	19	199
oveja	ganado	2.7	0.56668	175.5	14
cumpleaños	cita	1.4	0.055688	344	842
hueso	dientes	1.7	0.227395	298	503
frustración	enfado	4.5	0.504115	43.5	40
activo	valores	2.6	0.088826	186.5	814
ensayo	tarea	3.3	0.157449	117.5	687
cuidado	precaución	5.7	0.214091	3.5	538
cena	pollo	1.6	0.218174	312.5	524
palabra	literatura	2.4	0.239277	210	468
hijo	padre	1.6	0.386585	312.5	146
dado	cubo	3.5	0.278069	102.5	359
automóvil	coche	5.5	0.536165	6.5	22
madre	esposa	2.1	0.476302	254	54
arroz	niño	0	0.263583	842	401
inteligencia	lógica	3.4	0.545455	109.5	19

- Tomamos las palabras del SimLex (2M de asociaciones de palabras) e intentamos medir qué tan similares son según nuestro modelo.

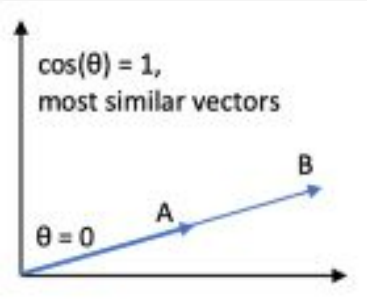
Proceso:

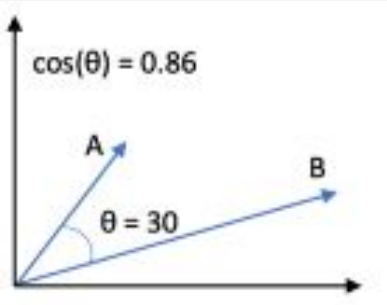
1 - Entrenar el modelo de SkipGram en los datos de asociación.

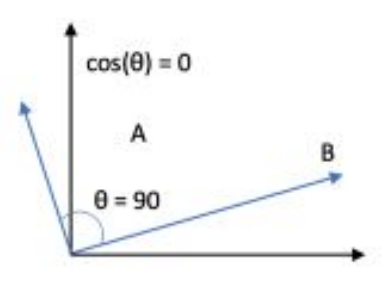
2 - Calcular distancia coseno entre los vectores obtenidos para cada par de palabras del SimLex basado en nuestro modelo.

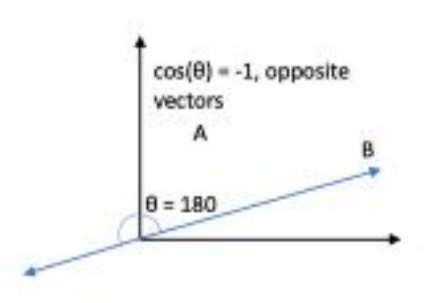
3 - Calcular correlación de Spearman entre la puntuación de SimLex y la distancia coseno obtenidas en el punto 2.

Medidas de rendimiento









- **Correlación de Spearman**: El coeficiente de correlación de Spearman es una medida de la correlación de rango (dependencia estadística del ranking entre dos variables). Se utiliza principalmente para el análisis de datos. Mide la fuerza y la dirección de la asociación entre dos variables clasificadas.
- **Distancia Coseno**: Mide la similitud entre dos vectores calculando el coseno del ángulo entre los dos vectores.
- **¿Por qué usar la similitud coseno?** Si observa la función coseno, es 1 en y= 0 y -1 en y= 180, eso significa que para dos vectores superpuestos, el coseno será el más alto y el más bajo para dos vectores exactamente opuestos.

Parámetros ajustables

Los parámetros buscan optimizar el modelo a través de su combinación

Se llevan a cabo distintos experimentos haciendo variar los parámetros.

n = 600 //Dimensión del vector
vocab_size = 65536//Tamaño del vocabulario en palabras
sequence_length = 10 //Tamaño de una frase en palabras
num_ns = 20 //muestreo negativo
window_size= 2 //Ventana de contexto de las palabras
BATCH_SIZE = 16384
BUFFER_SIZE = 10000
Epoch = 10 //Épocas de entrenamiento

Implementación de skipgram con Tensorflow

Tensorflow es una biblioteca de código abierto para aprendizaje automático.

Método supervisado de entrenamiento.

- +Performante
- +Facil de usar
- +Configurable

Fasttext

- Fasttext es una biblioteca para el aprendizaje de palabras y clasificación de textos

Basado en skipgram incluyendo información de la composición de las palabras.

Método no supervisado de entrenamiento.

+Codigo simple

Resultados de las corridas

	Tensorflow	Fasttext
Correlación	.69	.54
Fuente	2 Millones de asociaciones libres	1 Millón de asociaciones libres
*Tiempo**	200s	300s
Dimensión del vector	600	600
Observaciones	Se ejecutó en el cluster de la Fing con nodos de hasta 120 GB de RAM.	Exige más memoria que Tensor y hay errores en las librerías.

Resultados previos en base a otros métodos muestran una correlación máxima de 0.7 .

Resultados en base a fastText entrenados por terceros muestran correlaciones del orden de 0.5

*Duración del entrenamiento
**Basado en 10 épocas

Preguntas para Reflexionar

- ¿Es suficiente la cantidad de datos de entrada?
- ¿Qué se debe cambiar para mejorar la correlación?
- ¿Están bien elegidos los parámetros?

¡Muchas gracias!

Equipo CICADA

Sitio Web:
https://cicada.uy

Representaciones semánticas, tiempos de reacción y word embeddings;

¿Qué tanto se vinculan?

Objetivos del proyecto

General: Explorar la naturaleza de la relación entre las distancias semánticas y los tiempos de reacción.

Específicos:

1.Caracterizar datos de proyecto "lexicón".

2.Explorar relaciones entre cue y responses en función de los TR.

3.Analizar el peso de la asociación y su vínculo con los TR.

4.Indagar en la posible relación indirectamente proporcional de la frecuencia de las palabras y sus TR asociados.

5. Construir un modelo de representación semántica que considere los TR.

Sobre los datos a utilizar

Análisis de la información

En base a análisis de datos en R:

-Depuración de la base de datos. Descriptivos generales. Análisis de la información obtenida.

En base a lo mencionado previamente, empleando PLN:

construcción de W.E, descriptivos generales. análisis de distancias semánticas.

Proyecto lexicón

Estudio de Asociación Libre, realizado en el Centro de Investigación Básica en Psicología.

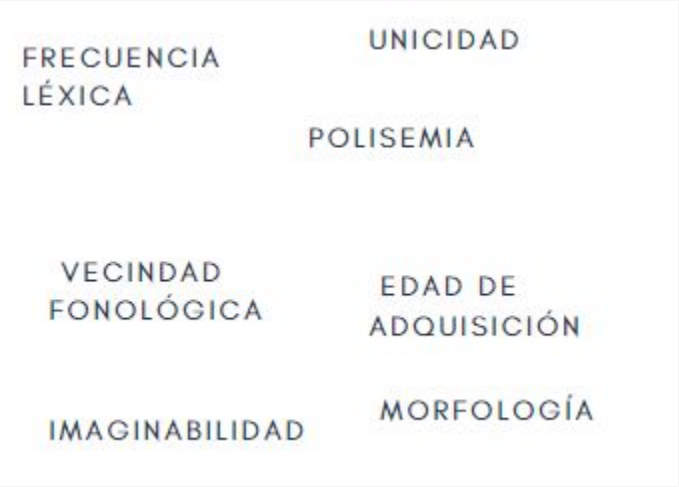
Datos OBTENIDOS

Participantes: 100.000 aprox. Lenguaje: español En su mayoría, Rioplatenses.

Sobre la muestra

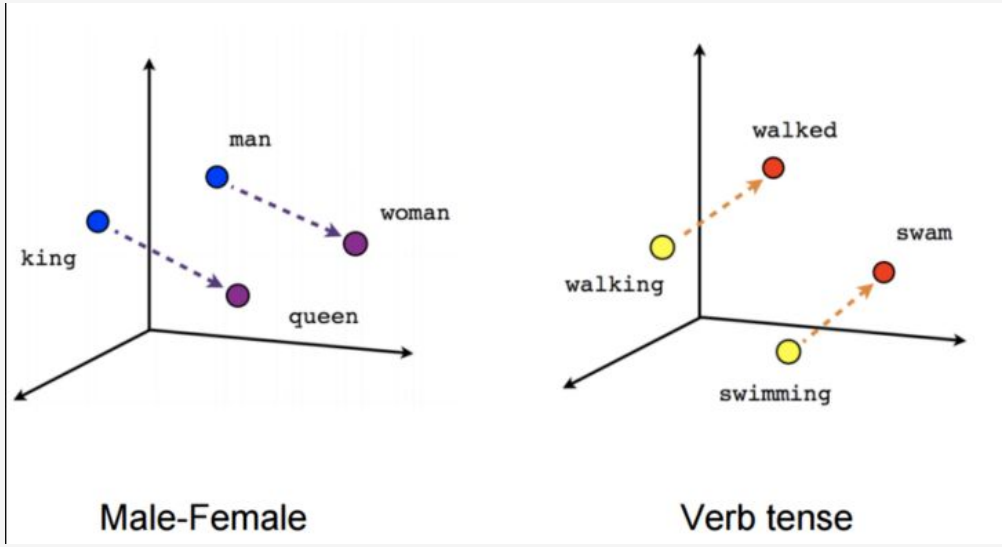
Se obtienen datos sociodemográficos, como sexo, edad, nivel educativo, lengua nativa.

Variables a considerar

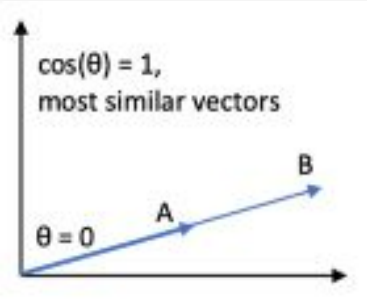


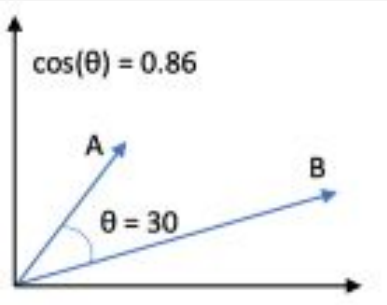
¿Por qué optar por la representación vectorial de palabras?

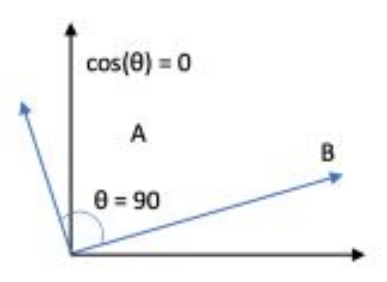
Recientemente se demostró que las palabras vectores capturan muchas regularidades lingüísticas, por ejemplo, operaciones vectoriales vector ('París') - vector ('Francia') + vector ('Italia') da como resultado un vector que está muy cerca de vector ('Roma '), y vector (' rey ') - vector (' hombre ') + vector (' mujer ') está cerca del vector (' reina ').

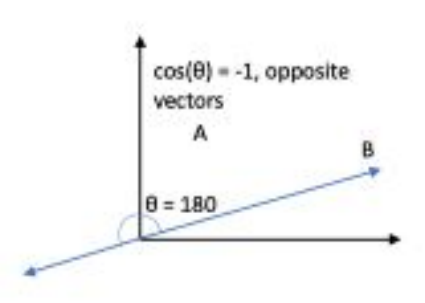


Medidas de rendimiento









- **Distancia Coseno**: Mide la similitud entre dos vectores calculando el coseno del ángulo entre los dos vectores.
- **¿Por qué usar la similitud coseno?** Si observa la función coseno, es 1 en y= 0 y -1 en y= 180, eso significa que para dos vectores superpuestos, el coseno será el más alto y el más bajo para dos vectores exactamente opuestos.
- **Correlación de Spearman**: El coeficiente de correlación de Spearman es una medida de la correlación de rango (dependencia estadística del ranking entre dos variables). Se utiliza principalmente para el análisis de datos. Mide la fuerza y la dirección de la asociación entre dos variables clasificadas.