Word Embeddings

Parte 1

Bibliografía

Jurafsky and Martin 3rd edition. Cap 6. https://web.stanford.edu/~jurafsky/slp3/

Curso Christopher Potts et al.: https://web.stanford.edu/class/cs224u/

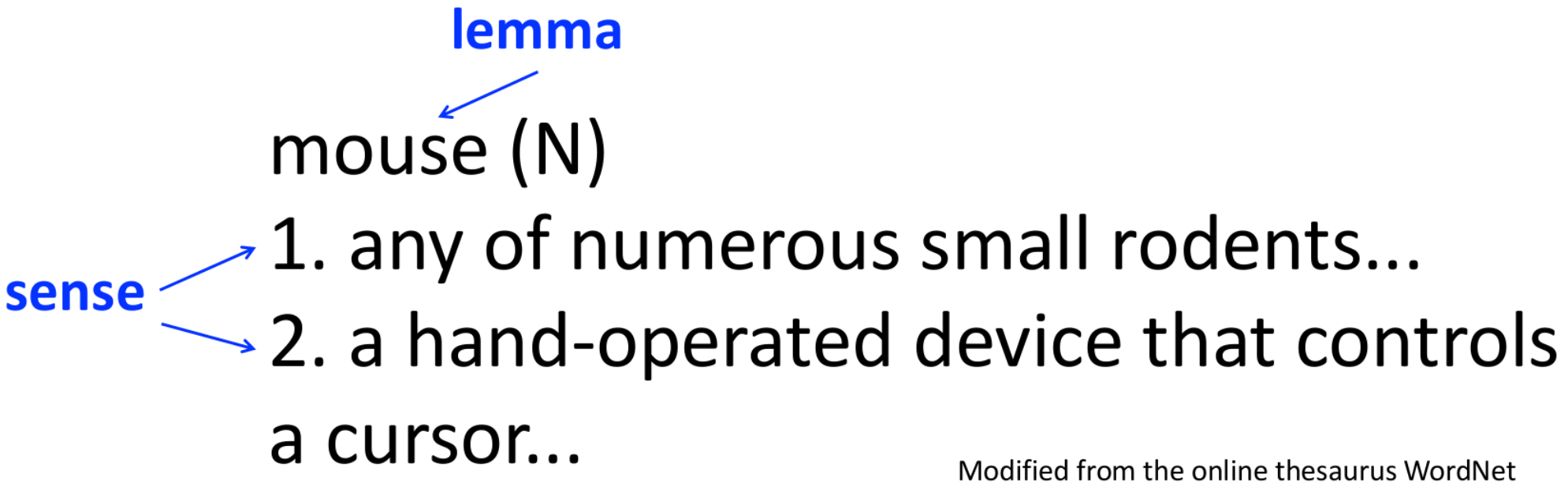
Bibliografía

Jurafsky and Martin 3rd edition. Cap 6. https://web.stanford.edu/~jurafsky/slp3/

Curso Christopher Potts et al.: https://web.stanford.edu/class/cs224u/

En esta presentación se usan slides de ambos recursos. Disculpen la mezcla de inglés y español.

Lemmas and senses



Modified from the online thesaurus WordNet

A sense or "concept" is the meaning component of a word
Lemmas can be polysemous (have multiple senses)

Relaciones de semántica léxica

Relaciones entre significados (sentidos)

Sinonimia perro/canino sillón/sofá automovil/auto

Hiperonimia perro/mamífero rojo/color alegría/emoción

Cohipónimos foca/delfin/ballena rojo/azul/verde alegría/sorpresa/miedo

Antonimia frío/calor alto/bajo rápido/lento

Similarity vs. Relatedness té/café vs té/taza

Campo semántico

Words that

- cover a particular semantic domain
- bear structured relations with each other.

hospitals

surgeon, scalpel, nurse, anaesthetic, hospital

restaurants

waiter, menu, plate, food, menu, chef

houses

door, roof, kitchen, family, bed

Extraído de https://web.stanford.edu/~jurafsky/slp3/

¿y a través del uso de las palabras?

“You shall know a word by the company it keeps” (Firth, 1954)

What does recent English borrowing *ongchoi* mean?

Suppose you see these sentences:

- Ong choi is delicious **sautéed with garlic**.
- Ong choi is superb **over rice**
- Ong choi **leaves** with salty sauces

And you've also seen these:

- ...spinach **sautéed with garlic over rice**
- Chard stems and **leaves** are **delicious**
- Collard greens and other **salty** leafy greens

Conclusion:

- Ongchoi is a leafy green like spinach, chard, or collard greens
  - We could conclude this based on words like "leaves" and "delicious" and "sauteed"

Suponga que ve:
Ongchoi is delicioso salteado con ajo.
Ongchoi es magnífico sobre arroz
Las hojas de ongchoi con salsa saladas.

Y también:
… spinaca salteada con ajo sobre arroz
Hojas y tallos de acelga son deliciosos
Col rizada y otras verduras de hoja verde saladas

Conclusion:
Ongchoi is una verdura con hojas como la espinaca, la acelga y la col rizada.

Se desprende de palabras y frases como hojas, saladas, delicioso, salteado con ajo, sobre arroz

Extraído de https://web.stanford.edu/~jurafsky/slp3/

“You shall know a word by the company it keeps” (Firth, 1954)

Ongchoi: *Ipomoea aquatica "Water Spinach"*

空心菜
kangkong
rau muống
…

Extraído de https://web.stanford.edu/~jurafsky/slp3/

Hipótesis distribucional

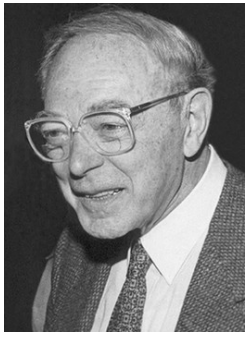
”Las palabras que se usan en los mismos contextos tienen significados parecidos.”

Word Embeddings

Idea 1: Defining meaning by linguistic distribution

(Ferdinand de Saussure, L. Bloomfield, Zellig Harris, John R. Firth)

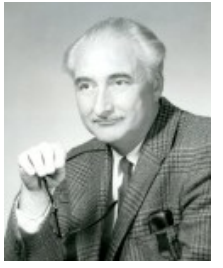
Structural Linguistics



Context of situation

Idea 2: Meaning as a point in space (Osgood et al. 1957)

Semantic Differential



Word Embeddings

Representar el significado de las palabras con vectores, usando los contextos de las palabras en texto.

Significados parecidos → Vectores parecidos

Pricipio Lingüístico de Contraste

Difference in form
→
difference in meaning

Vectores formados a partir de todos los contextos de las palabras.

The meaning of a word is its use in the language. (Wittgenstein, 1953)

De https://web.stanford.edu/~jurafsky/slp3/

Ventajas

- **Se necesita solamente texto y hay mucho disponible (auto-supervisado)**
- **Representaciones de elementos superiores a la palabra (frases, oraciones)**
- **Tareas de PLN (micro-features, transfer learning)**
- **Estudios sobre el lenguaje (diacronía, sesgos)**

¿Por qué nombre word embedding?

La semántica de las palabras "embebida" en un espacio vectorial.

Si consideramos la noción de embedding como función inyectiva que preserva la estructura:

A un **word embedding** lo podemos ver como una **función inyectiva** de una representación **localista (one-hot embedding)** a una representación **distribuida (R^n).**

Es decir, de un espacio vectorial binario donde cada palabra es representada por una componente del espacio.

"casa" - 0….010….000….0 un 1 en la componente de esa palabra y 0 en el resto

"árbol" - 0….000….010….0 vectores de dimensión el tamaño del vocabulario, ej. ~300.000

A **vectores de números reales** de dimensión mucho menor (ej. 300) "preservando" la semántica de las palabras.

"casa" - 1,21.. …. 0,878.. …. 0,369

"árbol" - 0,37.. …. 1,130.. …. 0,527 (vectores de dimensión 300)

El origen del nombre viene de las redes neuronales donde es utilizada la representación one-hot para la entrada de las palabras onteniéndose vectores densos mediante una *look-up table* durante el entrenamiento. Vamos a volver sobre esto más adelante.

Otros nombres: Distributional Semantic Model (DSM), Vector-Space Model, Word Vectors, …

PLN sin Word Embeddings

Palabras como el *string* de sus caracteres (o índices en un vocabulario)

Clases de palabras léxicos de sentimientos, emociones, humor, temporalidad

Bases Léxicas WordNet, Framenet

Reglas, Expresiones Regulares

N-gramas

Lematizar, stemming

Ingeniería de atributos con métodos de aprendizaje automático (ej. Reg. Logística, NB, SVM)

Atributos (*Features*) "diseñados a mano"

Ejemplos - La palabra empieza con in-

- Que la palabra esté precedida por un determinante.
- Pertenecer a cierta lista de palabras.

PLN con Word Embeddings

From now on:
Computing with meaning representations
instead of string representations

荃者所以在鱼，得鱼而忘荃	Nets are for fish;
	Once you get the fish, you can forget the net.
言者所以在意，得意而忘言	Words are for meaning;
	Once you get the meaning, you can forget the words

庄子(Zhuangzi), Chapter 26

Extraído de https://web.stanford.edu/~jurafsky/slp3/

PLN con Word Embeddings

Representaciones vectoriales del significado de las palabras a partir de texto

Transfer Learning. Representar datos de entrenamiento en ML.

Representaciones de unidades mayores a la palabra.

Multimodalidad

- **Distintas modalidades a espacios vectoriales, aprender funciones**
  - **Imágenes**
  - **Sonidos**
  - **Videos**

Bueno.. Empecemos con métodos basados en conteos de contextos

Matrices de frecuencia palabra-documento

El corpus tiene que estar separado en documentos

	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
against	0	0	0	1	0	0	3	2	3	0
age	0	0	0	1	0	3	1	0	4	0
agent	0	0	0	0	0	0	0	0	0	0
ages	0	0	0	0	0	2	0	0	0	0
ago	0	0	0	2	0	0	0	0	3	0
agree	0	1	0	0	0	0	0	0	0	0
ahead	0	0	0	1	0	0	0	0	0	0
ain't	0	0	0	0	0	0	0	0	0	0
air	0	0	0	0	0	0	0	0	0	0
aka	0	0	0	1	0	0	0	0	0	0

Los documento pueden ser:
- libros
- artículos de wikipedia
- párrafos
- etc.

TF-IDF: For a corpus of documents D:

- Term frequency (TF):

$$\frac{x_{ij}}{\textbf{colsum}(X, j)}$$

- Inverse document frequency (IDF):

$$\log_e\left(\frac{|D|}{\left|\{d \in D : w \in d\}\right|}\right)$$

- TF-IDF: TF · IDF

- Se tienen vectores para palabras (filas) y documentos (columnas)
- Se pueden considerar medidas como tf-idf
- Vectores con muchos 0s (sparse)
- Dimensión potencialmente "grande"

Extrído del curso https://web.stanford.edu/class/cs224u/ (Christopher Potts)

LSA (Latent Semtantic Analysis)

Deerwester et al. 1990. Matriz palabra-documento (tf-idf), SVD truncado en el k-ésimo valor singular.

	d1	d2	d3	d4	d5	d6
gnarly	1	0	1	0	0	0
wicked	0	1	0	1	0	0
awesome	1	1	1	1	0	0
lame	0	0	0	0	1	1
terrible	0	0	0	0	0	1

⇓⇑

Distance from *gnarly*
1. gnarly
2. awesome
3. terrible
4. wicked
5. lame

T(erm)					S(ingular values)					D(ocument)					
gnarly	0.41	0.00	0.71	0.00	-0.58	2.45	0.00	0.00	0.00	0.00	d1	0.50	-0.00	0.50	0.00	-0.71
wicked	0.41	0.00	-0.71	0.00	-0.58	0.00	1.62	0.00	0.00	0.00	d2	0.50	0.00	-0.50	0.00	0.00
awesome	0.82	-0.00	-0.00	-0.00	0.58	0.00	0.00	1.41	0.00	0.00	d3	0.50	-0.00	0.50	0.00	0.71
lame	0.00	0.85	0.00	-0.53	0.00	0.00	0.00	0.00	0.62	0.00	d4	0.50	-0.00	-0.50	-0.00	0.00
terrible	0.00	0.53	0.00	0.85	0.00	0.00	0.00	0.00	0.00	-0.00	d5	-0.00	0.53	0.00	-0.85	0.00
											d6	0.00	0.85	0.00	0.53	0.00

$\times$ $\times$ T

gnarly	0.41	0.00		2.45	0.00		gnarly	1.00	0.00
wicked	0.41	0.00		0.00	1.62		wicked	1.00	0.00
awesome	0.82	-0.00	$\times$			$=$	awesome	2.00	0.00
lame	0.00	0.85					lame	0.00	1.38
terrible	0.00	0.53					terrible	0.00	0.85

Distance from *gnarly*
1. gnarly
2. wicked
3. awesome
4. terrible
5. lame

Singular Value Decomposition (SVD)

$$A_{m\times n} = T_{m\times m} S_{m\times m} D^T_{n\times m}$$

$$\begin{pmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix} = \begin{pmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{pmatrix} \begin{pmatrix} \cdot & & \\ & \cdot & \\ & & \cdot \end{pmatrix} \begin{pmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{pmatrix}^T$$

$A_{3\times4} = T_{3\times3} \; S_{3\times3} \; D^T_{4\times3}$

- **Vectores densos (pocos 0s)**
- **Dimensión reducida (Ej. 300)**
- **La dimensión es hiperparám.**

Basado en https://web.stanford.edu/class/cs224u/ (Christopher Potts)

Matrices de frecuencia palabra-palabra

	:)	:/	:D	:|	;p	abandon	abc	ability	able	⋯
:)	74	1	0	0	0	1	0	2	2	
:/	1	306	0	0	0	0	0	0	17	
:D	0	0	16	0	0	0	6	1	1	
:|	0	0	0	120	0	0	0	1	9	
;p	0	0	0	0	516286	0	0	0	0	⋯
abandon	1	0	0	0	0	370	24	65	235	
abc	0	0	6	0	0	24	7948	77	291	
ability	2	0	1	1	0	65	77	4820	1807	
able	2	17	1	9	0	235	291	1807	14328	
⋮					⋮					

	from	swerve	of	shore	to	bend	of	bay	,	brings
Window: 3	4	3	2	1	0	1	2	3	4	5
Scaling: flat	0	1	1	1	1	1	1	1	0	0
Scaling: $\frac{1}{n}$	0	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{1}$	1	$\frac{1}{1}$	$\frac{1}{2}$	$\frac{1}{3}$	0	0

- **Ventanas “grandes” capturan información “mas semántica”**
- **Ventanas “chicas” compuran información “más sintáctica”**
- **Se pueden considar valores como PMI o PPMI**

- **Vectores de dimensión grande (#palabras)**
- **Vectores con muchas entradas en 0 (sparse)**

Basado en https://web.stanford.edu/class/cs224u/ (Christopher Potts)

HAL (Hyperspace Analogue to Language)

(Lund and Burgess, 1996) Producing high-dimensional semantic spaces from lexical co-occurrence

Matriz de frecuencia palabra-palabra NxN con N la cantidad de palabras

Celdas $f_{ij} = v - d_{ij}$ con v la distancia máxima de la ventana y d_{ij} la distancia entre las palabras

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}
0	7	6	5	4	3	2	1	0	

Se concancatenan filas y columnas obteniendo vectores de diemsión 2N

Reducción de dimensión con SVD

PPMI - SVD

Levy and Goldberg, 2014. Neural Word Embedding as Implicit Matrix Factorization

También en Baroni et al., 2014: Don't Count, Predict! A systematic comparison of context-counting vs contex-predicting semantic vectors.

Matriz PPMI palabra-palabra.

LG propone Shifted PPMI dando mejores resultados en algunas tareas

$$SPPMI_k(w, c) = \max (PMI(w, c) - \log k, 0)$$

Reducción de dimensión con SVD. LG propone Symmetric SVD (Sect 4.2)

GloVe

Pennington, Socher and Manning, 2014. GloVe: Global Vectors for Word Representations

Basado en que el cociente de probs. condicionadas codifica cierta información semántica

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

Pueden descargar vectores del sitio y una implementación de GloVe

Para codificar P(k|i)/P(k|j) con la resta proponen que

$$w_i^T \tilde{w}_k = \log(P_{ik}) = \log(X_{ik}) - \log(X_i)$$

Minimizando

$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2 \qquad f(x) = \begin{cases} (x/x_{\max})^\alpha & \text{if } x < x_{\max} \\ 1 & \text{otherwise} \end{cases}$$

Extraído de https://nlp.stanford.edu/pubs/glove.pdf

Related Words (vectore distance)

Distancia Euclídea

$$\textbf{euclidean}(u, v) = \sqrt{\sum_{i=1}^{n} |u_i - v_i|^2}$$

Distancia coseno

$$\textbf{cosine}(u, v) = 1 - \frac{\sum_{i=1}^{n} u_i \times v_i}{||u||_2 \times ||v||_2}$$

Similitud coseno

$$\frac{\vec{v} \cdot \vec{w}}{|\vec{v}||\vec{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2} \sqrt{\sum_{i=1}^{N} w_i^2}}$$

Palabras similares a la palabra “frog”

0. *frog*
1. frogs
2. toad
3. litoria
4. leptodactylidae
5. rana
6. lizard
7. eleutherodactylus

3. litoria

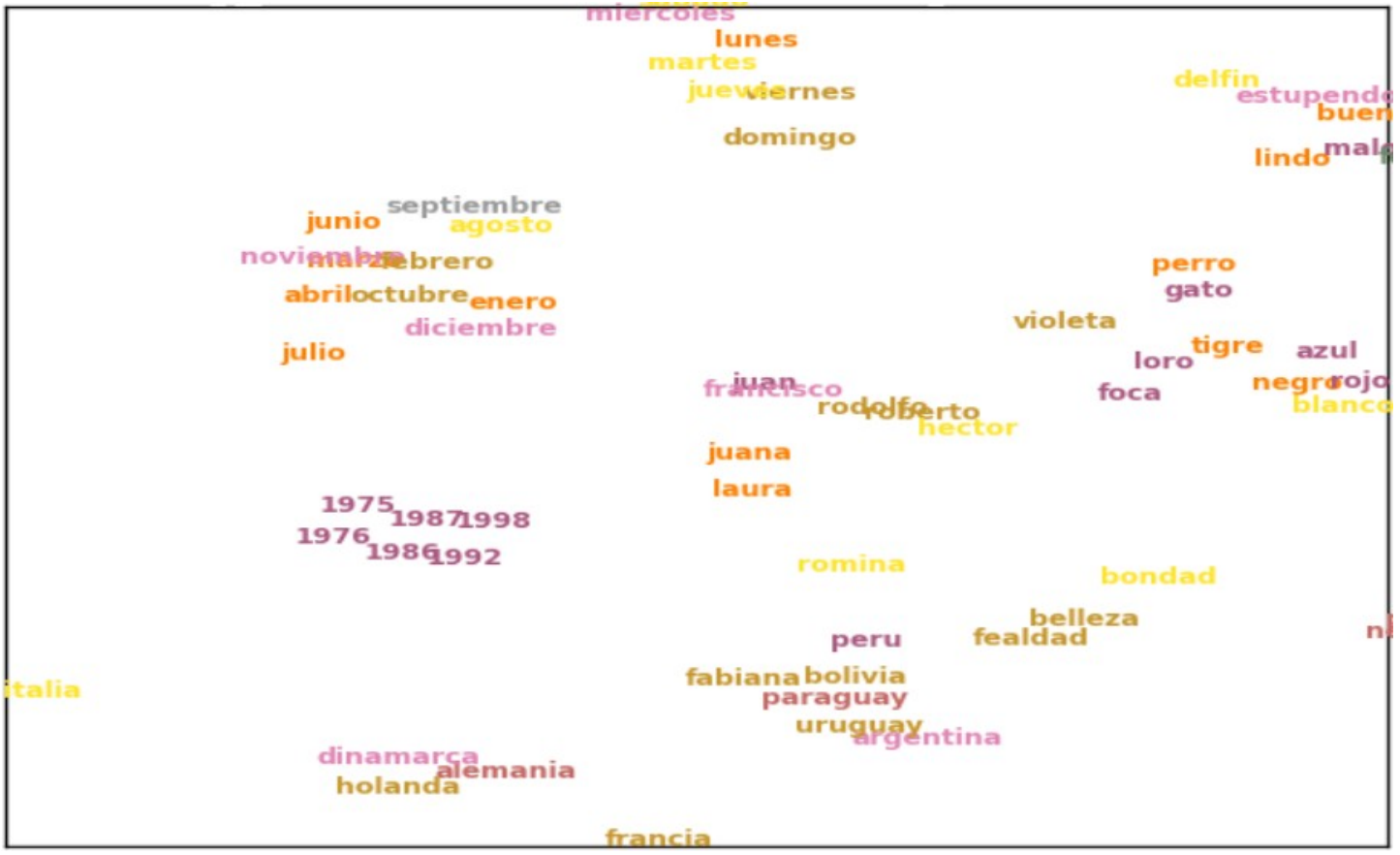
4. leptodactylidae

5. rana

7. eleutherodactylus

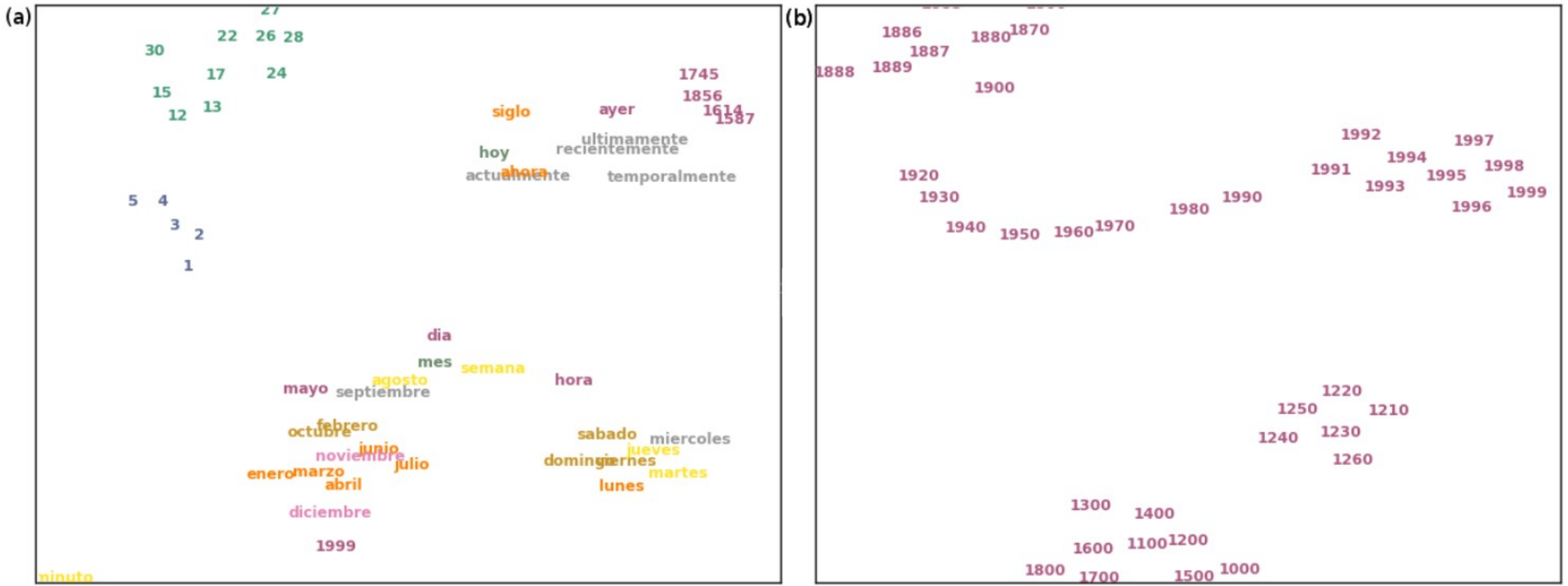
Extraído de https://nlp.stanford.edu/projects/glove/

Related Words (vector distance)



Extraído de https://aclanthology.org/L16-1584.pdf

Related Words (vector distance)



Extraído de https://drops.dagstuhl.de/opus/volltexte/2017/7925/pdf/LIPIcs-TIME-2017-12.pdf

Related Words (vector distance)

Primero	Segundo	Vigésimo	1853	1850	1700	1999
luego	tercer	trigésimo	1855	1840	1600	1998
segundo	primer	décimo	1854	1849	1800	1995
mismo	cuarto	cuadragésimo	1856	1870	1500	1997
último	último	noveno	1852	1860	1400	1996
primer	quinto	quincuagésimo	1851	1880	1200	2002
posteriormente	primero	octavo	1865	1851	1100	2003
después	tercero	quinto	1849	1830	1300	1994
...	...	...	...	...	...	...

Extraído de https://drops.dagstuhl.de/opus/volltexte/2017/7925/pdf/LIPIcs-TIME-2017-12.pdf

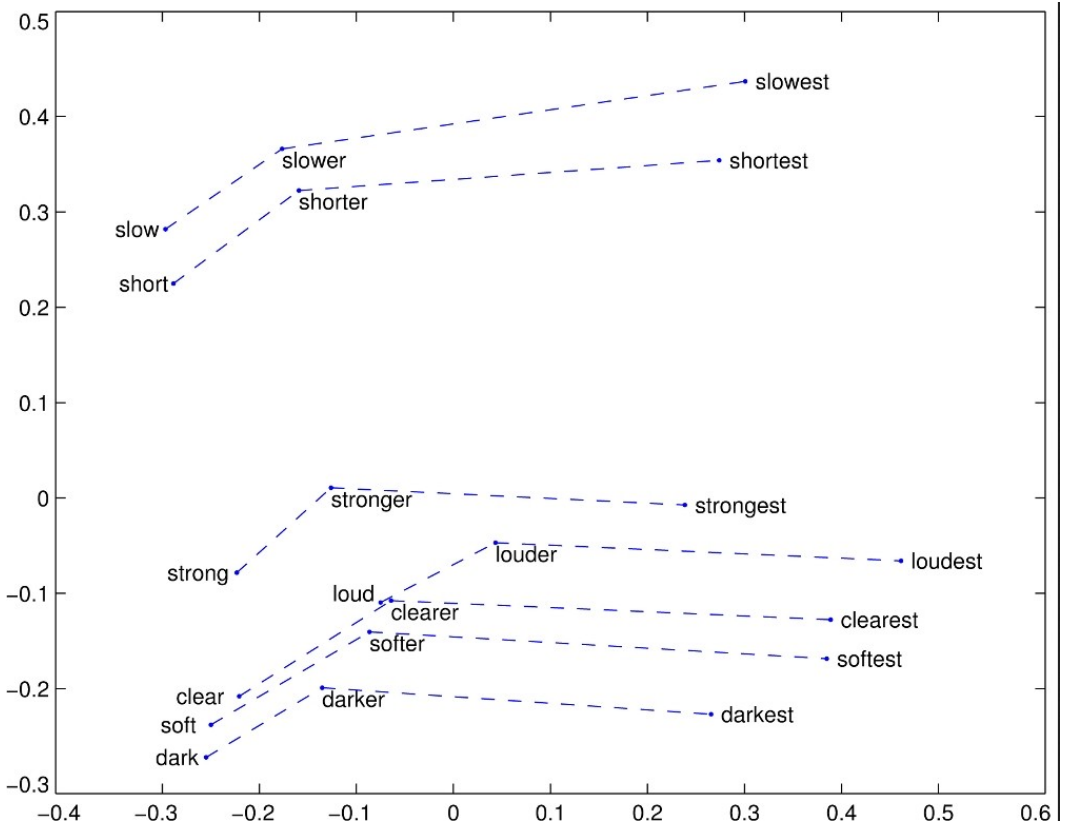
Related Words (vector distance)

Amanecer	Neolítico	Comienzo	Antes	Repentinamente	Apresuradamente
atardecer	paleolítico	inicio	después	súbitamente	marchar
mañana	mesolítico	dio	tras	muere	replegarse
noche	calcolítico	antes	ya	falleció	precipitadamente
día	neolítico	final	días	murió	desecaba
medianoche	datan	dando	luego	prematuramente	mudarse
anochece	pleistoceno	llegada	ese	trágicamente	periódicamente
mediodía	precerámico	finales	meses	tempranamente	dirigiera
ocaso	epipaleolítico	principio	tiempo	...	...
madrugada	bronce	momento	comenzar		
...	...	...	...		

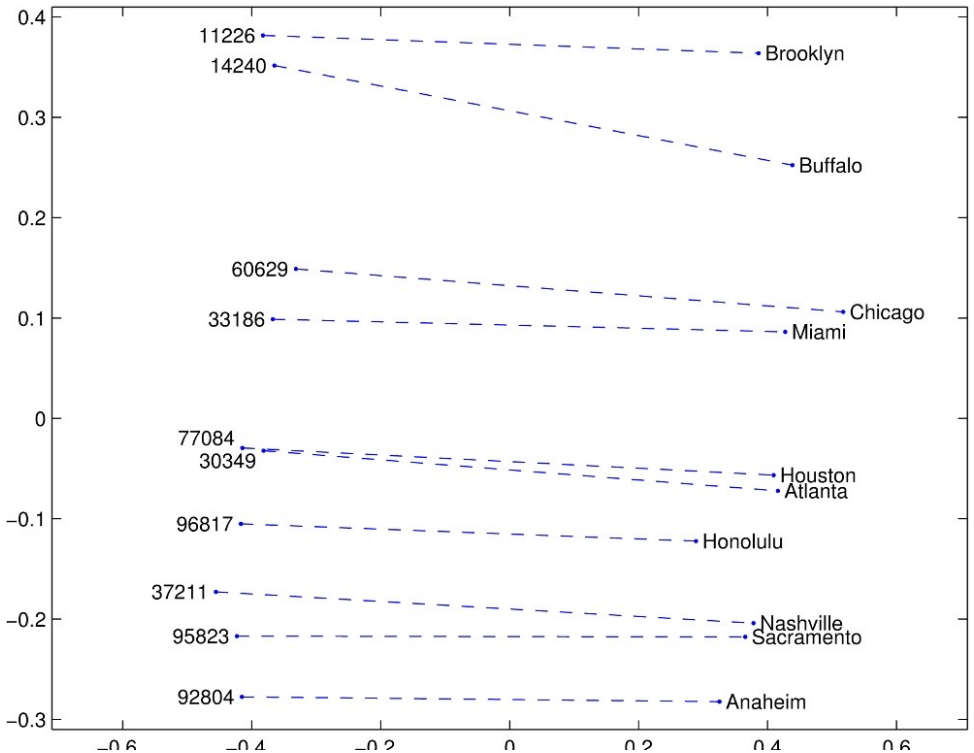
Extraído de https://drops.dagstuhl.de/opus/volltexte/2017/7925/pdf/LIPIcs-TIME-2017-12.pdf

Linear Substructures (vector difference)

Adjective – comparative – superlative



Zip Code – US City



Extraído de https://nlp.stanford.edu/projects/glove/

Los métodos anteriores pueden encontrarse como basados en conteo de contextos.
[(Baroni et al., 2014) Don’t count, predict!]

Consisten en una matriz de frecuencias y una reducción de la dimensión.

(GloVe podría verse como una caso “especial”)

En la que viene seguimos con word embeddings:

- Métodos basados en predicción de contextos

- Evaluación de word embeddings