Laboratorio de datos: web scraping y Procesamiento de Lenguaje Natural

Clase 7. Un acercamiento a los word embeddings

factor~data
EIDAES_UNSAM

Hipótesis distribucional

- "El significado deriva del uso de las palabras en el lenguaje" (Wittgenstein)
- Podemos captar el sentido de las palabras según su "compañía"
- Palabras cercanas tienen sentidos "cercanos"
- Ítems lingüísticos con distribuciones similares tienen significados similares"
- Idea de co-ocurrencia => términos que ocurren juntos

TFM Co-ocurrencia a nivel documento

Palabras, bigramas, trigramas, lemas, solo la raíz de la palabra...

F términos

$$\text{Matriz } M = \begin{array}{c} \\ \textbf{\textit{N documentos}} \end{array} \begin{bmatrix} V_{11} & V_{12} & V_{13} & \cdots & V_{1F} \\ & & & & \\ V_{N1} & V_{N2} & V_{N3} & \cdots & V_{NF} \end{bmatrix}$$

Frecuencia del término

Fuente: https://github.com/gefero/ws_text_mining/blob/master/slides/04_pinto.pdf

- La matriz de documentos-términos suele tener muchos ceros
- Problema: se hace difícil medir la relación entre los distintos documentos o términos

	Palabra 1	Palabra 2	Palabra 3	Palabra 4	Palabra 5	
Relato 1	0	0.12	0.01	0	0	
Relato 2	0	0	0.44	0.15	0.65	
Relato 3	0.11	0.31	0.28	0	0	(...)
Relato 4	0	0	0.05	0.21	0	
Relato 5	0	0.13	0	0.07	0	
			(...)			

La correlación lineal entre filas nos da una idea de la similitud del significado entre relatos

La correlación lineal entre columnas nos da una idea de la similitud del significado entre palabras

Pero hay un problema: la mayor parte de los valores son 0

"Sobre la mesa hay un florero con margaritas y jazmines"

"El vaso lleno de flores está apoyado sobre una mesada"

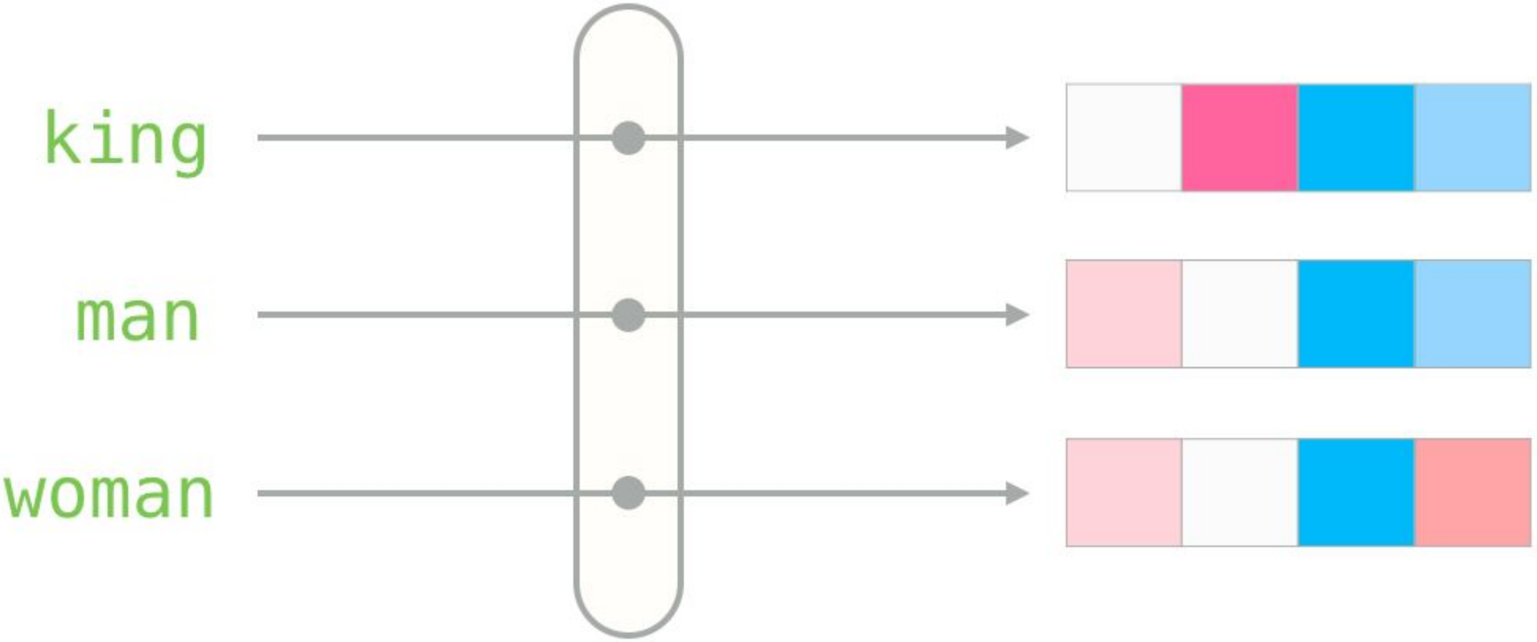
- Mismo sentido pero ninguna palabra en común
- Una solución ya la vimos: LDA, STM => detección de tópicos

- **Otra solución: word embeddings**

Word embeddings => idea general

- Reducir la dimensión del vocabulario
  - ~50.000 palabras a ~100 => representación no “esparsa” sino densa
- Flexibilizar supuestos de BoW: cada columna/término/dimensión es un término y se asume independencia
- Hay interacción entre palabras => es esperable que la dimensionalidad sea menor
- Lograr introducir una métrica de distancia para que palabras “cerca” en el nuevo espacio estén “cerca” semánticamente estén cerca.

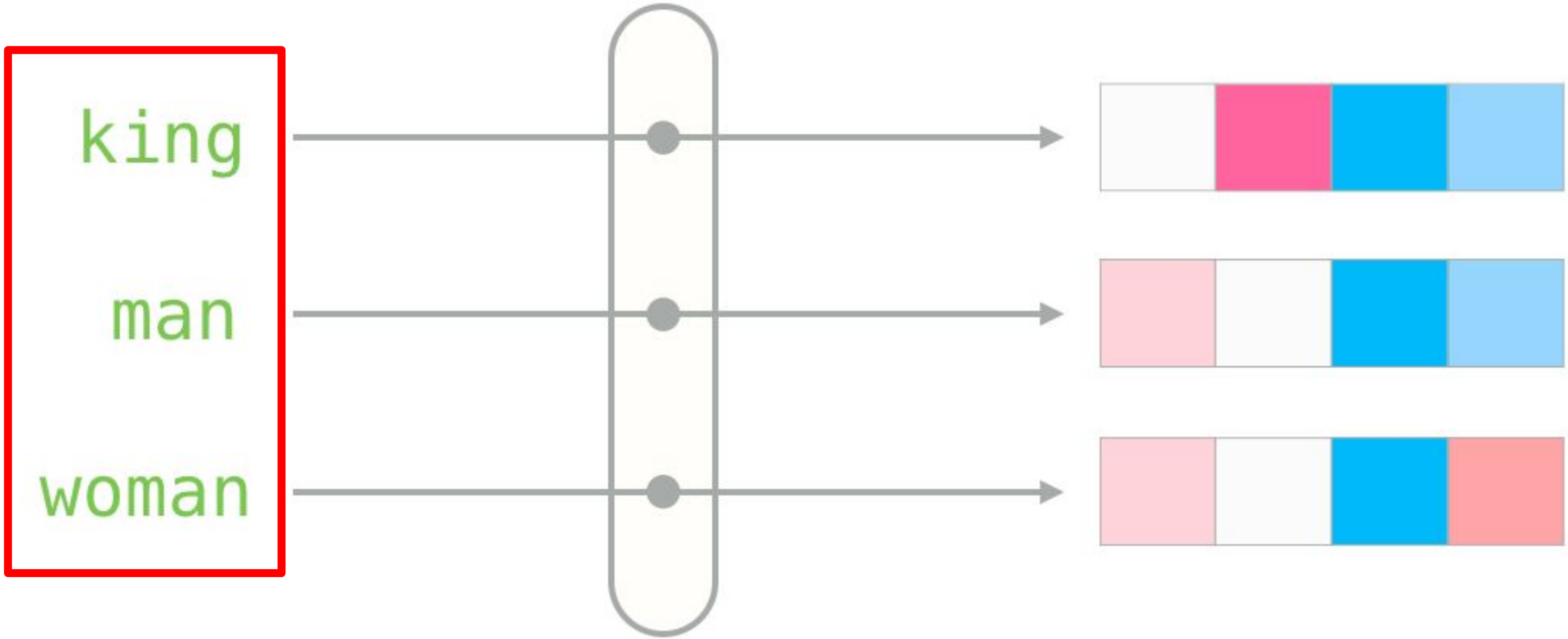
word2vec



factor~data
EIDAES_UNSAM

Fuente: https://jalammar.github.io/illustrated-word2vec/

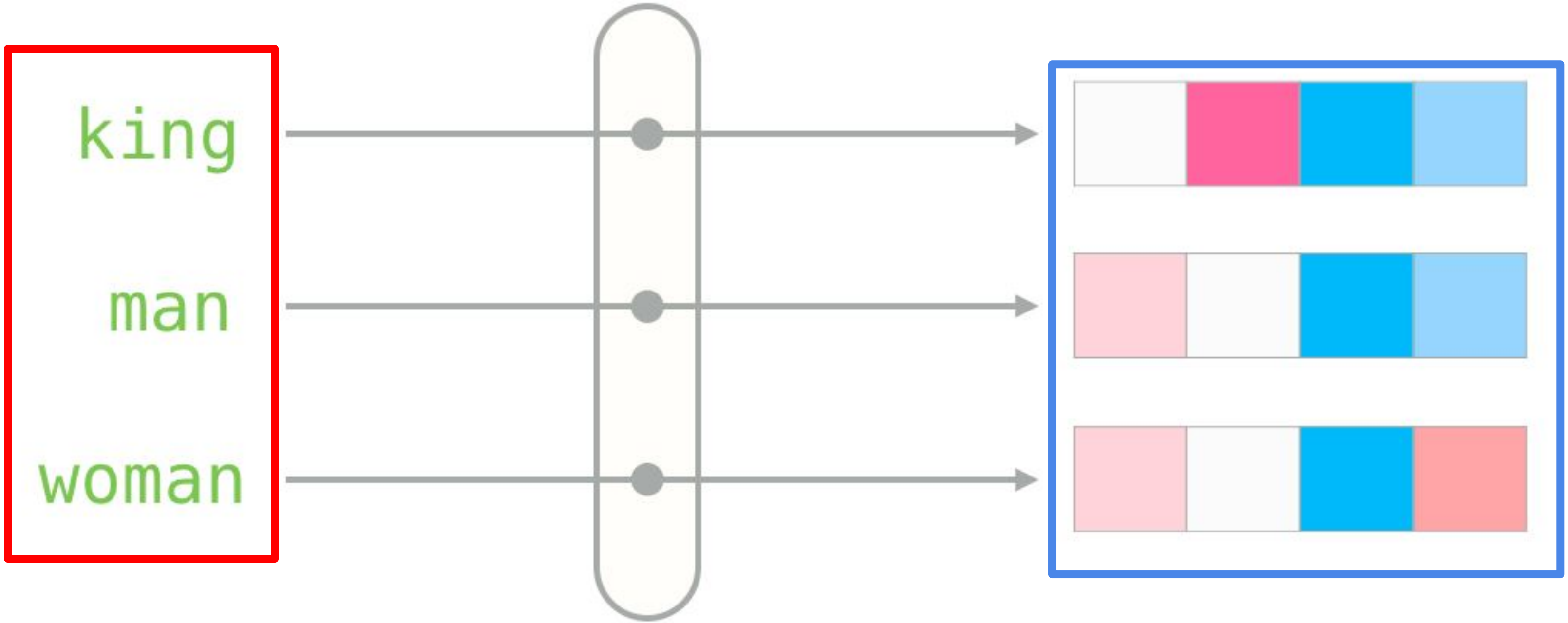
word2vec



factor~data
EIDAES_UNSAM

Fuente: https://jalammar.github.io/illustrated-word2vec/

word2vec

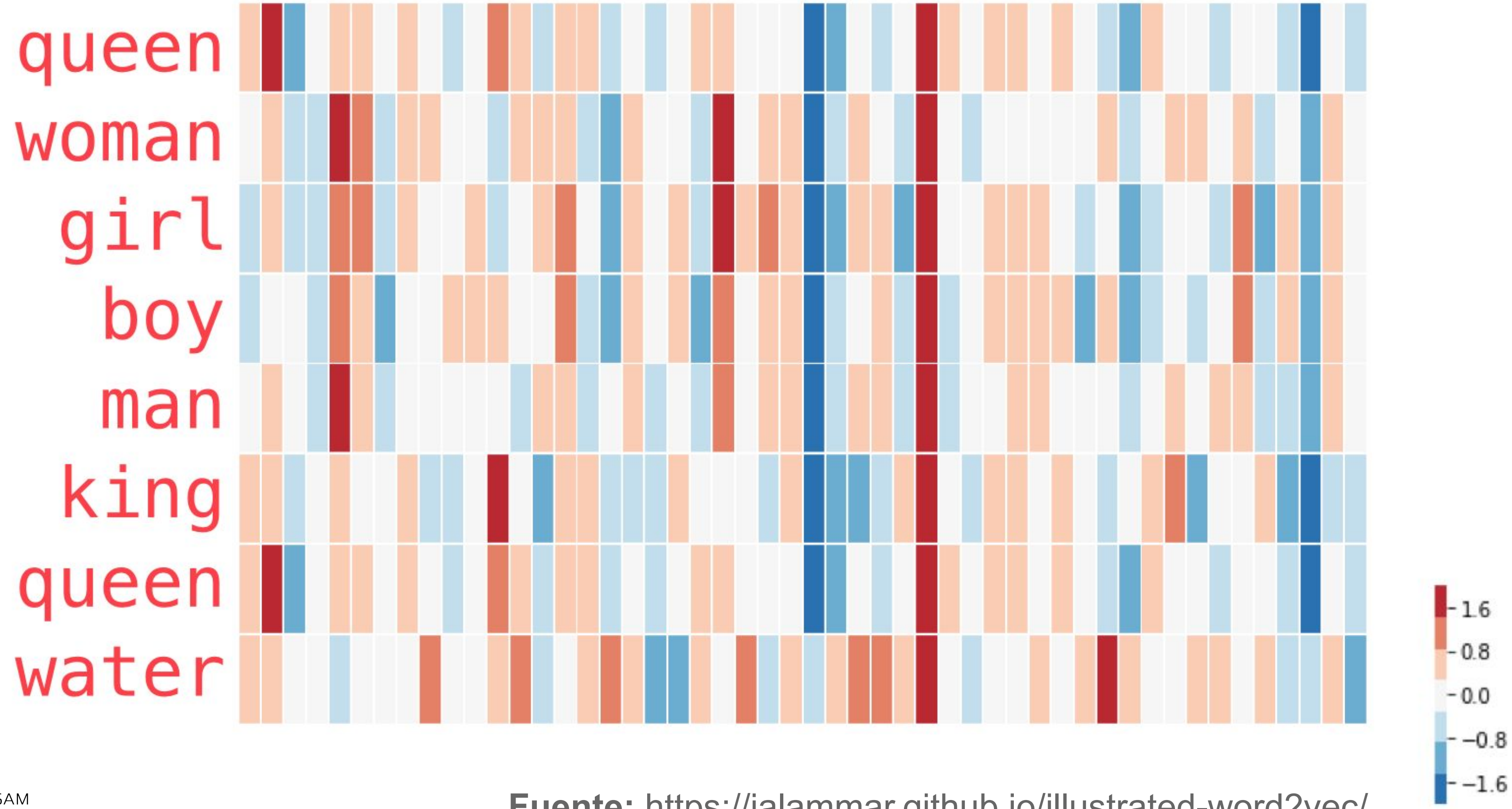


Fuente: https://jalammar.github.io/illustrated-word2vec/

word2vec

Fuente: https://jalammar.github.io/illustrated-word2vec/

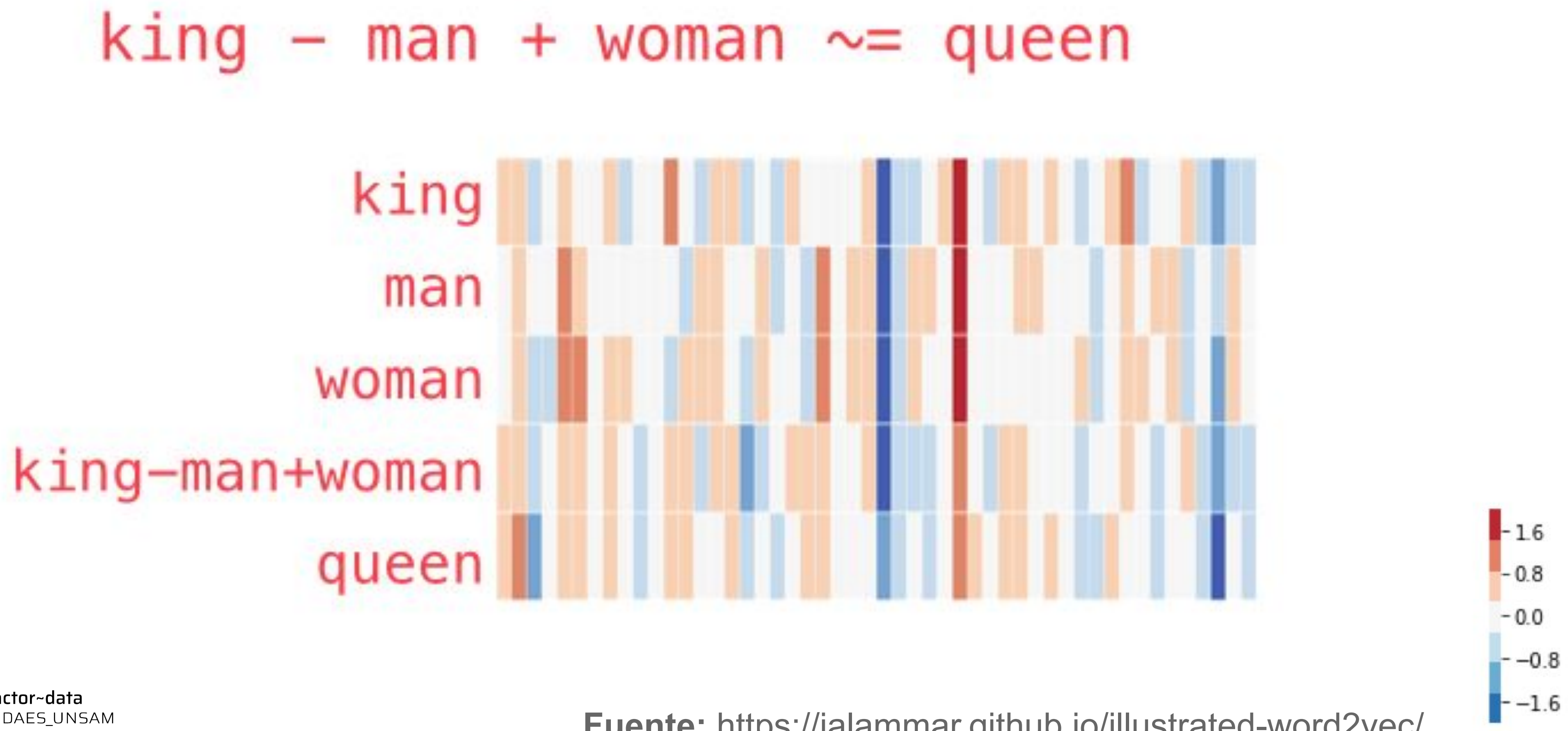
word2vec



Fuente: https://jalammar.github.io/illustrated-word2vec/

factor~data
EIDAES_UNSAM

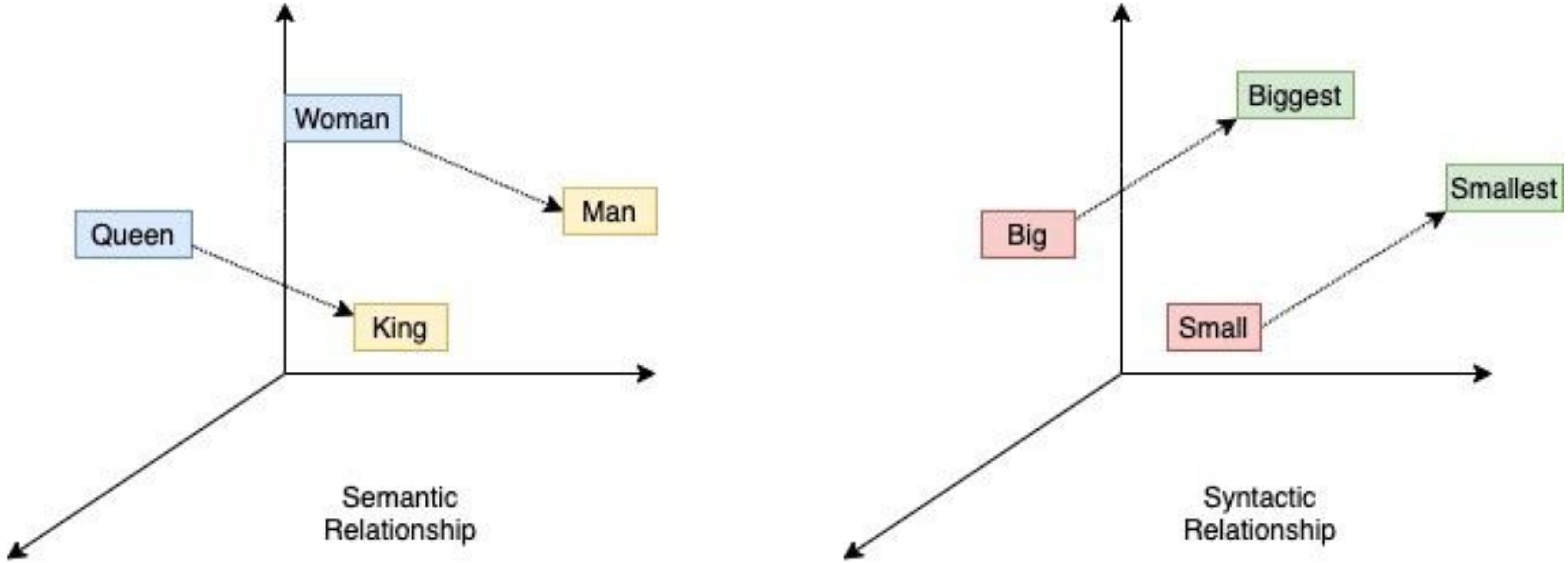
word2vec



Fuente: https://jalammar.github.io/illustrated-word2vec/

factor~data
EIDAES_UNSAM

word2vec



factor~data
EIDAES_UNSAM

Evaluación de embeddings

Table 1: *Examples of five types of semantic and nine types of syntactic questions in the Semantic-Syntactic Word Relationship test set.*

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

Fuente: Efficient Estimation of Word Representations in Vector Space (2013) https://arxiv.org/abs/1301.3781 Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean

Evaluación de embeddings

Table 4: *Comparison of publicly available word vectors on the Semantic-Syntactic Word Relationship test set, and word vectors from our models. Full vocabularies are used.*

Model	Vector Dimensionality	Training words	Accuracy [%]		
			Semantic	Syntactic	Total
Collobert-Weston NNLM	50	660M	9.3	12.3	11.0
Turian NNLM	50	37M	1.4	2.6	2.1
Turian NNLM	200	37M	1.4	2.2	1.8
Mnih NNLM	50	37M	1.8	9.1	5.8
Mnih NNLM	100	37M	3.3	13.2	8.8
Mikolov RNNLM	80	320M	4.9	18.4	12.7
Mikolov RNNLM	640	320M	8.6	36.5	24.6
Huang NNLM	50	990M	13.3	11.6	12.3
Our NNLM	20	6B	12.9	26.4	20.3
Our NNLM	50	6B	27.9	55.8	43.2
Our NNLM	100	6B	34.2	**64.5**	50.8
CBOW	300	783M	15.5	53.1	36.1
Skip-gram	300	783M	**50.0**	55.9	**53.3**

Fuente: Efficient Estimation of Word Representations in Vector Space (2013) https://arxiv.org/abs/1301.3781 Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean

Evaluación de embeddings

Table 8: *Examples of the word pair relationships, using the best word vectors from Table 4 (Skip-gram model trained on 783M words with 300 dimensionality).*

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

Fuente: Efficient Estimation of Word Representations in Vector Space (2013) https://arxiv.org/abs/1301.3781 Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean

Usos posibles

- Similitud entre palabras y documentos
- Similitud entre palabras “target” y palabras de contexto al resultado

- Autocompletado
- Traducción automática
- Encontrar clusters de palabras con significados similares
- Buscar analogías entre palabras

- Modelo semántico del lenguaje para comparar con procesamiento del lenguaje hecho por humanos

Aplicaciones en Ciencias Sociales - Estereotipos

American Sociological Review
2019, Vol. 84(5) 905–949
© American Sociological
Association 2019
DOI: 10.1177/0003122419877135
journals.sagepub.com/home/asr
SAGE

The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings

Austin C. Kozlowski,[a] Matt Taddy,[b]
and James A. Evans[a,c]

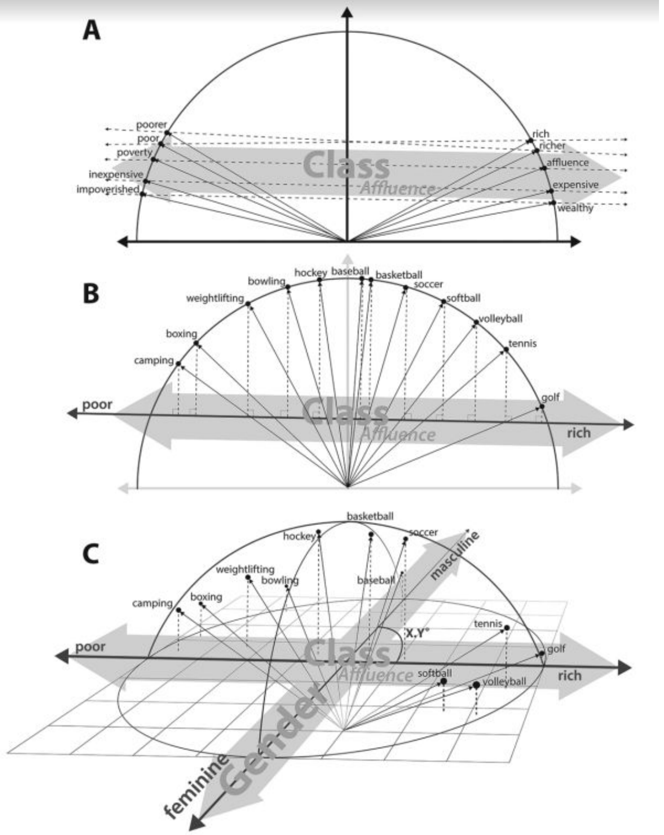
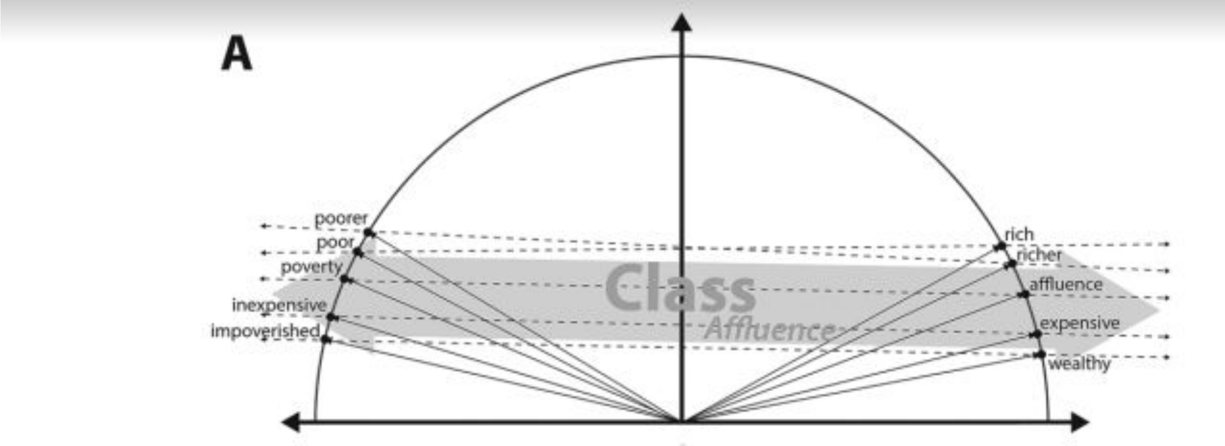


Figure 2. Conceptual Diagram of (A) the Construction of a Cultural Dimension; (B) the Projection of Words onto That Dimension; and (C) the Simultaneous Projection of Words onto Multiple Dimensions

factor~data
EIDAES_UNSAM

Aplicaciones en Ciencias Sociales - Estereotipos

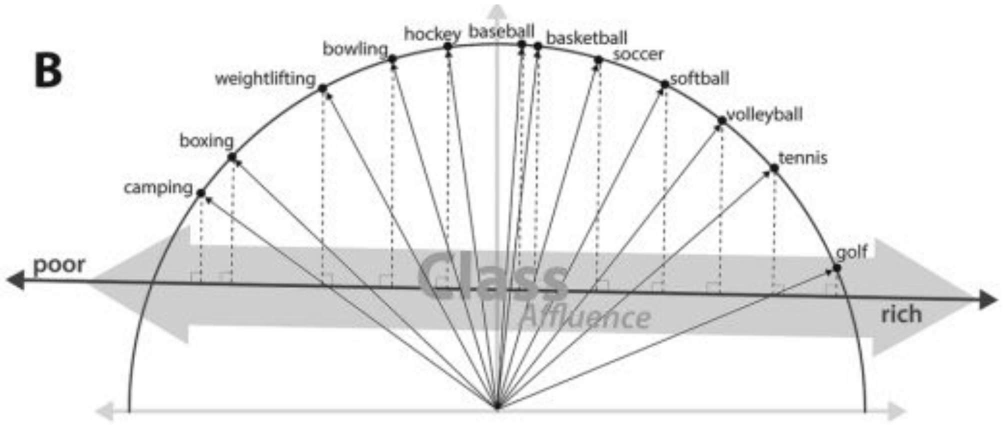


Measuring Cultural Dimensions

To identify cultural dimensions in word embedding models, we average numerous pairs of antonym words. Cultural dimensions are calculated by simply taking the mean of all word pair differences that approximate a given dimension, $\frac{\sum_{p}^{|P|} \overrightarrow{p_1} - \overrightarrow{p_2}}{|P|}$, where p are all antonym word pairs in relevant set P, and $\overrightarrow{p_1}$ and $\overrightarrow{p_2}$ are the first and second word vectors of each pair.[17] The projection of a normalized word vector onto a cultural dimension is calculated with cosine similarity, as is the angle between cultural dimensions.

factor~data
EIDAES_UNSAM

Aplicaciones en Ciencias Sociales - Estereotipos



Aplicaciones en Ciencias Sociales - Estereotipos

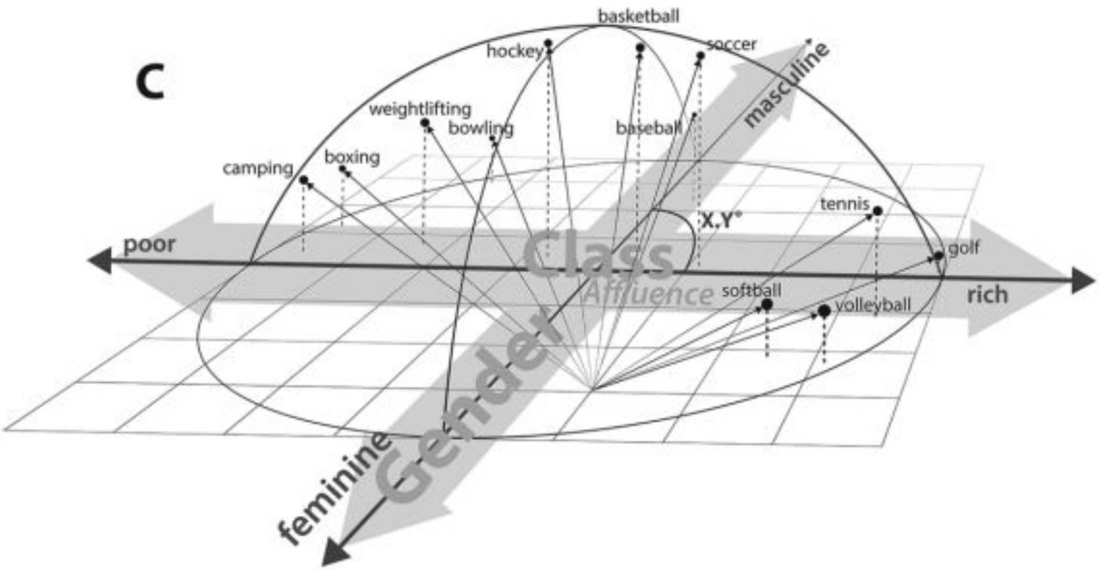


Figure 2. Conceptual Diagram of (A) the Construction of a Cultural Dimension; (B) the Projection of Words onto That Dimension; and (C) the Simultaneous Projection of Words onto Multiple Dimensions

Aplicaciones en Ciencias Sociales - Estereotipos

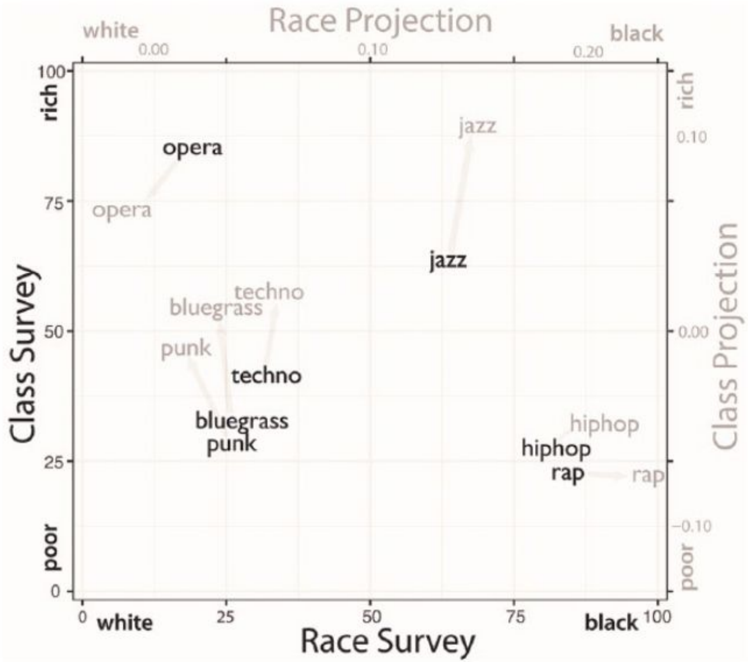


Figure 3. Projection of Music Genres onto Race and Class Dimensions of the Google News Word Embedding (Gray) and Average Survey Ratings for Race and Class Associations (Black)

factor~data
EIDAES_UNSAM

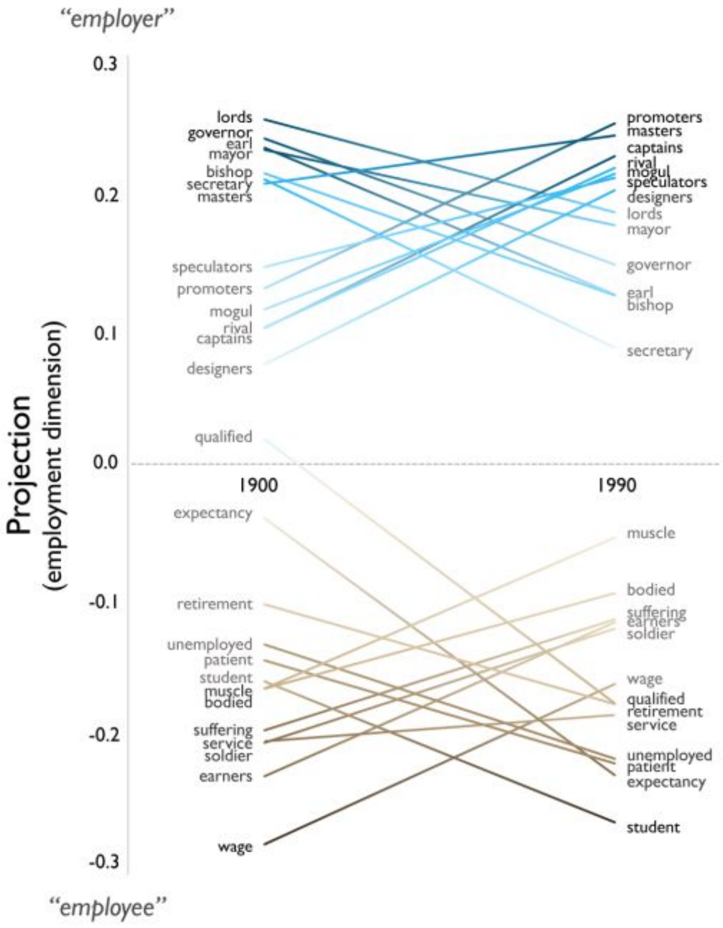


Figure 10. Words That Project High and Low on the Employment Dimension of Word Embedding Models Trained on Texts Published at the Beginning and End of the Twentieth Century; 1900–1919 and 1980–1999 Google Ngrams Corpus

Aplicaciones en Ciencias Sociales

npj Schizophrenia
www.nature.com/npjschz

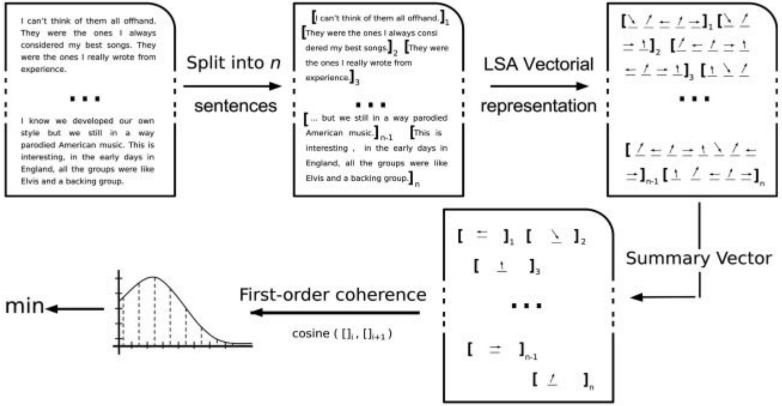
ARTICLE OPEN

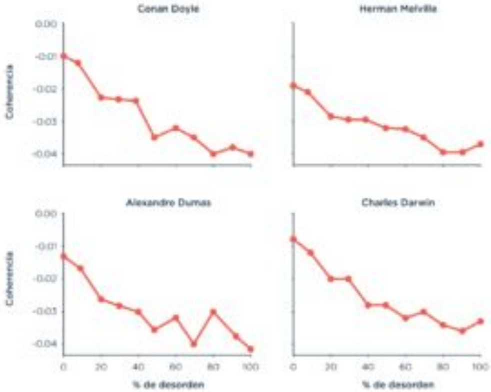
Automated analysis of free speech predicts psychosis onset in high-risk youths

Gillinder Bedi[1,2,9], Facundo Carrillo[3,9], Guillermo A Cecchi[4], Diego Fernández Slezak[3], Mariano Sigman[5], Natália B Mota[6], Sidarta Ribeiro[6], Daniel C Javitt[1,7], Mauro Copelli[8] and Cheryl M Corcoran[1,7]

BACKGROUND/OBJECTIVES: Psychiatry lacks the objective clinical tests routinely used in other specializations. Novel computerized methods to characterize complex behaviors such as speech could be used to identify and predict psychiatric illness in individuals.
AIMS: In this proof-of-principle study, our aim was to test automated speech analyses combined with Machine Learning to predict later psychosis onset in youths at clinical high-risk (CHR) for psychosis.
METHODS: Thirty-four CHR youths (11 females) had baseline interviews and were assessed quarterly for up to 2.5 years; five transitioned to psychosis. Using automated analysis, transcripts of interviews were evaluated for semantic and syntactic features predicting later psychosis onset. Speech features were fed into a convex hull classification algorithm with leave-one-subject-out cross-validation to assess their predictive value for psychosis outcome. The canonical correlation between the speech features and prodromal symptom ratings was computed.
RESULTS: Derived speech features included a Latent Semantic Analysis measure of semantic coherence and two syntactic markers of speech complexity: maximum phrase length and use of determiners (e.g., *which*). These speech features predicted later psychosis development with 100% accuracy, outperforming classification from clinical interviews. Speech features were significantly correlated with prodromal symptoms.
CONCLUSIONS: Findings support the utility of automated speech analysis to measure subtle, clinically relevant mental state changes in emergent psychosis. Recent developments in computer science, including natural language processing, could provide the foundation for future development of objective clinical tests for psychiatry.

npj Schizophrenia (2015) **1**, Article number: 15030; doi:10.1038/npjschz.2015.30; published online 26 August 2015





factor~data
EIDAES_UNSAM

Aplicaciones en Ciencias Sociales - Estereotipos

Semantics derived automatically from language corpora contain human-like biases

Aylin Caliskan,[1*] Joanna J. Bryson,[1,2*] Arvind Narayanan[1*]

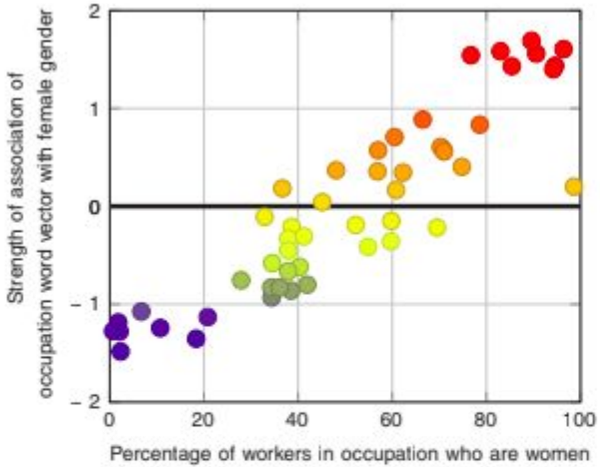


Fig. 1. Occupation-gender association. Pearson's correlation coefficient $\rho = 0.90$ with $P < 10^{-18}$.

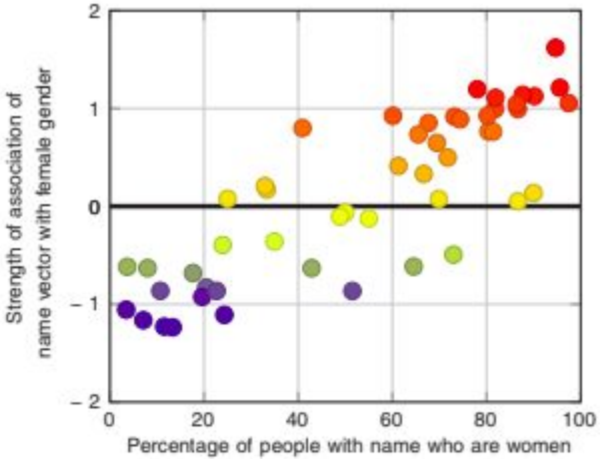


Fig. 2. Name-gender association. Pearson's correlation coefficient $\rho = 0.84$ with $P < 10^{-13}$.

Aplicaciones en Ciencias Sociales - Trayectorias

- Arquitectura de embeddings / transformers como forma de generar representaciones comprimidas de trayectorias en múltiples dimensiones.
- Life2Vec
  - Dataset n~3.000.000 de habitantes
    - Historias laborales / ingresos
    - Historias migratorias
    - Historias de salud
- Tarea: predicción de fallecimiento
- Usan algo similar a lo que funciona por detrás de chatGPT

nature computational science

Article

https://doi.

Using sequences of life-events to predict human lives

Received: 6 June 2023

Accepted: 15 November 2023

Published online: 18 December 2023

Check for updates

Germans Savcisens[1], Tina Eliassi-Rad[2,3], Lars Kai Hansen[1], Laust Hvas Mortensen[4,5], Lau Lilleholt[6,7], Anna Rogers[8], Ingo Zettler[6,7] & Sune Lehmann[1,7]

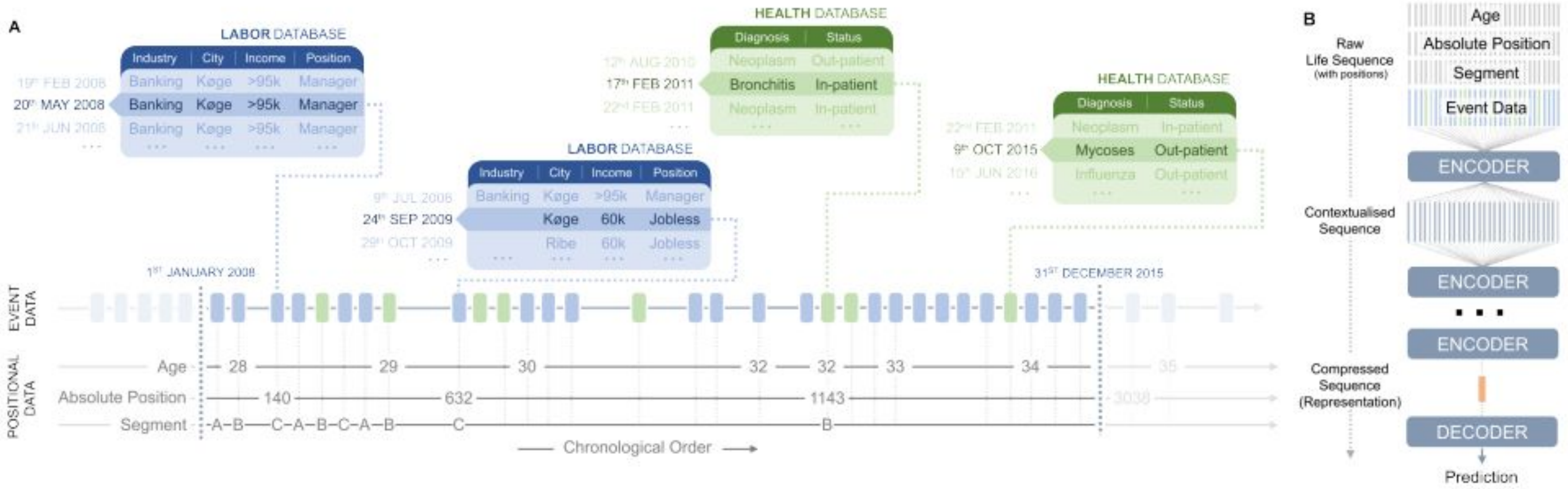
Here we represent human lives in a way that shares structural similarity to language, and we exploit this similarity to adapt natural language processing techniques to examine the evolution and predictability of human lives based on detailed event sequences. We do this by drawing on a comprehensive registry dataset, which is available for Denmark across several years, and that includes information about life-events related to health, education, occupation, income, address and working hours, recorded with day-to-day resolution. We create embeddings of life-events in a single vector space, showing that this embedding space is robust and highly structured. Our models allow us to predict diverse outcomes ranging from early mortality to personality nuances, outperforming state-of-the-art models by a wide margin. Using methods for interpreting deep learning models, we probe the algorithm to understand the factors that enable our predictions. Our framework allows researchers to discover potential mechanisms that impact life outcomes as well as the associated possibilities for personalized interventions.

We live in the age of algorithm-driven prediction of human behavior. The predictions range from those at the global and population level, with societies allocating vast resources to predicting phenomena such as global warming[1] or the spread of infectious diseases[2], all the way to the constant flow of individual micro-predictions that shape our reality and behavior as we use social media[3]. When it comes to individual life outcomes, however, the picture is more complex. Sociodemographic

decade interval, we show that accurate individual predictions are indeed possible. Our dataset includes a host of indicators, such as health, professional occupation and affiliation, income level, residency, working hours and education (Dataset section).

The main reason why we are currently experiencing this 'age of human prediction' is the advent of massive datasets and powerful machine learning algorithms[8,9]. Over the past decade, machine learning

Aplicaciones en Ciencias Sociales - Trayectorias



[CLS] [MALE] [YEAR=1957] [SEP] ... [MANAGER] [SEP] ... [FRACTURE] [IN-PATIENT] [SEP]

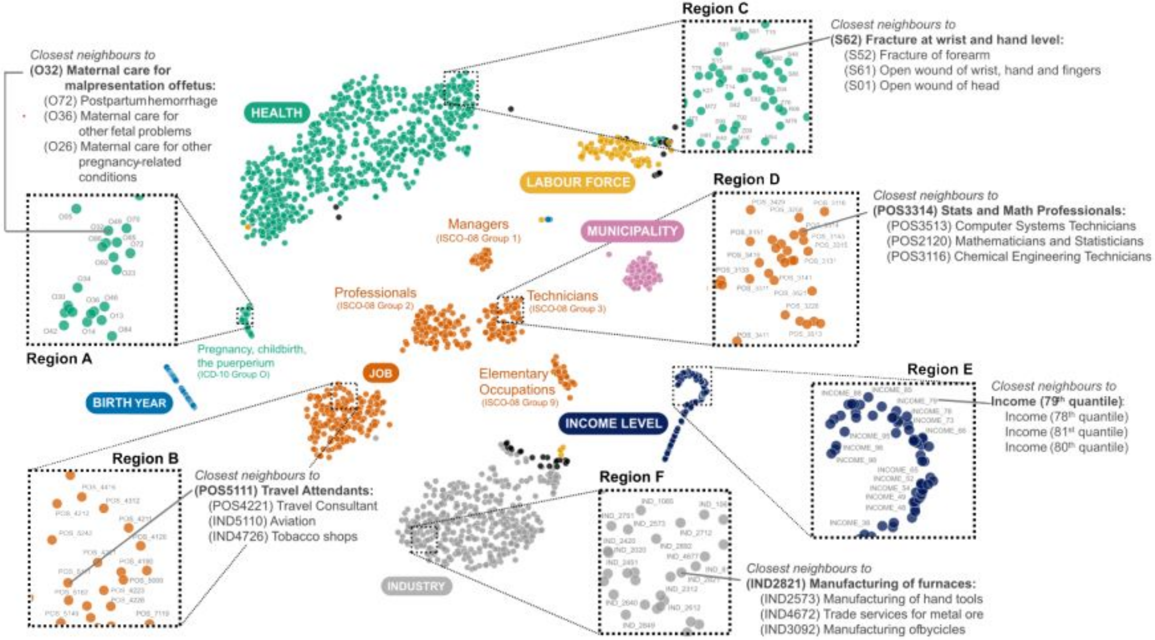


Figure 4: Two-dimensional projection of the concept space (using the PaCMAP [72]). Each point corresponds to a concept token in the vocabulary. Points are colored based on the concept types (several types are omitted - black points). Each region provides a closer look at several parts of the concept space. You can also see the top three closest neighbors for selected tokens (based on the cosine distance). (**A**) Diagnoses related to Pregnancy, childbirth, and the puerperium in ICD-10 [40]. (**B**) Job concepts related to Service and Sales Workers (corresponds to Job Category 5 of ISCO-08 [38]). (**C**) Injury-related diagnoses in ICD-10 [40]. (**D**) Job concepts related to Technicians and Associate Professionals (corresponds to Job Category 3 of ISCO-08 [38]). (**E**) Income-related concepts. `life2vec` arranges these concepts in increasing ordinal order. (**F**) Concepts related to the manufacturing industry in DB07 [39].

Aplicaciones en otras disciplinas

Google DeepMind 2025-7-31

AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data

Christopher F. Brown[*,1], Michal R. Kazmierski[*,1], Valerie J. Pasquarella[*,2], William J. Rucklidge[2], Masha Samsikova[1], Chenhui Zhang[1], Evan Shelhamer[1], Estefania Lahera[2], Olivia Wiles[1], Simon Ilyushchenko[2], Noel Gorelick[2], Lihui Lydia Zhang[1], Sophia Alj[1], Emily Schechter[2], Sean Askay[2], Oliver Guinan[2], Rebecca Moore[2], Alexis Boukouvalas[1] and Pushmeet Kohli[1]

[*]Equal contributions, [1]Google DeepMind, [2]Google

arXiv:2507.22291v1 [cs.CV] 29 Jul 2025

Unprecedented volumes of Earth observation data are continually collected around the world, but high-quality labels remain scarce given the effort required to make physical measurements and observations. This has led to considerable investment in bespoke modeling efforts translating sparse labels into maps. Here we introduce AlphaEarth Foundations, an embedding field model yielding a highly general, geospatial representation that assimilates spatial, temporal, and measurement contexts across multiple sources, enabling accurate and efficient production of maps and monitoring systems from local to global scales. The embeddings generated by AlphaEarth Foundations are the only to consistently outperform all previous featurization approaches tested on a diverse set of mapping evaluations without re-training. We will release a dataset of global, annual, analysis-ready embedding field layers from 2017 through 2024.

Introduction

Management of global food supplies, public health, and disaster response all start from maps that geographically anchor questions like "which forests pose an unacceptable wildfire risk?" or "where are soybeans grown?". The launch of the first Landsat satellite in 1972 marked the dawn of an era where spaceborne monitoring could serve the interests of global environmental policy-making and provide critical insights into our changing planet (Cohen and Goward, 2004). Over the following decades Earth observation (EO) data became widely available, and streams from both historic and modern EO instruments are now routinely used to create maps that answer questions about the past, present, and future of Earth's ecosystems and climate

embedding model that solves fundamental challenges in the institution of mapping through the generation of a universal feature space. The features produced by our model consistently achieve top performance in all application domains tested when compared to other general and even domain specific approaches (Figure 1A). This marks a shift from the previous state-of-the-art for which no single approach was dominant.

From sparse labels to maps

High-quality maps depend on high-quality labeled data, yet when working at global scales, a balance must be struck between measurement precision and spatial coverage. Many global mapping efforts focus on individual ecosystems like forests (Hansen et al., 2013), water (Pekel et al.,

Satellite Embedding V1

Dataset Availability

2017-01-01T00:00:00Z–2024-01-01T00:00:00Z

Dataset Provider

Google Earth Engine Google DeepMind

Earth Engine Snippet

```
ee.ImageCollection("GOOGLE/SATELLITE_EMBEDDING/V1/ANNUAL")
```

Description | Bands | Image Properties | Terms of Use

The Google Satellite Embedding dataset is a global, analysis-ready collection of learned geospatial embeddings. Each 10-meter pixel in this dataset is a 64-dimensional representation, or "embedding vector," that encodes temporal trajectories of surface conditions at and around that pixel as measured by various Earth observation instruments and datasets, over a single calendar year. Unlike conventional spectral inputs and indices, where bands reflect physical measurements, embeddings are feature vectors that summarize relationships across multi-source, multi-modal observations in a less directly interpretable, but more powerful way.

Aplicaciones en otras disciplinas

Google DeepMind 2025-7-31

AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data

Christopher F. Brown[*,1], Michal R. Kazmierski[*,1], Valerie J. Pasquarella[*,2], William J. Rucklidge[2], Masha Samsikova[1], Chenhui Zhang[1], Evan Shelhamer[1], Estefania Lahera[2], Olivia Wiles[1], Simon Ilyushchenko[2], Noel Gorelick[2], Lihui Lydia Zhang[1], Sophia Alj[1], Emily Schechter[2], Sean Askay[2], Oliver Guinan[2], Rebecca Moore[2], Alexis Boukouvalas[1] and Pushmeet Kohli[1]
[*]Equal contributions, [1]Google DeepMind, [2]Google

Unprecedented volumes of Earth observation data are continually collected around the world, but high-quality labels remain scarce given the effort required to make physical measurements and observations. This has led to considerable investment in bespoke modeling efforts translating sparse labels into maps. Here we introduce AlphaEarth Foundations, an embedding field model yielding a highly general, geospatial representation that assimilates spatial, temporal, and measurement contexts across multiple sources, enabling accurate and efficient production of maps and monitoring systems from local to global scales. The embeddings generated by AlphaEarth Foundations are the only to consistently outperform all previous featurization approaches tested on a diverse set of mapping evaluations without re-training. We will release a dataset of global, annual, analysis-ready embedding field layers from 2017 through 2024.

Introduction

Management of global food supplies, public health, and disaster response all start from maps that geographically anchor questions like "which forests pose an unacceptable wildfire risk?" or "where are soybeans grown?". The launch of the first Landsat satellite in 1972 marked the dawn of an era where spaceborne monitoring could serve the interests of global environmental policy-making and provide critical insights into our changing planet (Cohen and Goward, 2004). Over the following decades Earth observation (EO) data became widely available, and streams from both historic and modern EO instruments are now routinely used to create maps that answer questions about the past, present, and future of Earth's ecosystems and climate

embedding model that solves fundamental challenges in the institution of mapping through the generation of a universal feature space. The features produced by our model consistently achieve top performance in all application domains tested when compared to other general and even domain specific approaches (Figure 1A). This marks a shift from the previous state-of-the-art for which no single approach was dominant.

From sparse labels to maps

High-quality maps depend on high-quality labeled data, yet when working at global scales, a balance must be struck between measurement precision and spatial coverage. Many global mapping efforts focus on individual ecosystems like forests (Hansen et al., 2013), water (Pekel et al.,

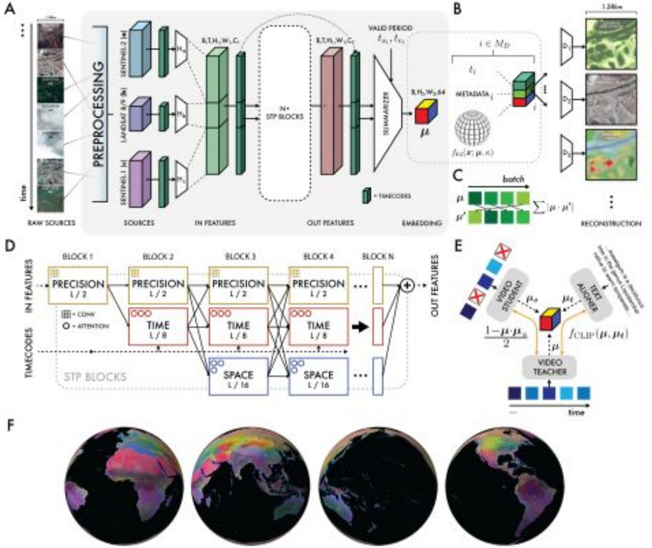


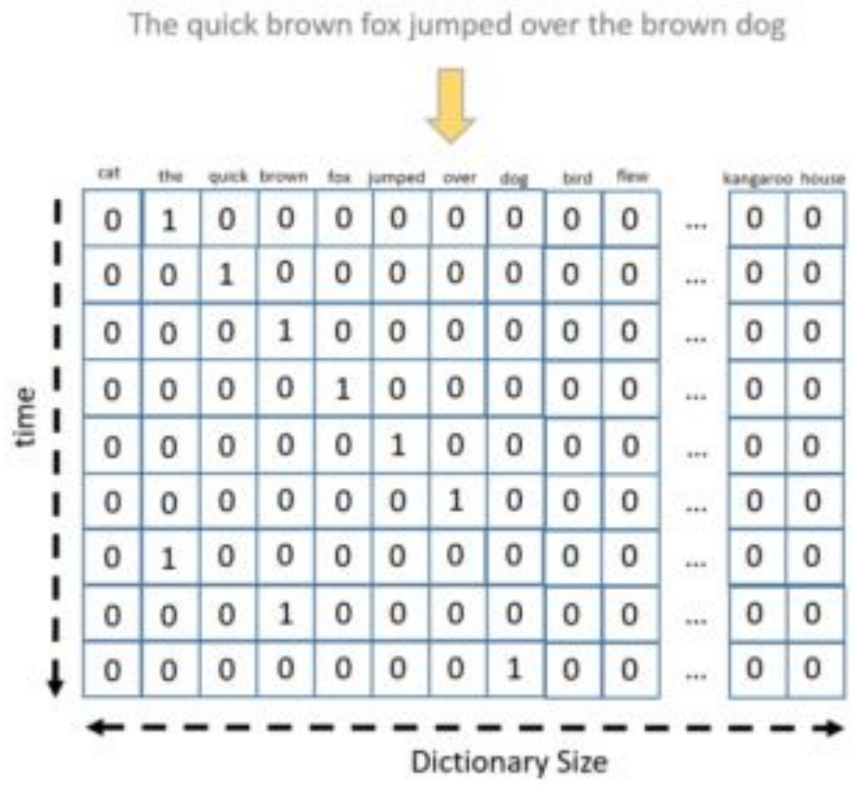
Figure 2 | **AlphaEarth Foundations.** (A) Block diagram of the overall network architecture used for video analysis. Preprocessing converts raw observation data via normalization using global statistics, and acquisition timestamps are converted to sinusoidal timecodes. Individual source encoders transform inputs to the same latent space before entering the bulk of the model. Outputs are summarized using conditional timecodes or "summary periods", unique to each decoded source and contrastive learning task. μ refers to the embedding outputs of the model. (B) Model outputs are treated as the mean direction of a von Mises-Fisher distribution, and decoding proceeds by sampling this distribution, and concatenating it with sensor geometry metadata and a timecode indicating the relative position in the valid period to decode. Decoding proceeds for all sources, with losses dependent on the characteristics of each source (see supplemental materials S1). (C) To prevent collapse and improve performance, embeddings are compared to equivalent batch-rotated embeddings using a dot product. The absolute value of this quantity is minimized as a necessary condition for an empirically uniform distribution in S^{63}. (D) Block diagram of the model bulk, consisting of simultaneous pathways at different resolutions to maintain efficiency and spatial precision. (E) Contrastive learning between the video teacher and student model, and text encoder. (F) Complete 360° view of 2023 annual embedding field covering Earth's land surface including minor islands over approximately ± 82°.

factor~data
EIDAES_UNSAM

¿Cómo sucede la magia?

One hot encoding

- Eje Y = tiempo
- Eje X = vocabulario
- Celdas: 1 si la palabra aparece en ese “momento”; 0 si no aparece



cat	the	quick	brown	fox	jumped	over	dog	bird	flew		kangaroo	house
0	1	0	0	0	0	0	0	0	0	...	0	0
0	0	1	0	0	0	0	0	0	0	...	0	0
0	0	0	1	0	0	0	0	0	0	...	0	0
0	0	0	0	1	0	0	0	0	0	...	0	0
0	0	0	0	0	1	0	0	0	0	...	0	0
0	0	0	0	0	0	1	0	0	0	...	0	0
0	1	0	0	0	0	0	0	0	0	...	0	0
0	0	0	1	0	0	0	0	0	0	...	0	0
0	0	0	0	0	0	0	1	0	0	...	0	0

Skip-gram

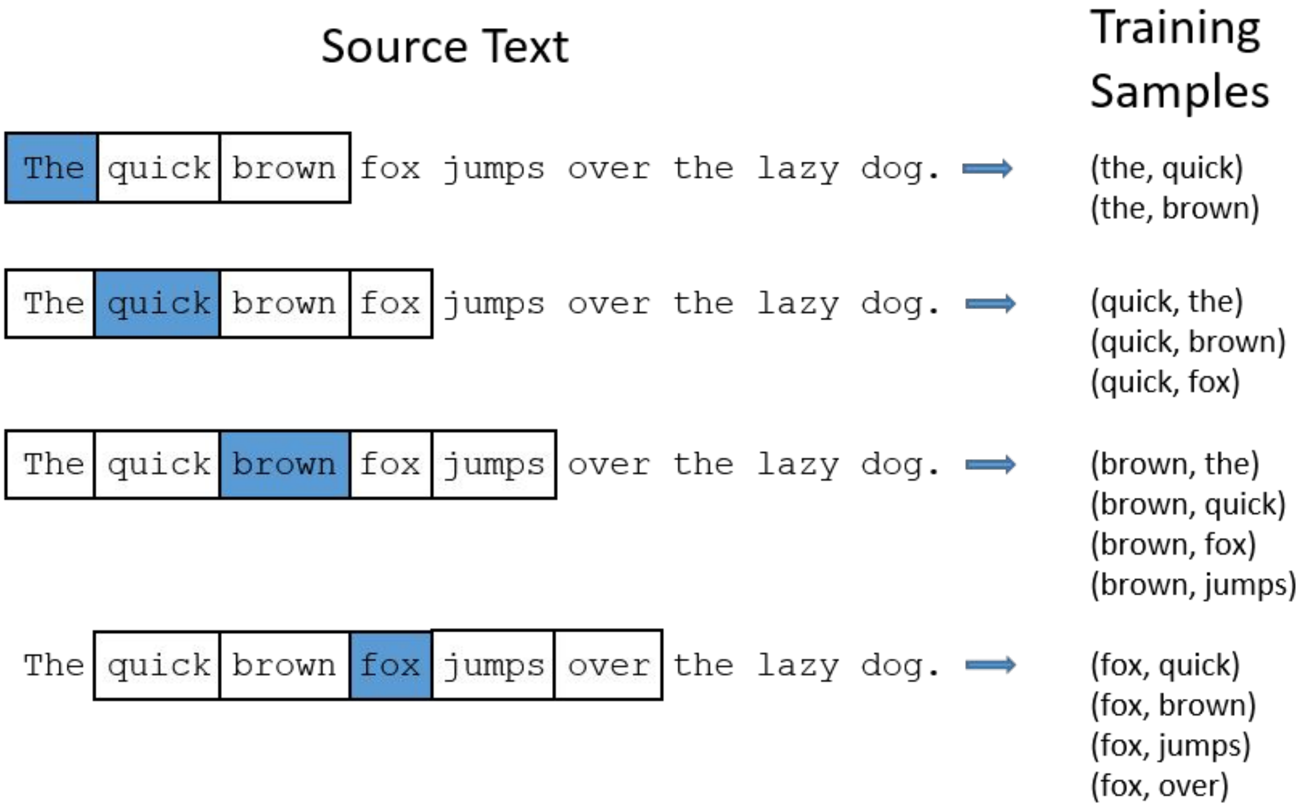
Cambia la unidad

Ahora el corpus es visto como un todo continuo…

No se ven los documentos por separado

Un parámetro importante: el tamaño de la ventana…

Otro metodo: CBOW (al revés)



Skip-gram

Contexte				Mot Cible
The	Quick	Fox	Jump	**Brown**
Quick	Brown	Jumps	Over	**Fox**
Brown	Fox	Over	The	**Jumps**

Skip-gram - Matriz de co-ocurrencias

	brown	dog	fox	jumps	lazy	over	quick	the
brown	0	0	0	0	0	0	1	1
dog	0	0	0	0	1	0	0	1
fox	1	0	0	0	0	0	1	0
jumps	1	0	1	0	0	0	0	0
lazy	0	0	0	0	0	1	0	1
over	0	0	1	1	0	0	0	0
quick	0	0	0	0	0	0	0	1
the	0	0	0	1	0	1	0	0

Skip-gram (otro ejemplo)

Thou shalt not make a machine in the likeness of a human mind

thou	shalt	not	make	a	machine	in	the	…

input word	target word
not	thou
not	shalt
not	make
not	a

Cambia la unidad

Ahora el corpus es visto como un todo continuo…

No se ven los documentos por separado

Un parámetro importante: el tamaño de la ventana…

Otro metodo: CBOW (al revés)

Fuente: https://jalammar.github.io/illustrated-word2vec/

Skip-gram (otro ejemplo)

Thou shalt not make a machine in the likeness of a human mind

thou	shalt	not	make	a	machine	in	the	…

thou	shalt	not	make	a	machine	in	the	…

input word	target word
not	thou
not	shalt
not	make
not	a
make	shalt
make	not
make	a
make	machine

Fuente: https://jalammar.github.io/illustrated-word2vec/

Skip-gram (otro ejemplo)

Thou shalt not make a machine in the likeness of a human mind

thou	shalt	not	make	a	machine	in	the	…
thou	shalt	not	make	a	machine	in	the	…
thou	shalt	not	make	a	machine	in	the	…
thou	shalt	not	make	a	machine	in	the	…
thou	shalt	not	make	a	machine	in	the	…

input word	target word
not	thou
not	shalt
not	make
not	a
make	shalt
make	not
make	a
make	machine
a	not
a	make
a	machine
a	in
machine	make
machine	a
machine	in
machine	the
in	a
in	machine
in	the
in	likeness

Fuente: https://jalammar.github.io/illustrated-word2vec/

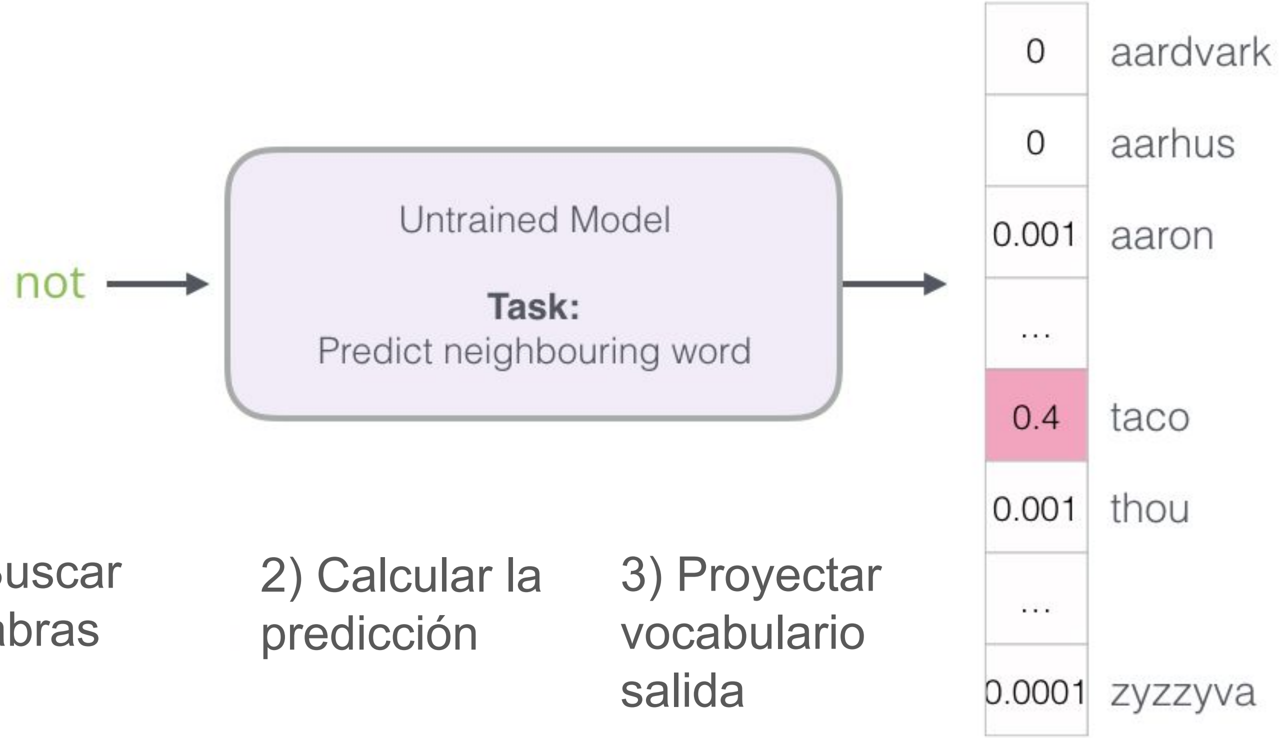
Modelando con skipgram

input word	target word
not	thou
not	shalt
not	make
not	a
make	shalt
make	not
make	a
make	machine
a	not
a	make
a	machine
a	in
machine	make
machine	a
machine	in
machine	the
in	a
in	machine
in	the
in	likeness

factor~data
EIDAES_UNSAM

Fuente: https://jalammar.github.io/illustrated-word2vec/

Modelando con skipgram



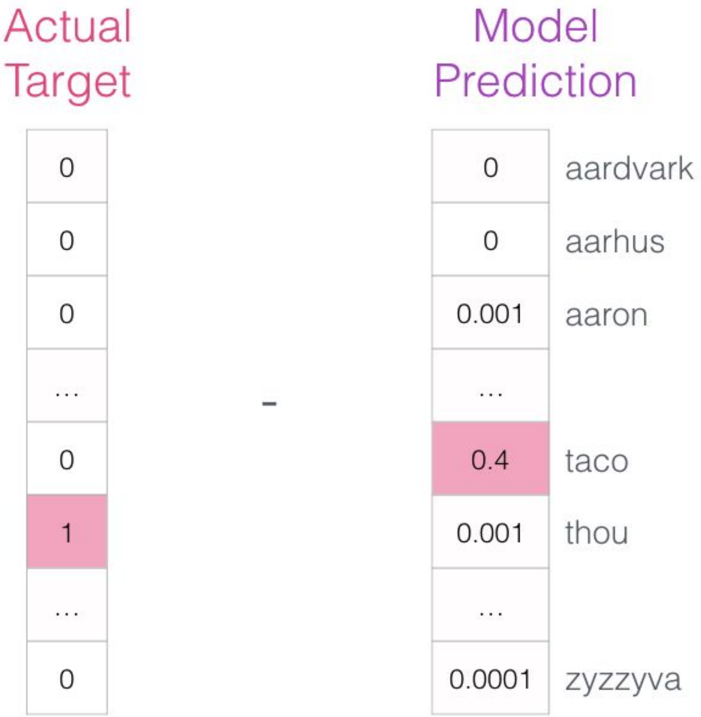
1) Buscar palabras

2) Calcular la predicción

3) Proyectar vocabulario salida

Fuente: https://jalammar.github.io/illustrated-word2vec/

Modelando con skipgram



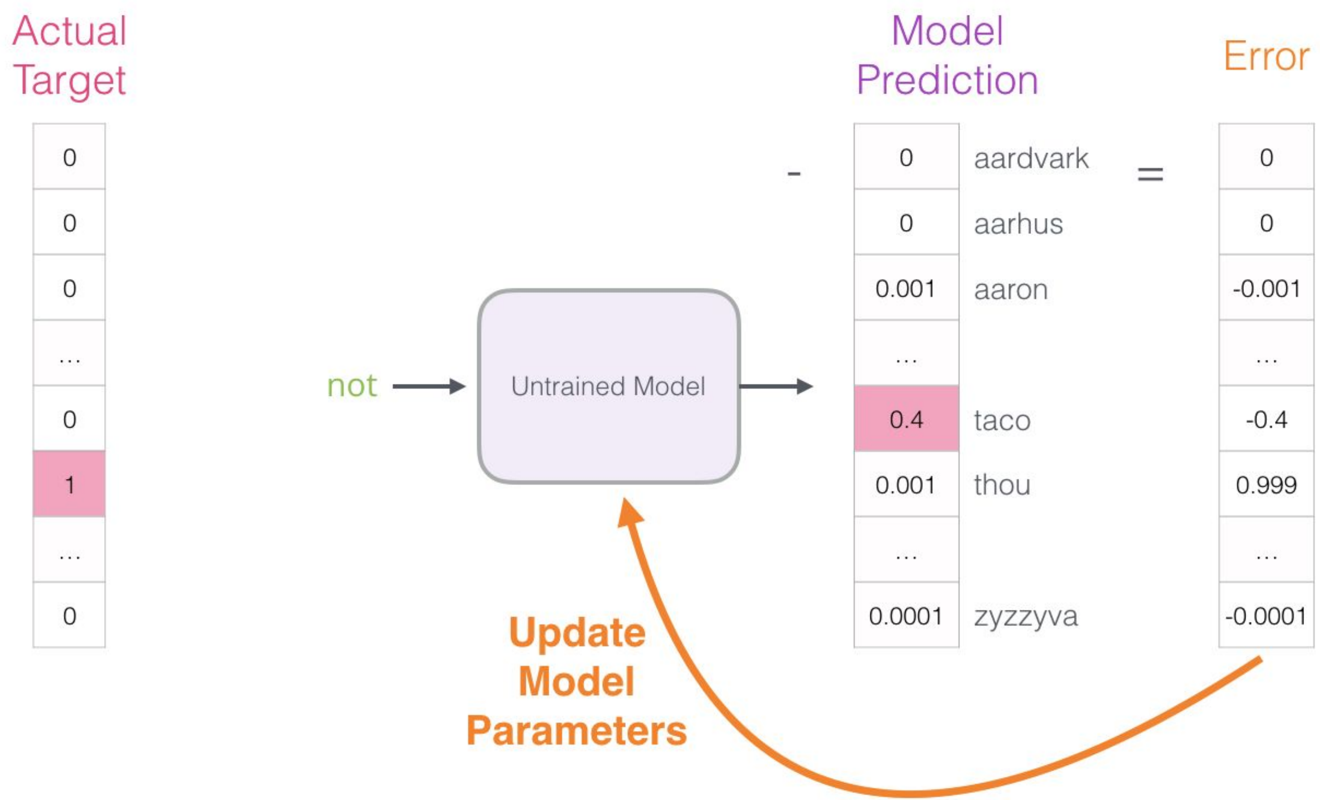
Fuente: https://jalammar.github.io/illustrated-word2vec/

Modelando con skipgram

Actual Target		Model Prediction			Error
0		0	aardvark		0
0		0	aarhus		0
0		0.001	aaron		-0.001
…	-	…		=	…
0		0.4	taco		-0.4
1		0.001	thou		0.999
…		…			…
0		0.0001	zyzzyva		-0.0001

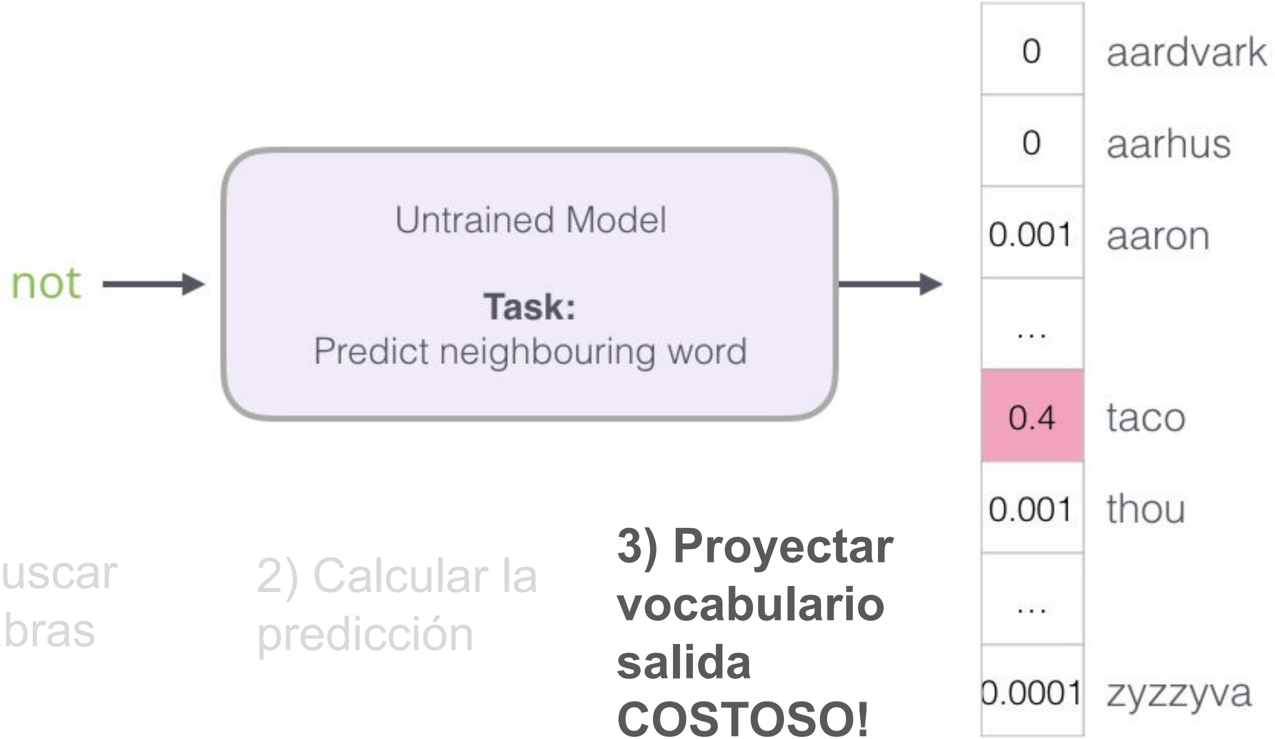
Fuente: https://jalammar.github.io/illustrated-word2vec/

Modelando con skipgram



factor~data
EIDAES_UNSAM

Modelando con skipgram => PROBLEMA



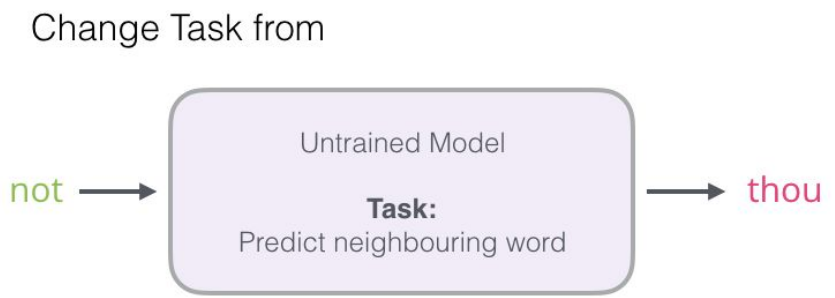
1) Buscar palabras

2) Calcular la predicción

3) Proyectar vocabulario salida COSTOSO!

Fuente: https://jalammar.github.io/illustrated-word2vec/

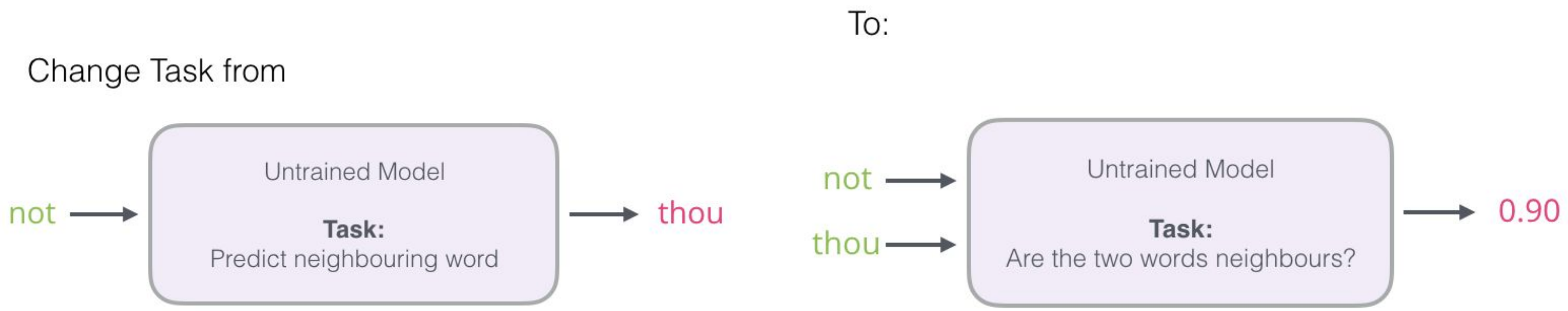
Modelando con skipgram => PROBLEMA



factor~data
EIDAES_UNSAM

Fuente: https://jalammar.github.io/illustrated-word2vec/

Modelando con skipgram => PROBLEMA

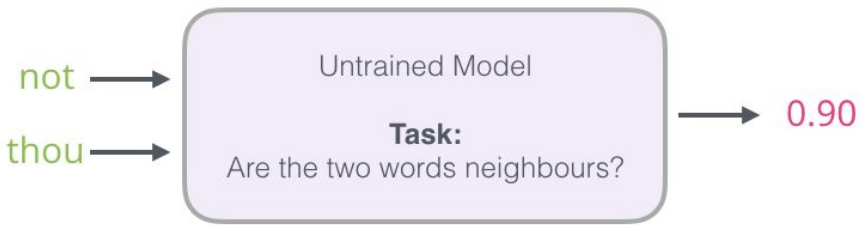


Fuente: https://jalammar.github.io/illustrated-word2vec/

Modelando con skipgram => PROBLEMA

Change Task from

To:



input word	target word
not	thou
not	shalt
not	make
not	a
make	shalt
make	not
make	a
make	machine

input word	output word	target
not	thou	1
not	shalt	1
not	make	1
not	a	1
make	shalt	1
make	not	1
make	a	1
make	machine	1

Problema!
Todos ejemplos positivos…

OVERFITTING

Fuente: https://jalammar.github.io/illustrated-word2vec/

Negative sampling

input word	output word	target
not	thou	1
not		0
not		0
not	shalt	1
not	make	1

Negative examples

Fuente: https://jalammar.github.io/illustrated-word2vec/

Negative sampling

Pick randomly from vocabulary
(random sampling)

input word	output word	target
not	thou	1
not	aaron	0
not	taco	0
not	shalt	1
not	make	1

Word	Count	Probability
aardvark		
aarhus		
aaron		
taco		
thou		
zyzzyva		

Fuente: https://jalammar.github.io/illustrated-word2vec/

La fórmula mágina de w2vec

Skipgram

shalt	not	make	a	machine

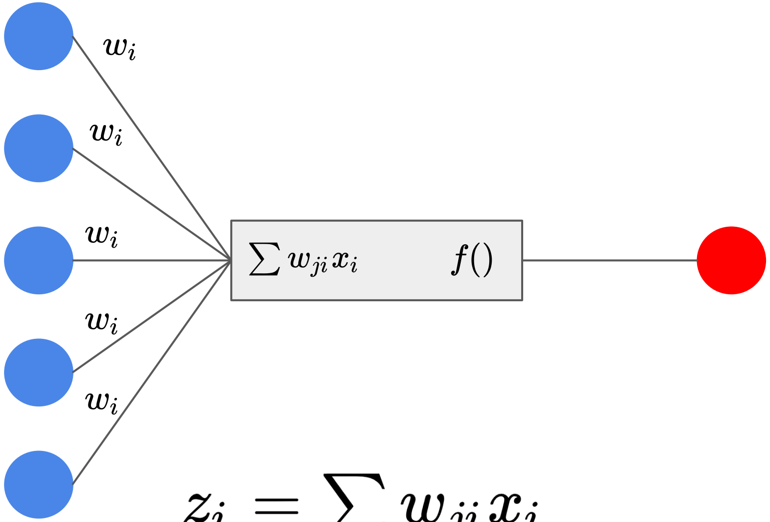
input	output
make	shalt
make	not
make	a
make	machine

Negative Sampling

input word	output word	target
make	shalt	1
make	aaron	0
make	taco	0

Fuente: https://jalammar.github.io/illustrated-word2vec/

Regresión logística en forma de red neuronal

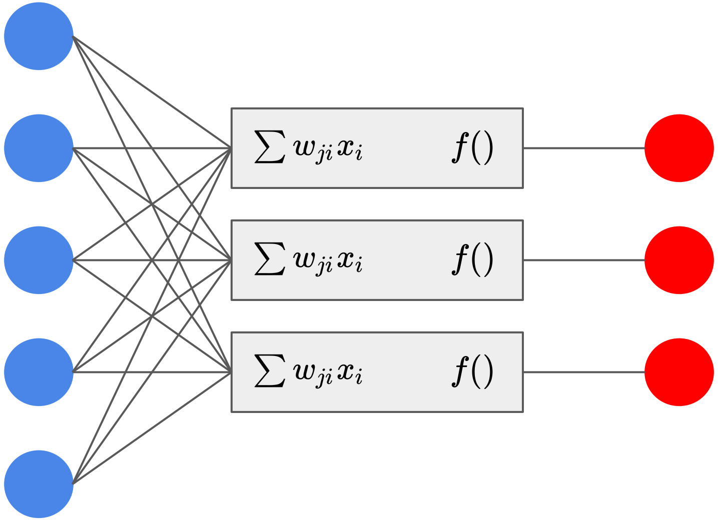


$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_{\acute{\imath}}}}$$

$$z_i = \sum w_{ji} x_i$$

$$x_i$$

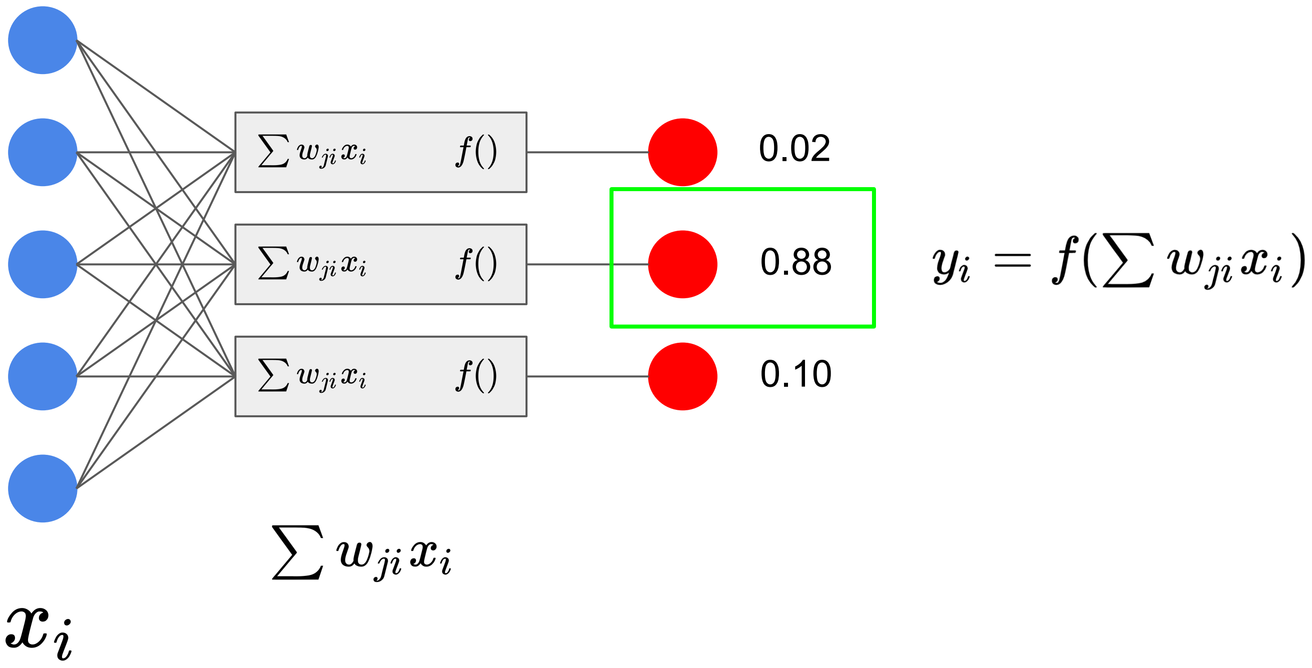
Redes neuronales (intuición)



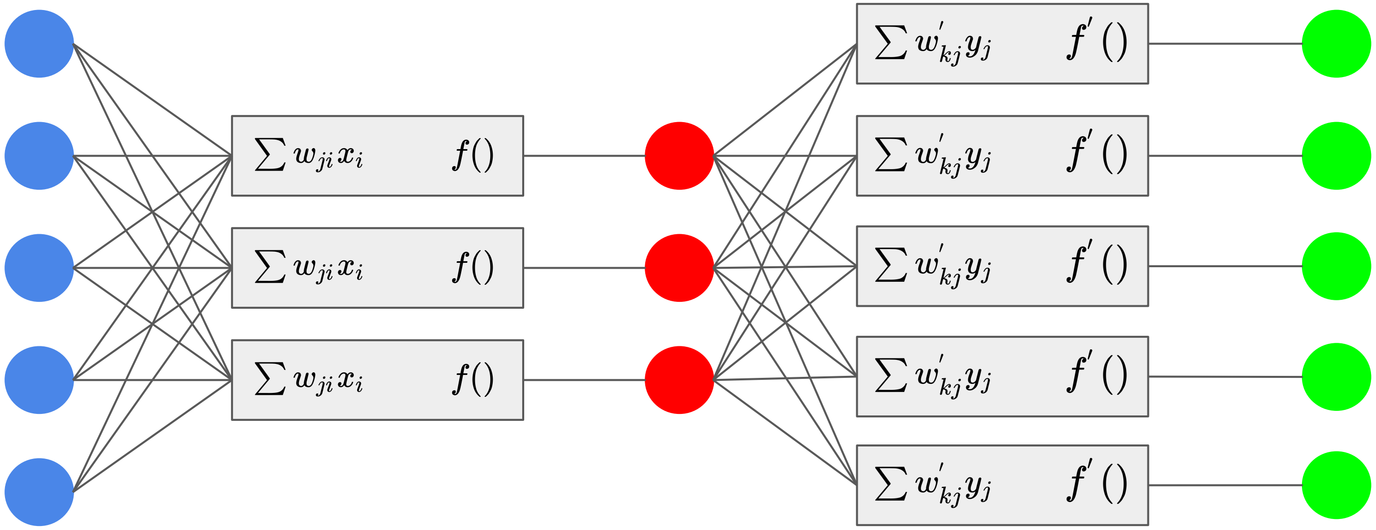
$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_i}}$$

$$z_i = \sum w_{ji} x_i$$

Redes neuronales (intuición)



Ahora sí… word2vec



$\sum w_{ji}x_i$

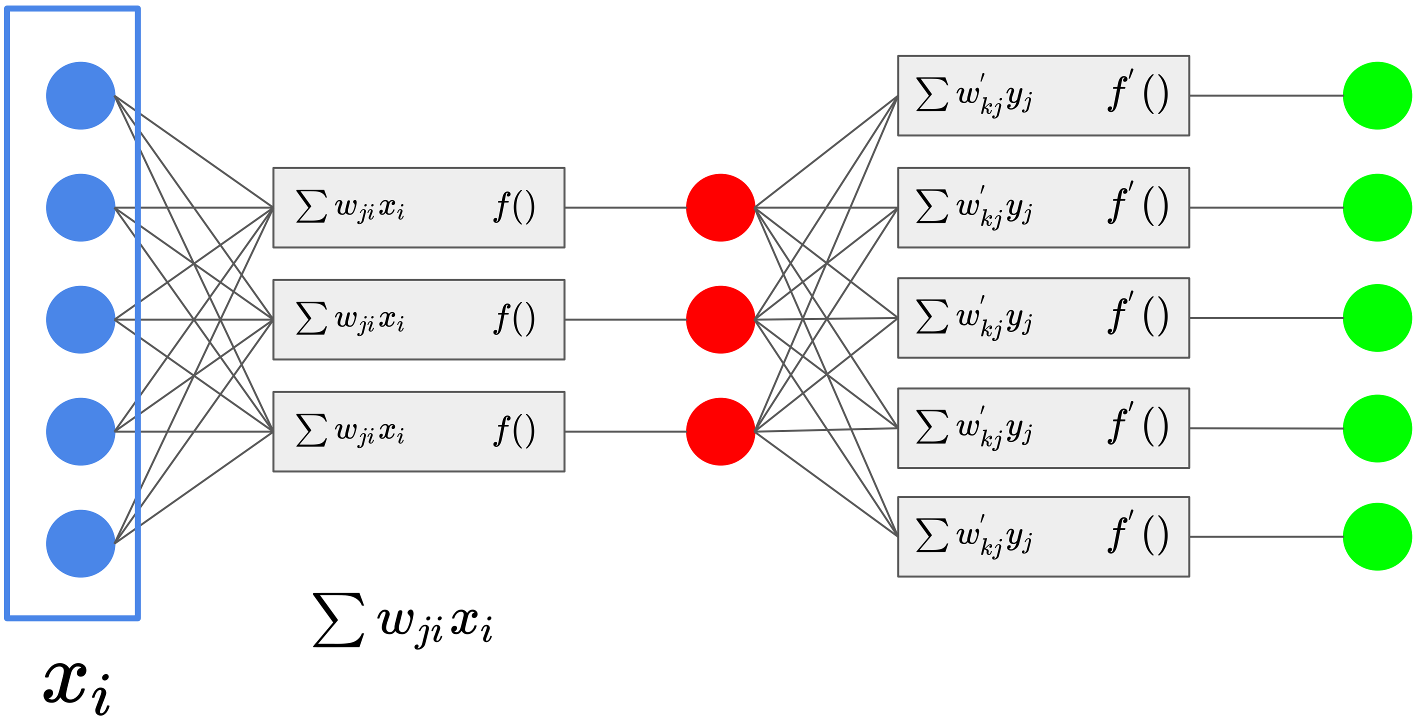
x_i

$$y_i = f(\sum w_{ji}x_i) \qquad z_k = f^{'}(\sum w^{'}_{kj}y_j)$$

Ahora sí… word2vec

Una “unidad” por palabra en el vocabulario => One hot encoded

1 x 5



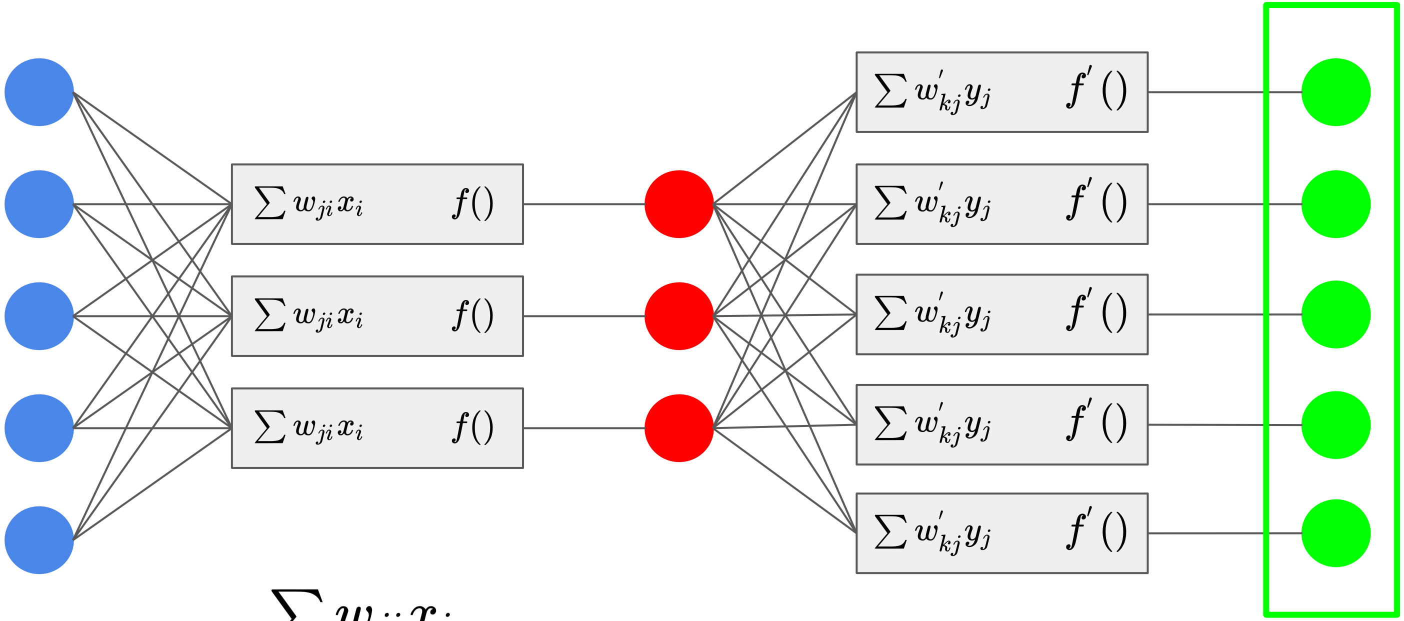
$$y_i = f(\sum w_{ji} x_i)$$

$$z_k = f'(\sum w'_{kj} y_j)$$

Ahora sí… word2vec

Una “unidad” por palabra en el vocabulario => One hot encoded 1 x 5

Una “unidad” por palabra en el vocabulario => One hot encoded



$\sum w_{ji}x_i$

x_i

$$y_i = f(\sum w_{ji}x_i)$$

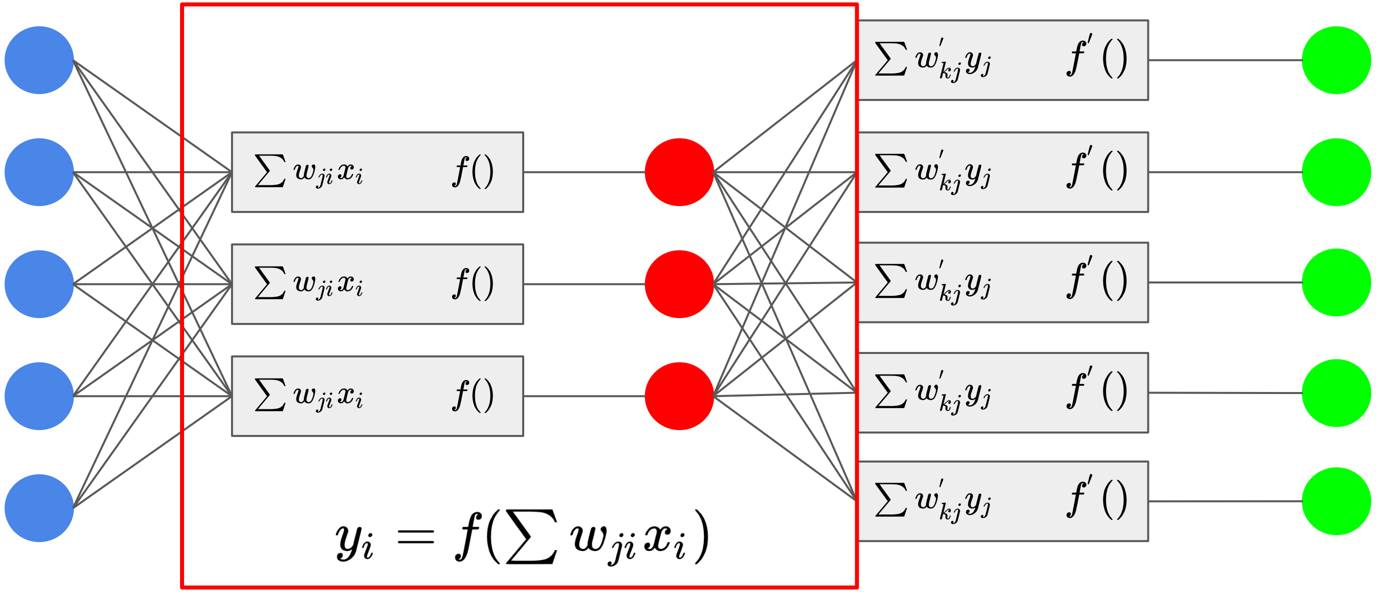
$$z_k = f'(\sum w'_{kj}y_j)$$

factor~data
EIDAES_UNSAM

Ahora sí… word2vec

Este es el **embedding**. Es la representación de baja dimensionalidad de una palabra
1 x 3

Una “unidad” por palabra en el vocabulario => One hot encoded



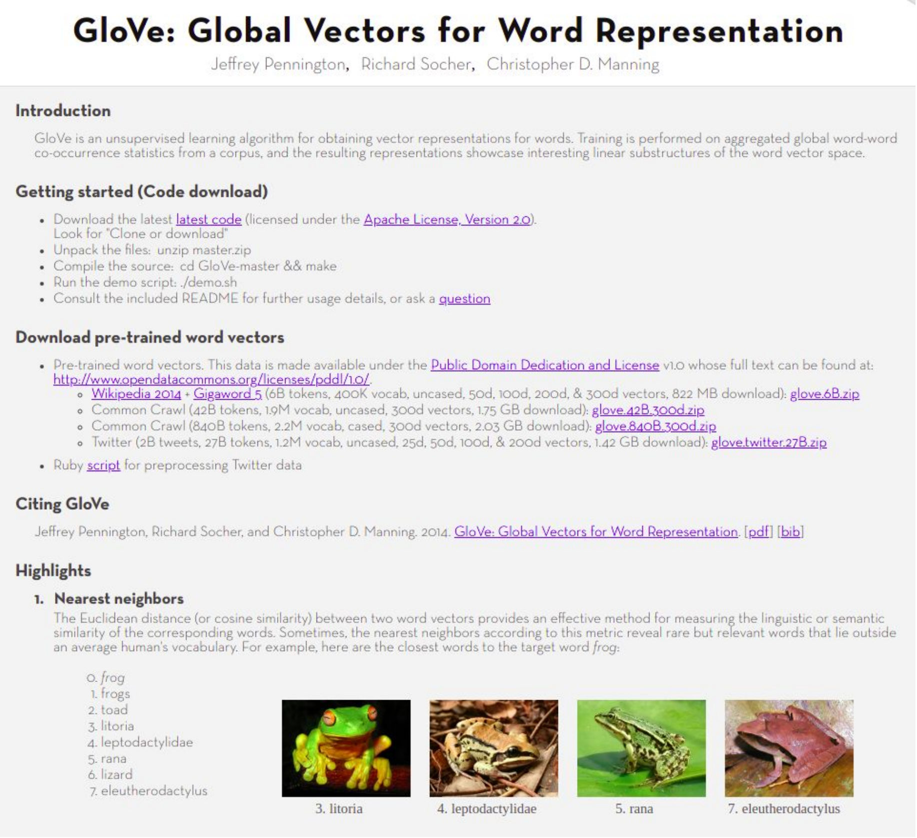
x_i

$$z_k = f'(\sum w'_{kj} y_j)$$

Otros métodos para construir embeddings

- word2vec fue pionero (2013) pero hoy hay métodos mejores

- GloVe: trabaja directamente sobre la matriz de co-ocurrencias



GloVe: Global Vectors for Word Representation

Jeffrey Pennington, Richard Socher, Christopher D. Manning

Introduction

GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.

Getting started (Code download)

- Download the latest latest code (licensed under the Apache License, Version 2.0). Look for "Clone or download"
- Unpack the files: unzip master.zip
- Compile the source: cd GloVe-master && make
- Run the demo script: ./demo.sh
- Consult the included README for further usage details, or ask a question

Download pre-trained word vectors

- Pre-trained word vectors. This data is made available under the Public Domain Dedication and License v1.0 whose full text can be found at: http://www.opendatacommons.org/licenses/pddl/1.0/.
  - Wikipedia 2014 + Gigaword 5 (6B tokens, 400K vocab, uncased, 50d, 100d, 200d, & 300d vectors, 822 MB download): glove.6B.zip
  - Common Crawl (42B tokens, 1.9M vocab, uncased, 300d vectors, 1.75 GB download): glove.42B.300d.zip
  - Common Crawl (840B tokens, 2.2M vocab, cased, 300d vectors, 2.03 GB download): glove.840B.300d.zip
  - Twitter (2B tweets, 27B tokens, 1.2M vocab, uncased, 25d, 50d, 100d, & 200d vectors, 1.42 GB download): glove.twitter.27B.zip
- Ruby script for preprocessing Twitter data

Citing GloVe

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. [pdf] [bib]

Highlights

1. **Nearest neighbors**

The Euclidean distance (or cosine similarity) between two word vectors provides an effective method for measuring the linguistic or semantic similarity of the corresponding words. Sometimes, the nearest neighbors according to this metric reveal rare but relevant words that lie outside an average human's vocabulary. For example, here are the closest words to the target word *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

factor~data
EIDAES_UNSAM