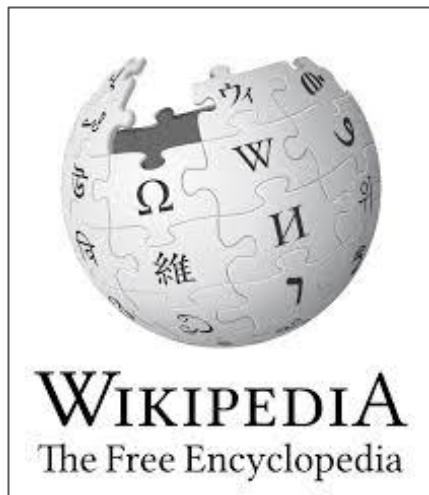
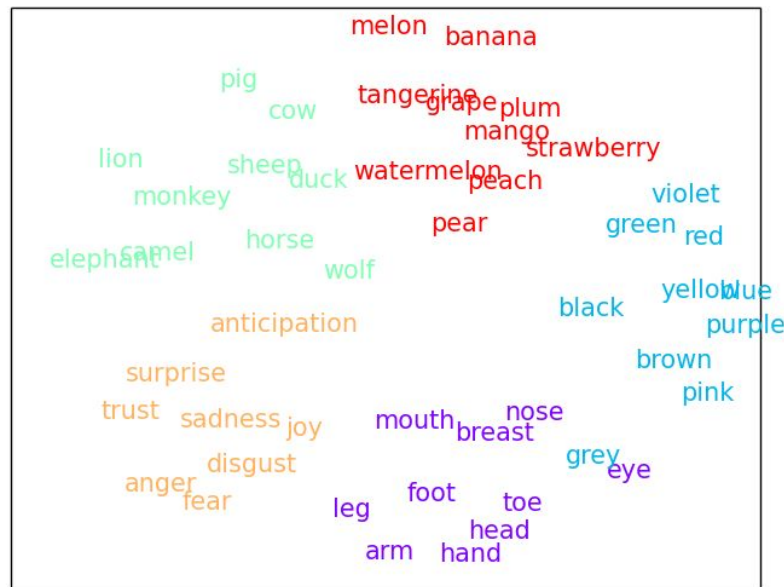


Word-embeddings

Corpus de textos



Word-embeddings



Latent Semantic Analysis

Ejemplo

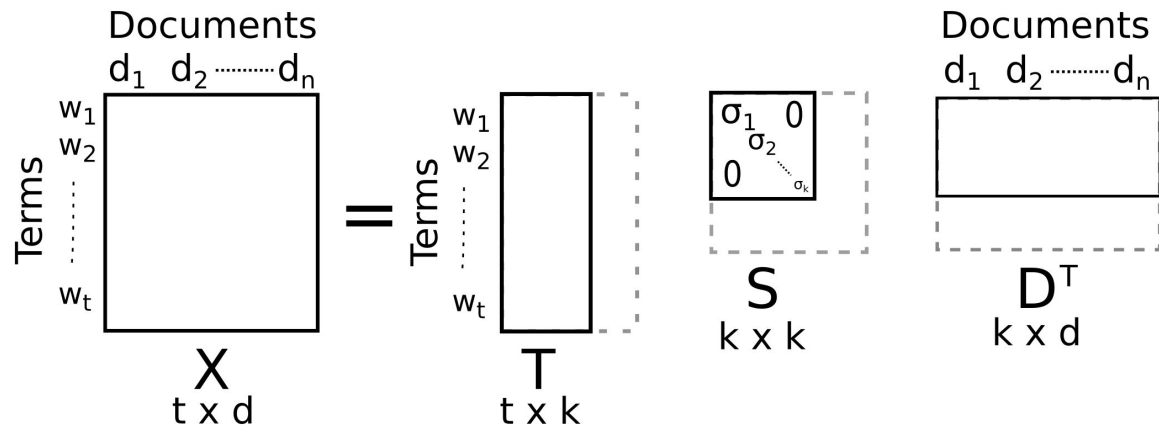
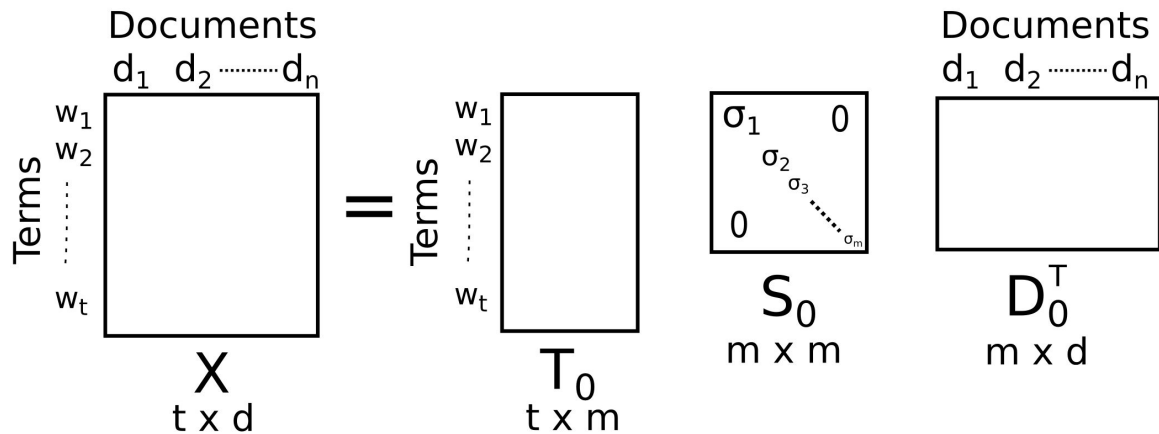
- | | |
|-------------------------------|--|
| Human-computer
interaction | <ul style="list-style-type: none">● c1: Human machine interface for ABC computer applications.● c2: A survey of user opinion of computer system response time.● c3: The EPS user interface management system.● c4: System and human system engineering testing of EPS.● c5: Relation of user perceived response time to error measurement. |
| Graphs theory | <ul style="list-style-type: none">● m1: The generation of random, binary, ordered trees.● m2: The intersection graph of paths in trees.● m3: Graph minors IV: Widths of trees and well-quasi-ordering.● m4: Graph minors: A survey. |

Term-Documents Matrix

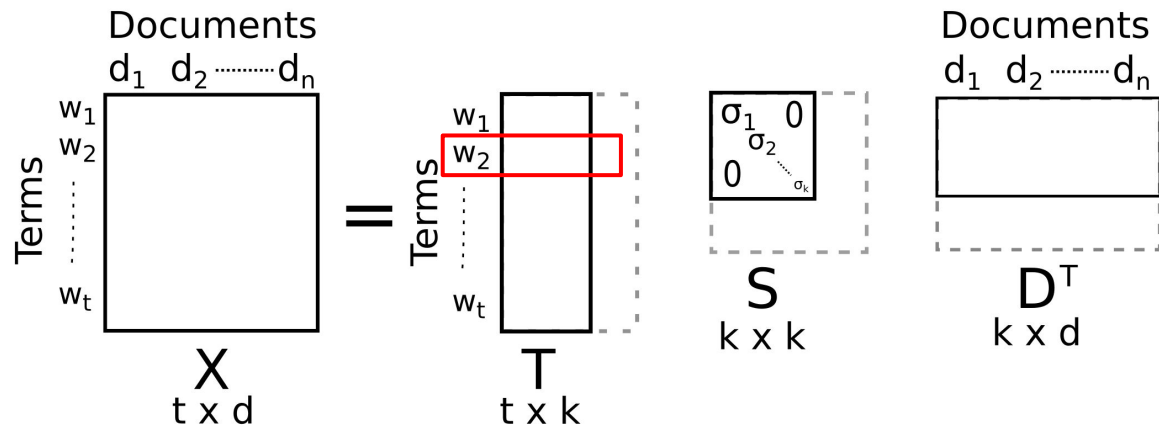
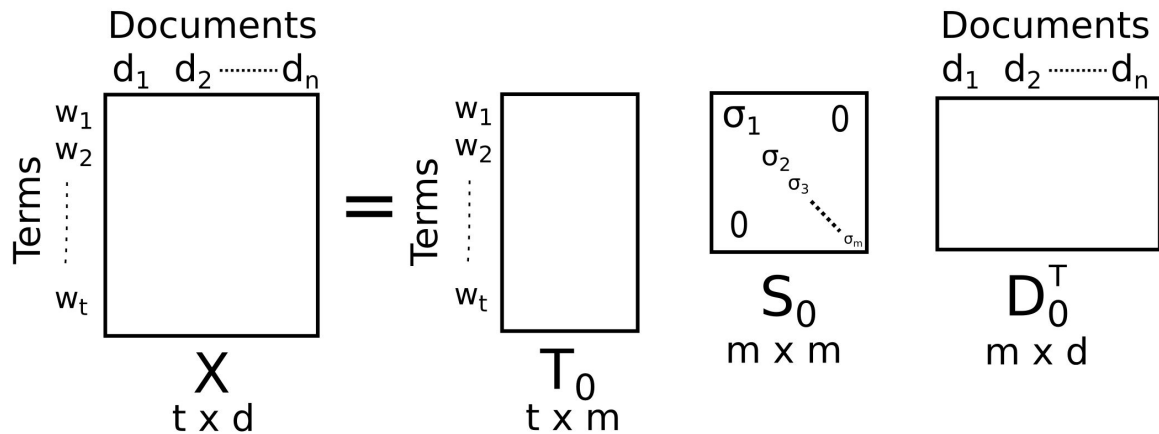
$$X =$$

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

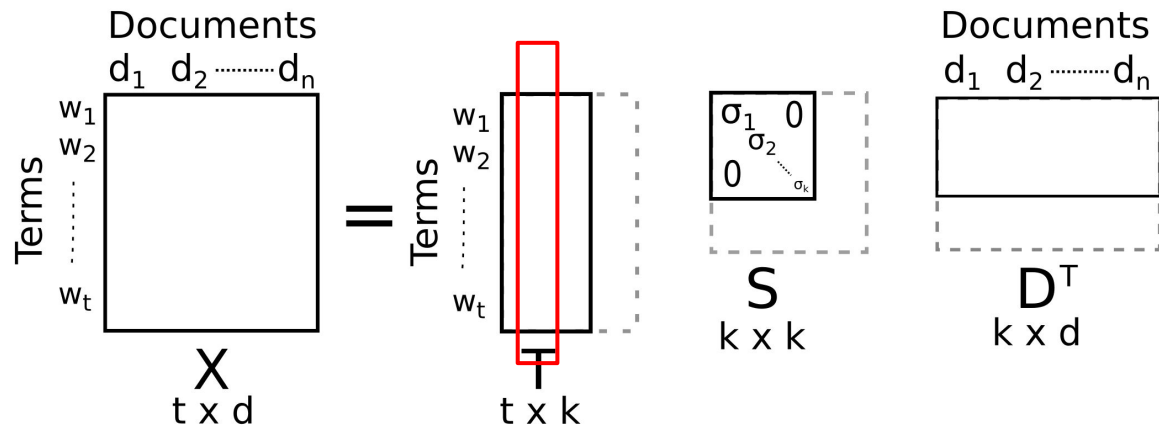
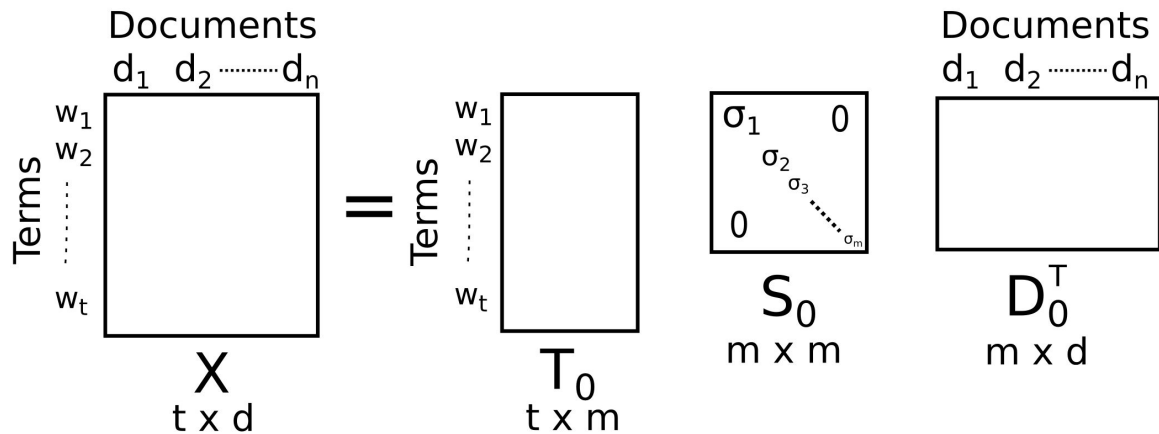
SVD



SVD



SVD



[illegible]

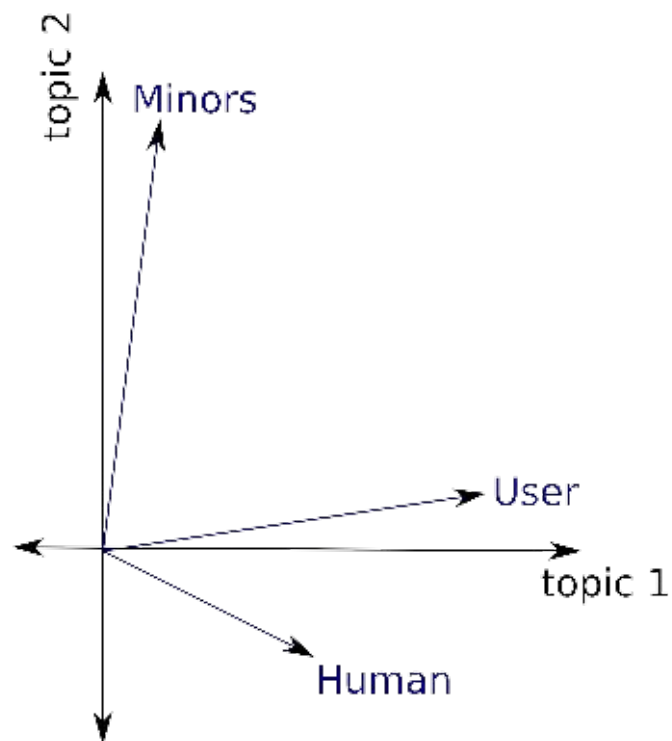
	W	S	C
	0.22 -0.11	3.34 0.00	0.20 0.60 0.46 0.54 0.28 0.00 0.02 0.02 0.08
	0.20 -0.07	0.00 2.54	-0.06 0.17 -0.13 0.53 -0.23 0.11 0.19 0.44 0.62
	0.24 0.04		
user	0.40 0.06		
	0.64 -0.17		
	0.27 0.11		
=	0.27 0.11		
	0.30 -0.14		
	0.21 0.27		
	0.01 0.49		
	0.04 0.62		
	0.03 0.45		

W

S

C

Human	→	0.22 -0.11	3.34 0.00	0.20 0.60 0.46 0.54 0.28 0.00 0.02 0.02 0.08
		0.20 -0.07	0.00 2.54	-0.06 0.17 -0.13 0.53 -0.23 0.11 0.19 0.44 0.62
		0.24 0.04		
User	→	0.40 0.06		
		0.64 -0.17		
		0.27 0.11		
		0.27 0.11		
		0.30 -0.14		
		0.21 0.27		
		0.01 0.49		
		0.04 0.62		
Minors	→	0.03 0.45		

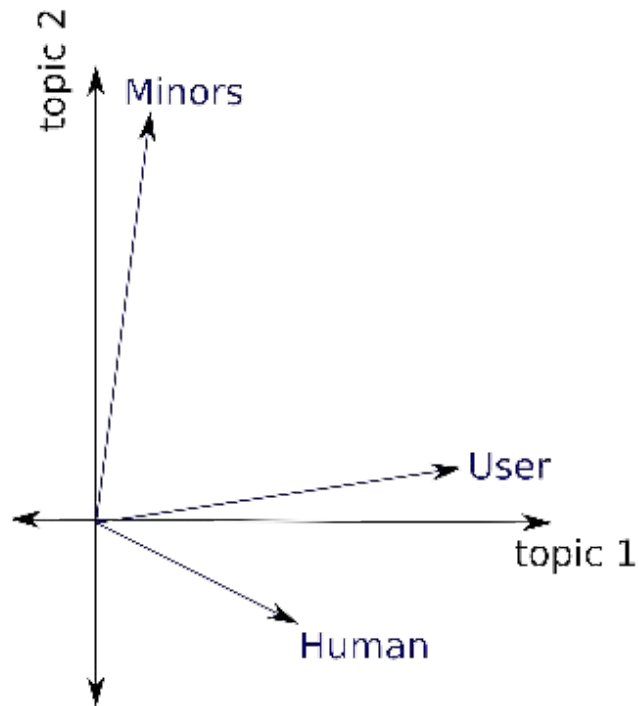


	W	S	C
Human →	0.22 -0.11 0.20 -0.07 0.24 0.04	3.34 0.00 0.00 2.54	0.20 0.60 0.46 0.54 0.28 0.00 0.02 0.02 0.08 -0.06 0.17 -0.13 0.53 -0.23 0.11 0.19 0.44 0.62
User →	0.40 0.06 0.64 -0.17 0.27 0.11 0.27 0.11 0.30 -0.14 0.21 0.27 0.01 0.49 0.04 0.62		
Minors →	0.03 0.45		

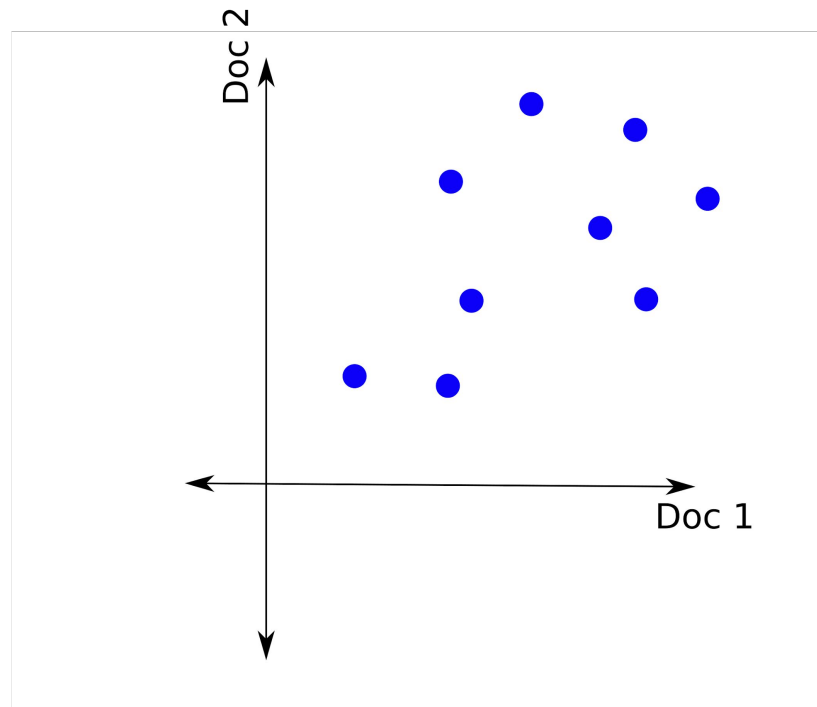
Cosine similarity

$$\text{cossim}(\bar{v}_1, \bar{v}_2) = \cos(\text{angle}) = \frac{\bar{v}_1 \cdot \bar{v}_2}{|\bar{v}_1| \cdot |\bar{v}_2|}$$

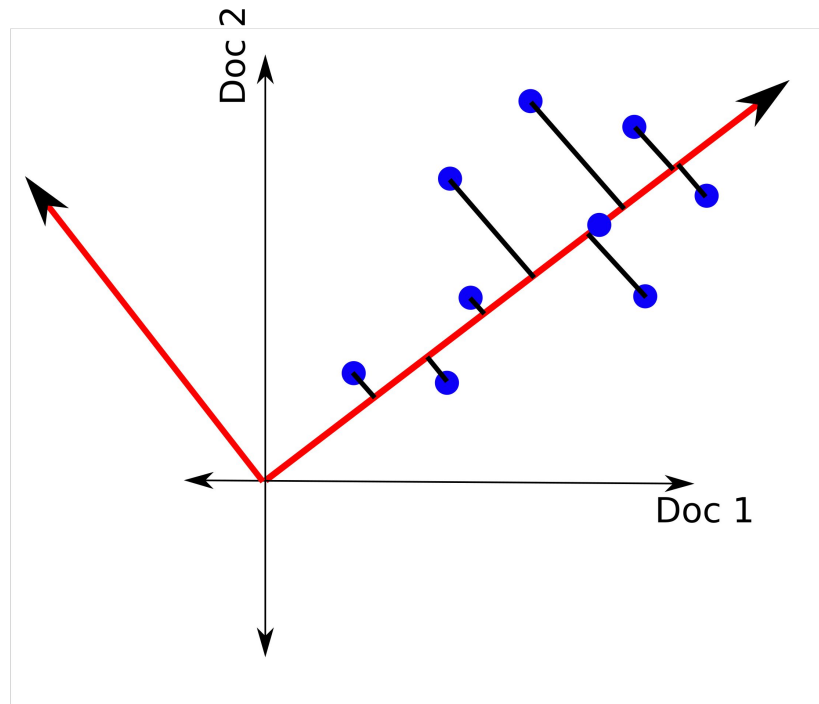
$$\text{cossim}(\bar{v}_1, \bar{v}_2) \in [-1, 1]$$



Más sobre SVD



Más sobre SVD



Antes de usar el SVD se puede aplicar una transformación

TF-IDF

$$\text{tf-idf}(t, d) = \text{tf}(t, d) \cdot \text{idf}(t)$$

$$\text{idf}(t) = \log \frac{1+|D|}{1+|\{d:t \in d\}|} + 1$$

Log-Entropy

$$\text{LogEnt}(t, d) = \log(\text{tf}(t, d) + 1) \cdot W_g$$

$$W_g = 1 + \frac{\sum_d P(t, d) \log(P(t, d))}{\log(|D| + 1)}$$

$$\text{con} \quad P(t, d) = \frac{\text{tf}(t, d)}{\sum_d \text{tf}(t, d)}$$

Un LSA alternativo:

Term-context matrix

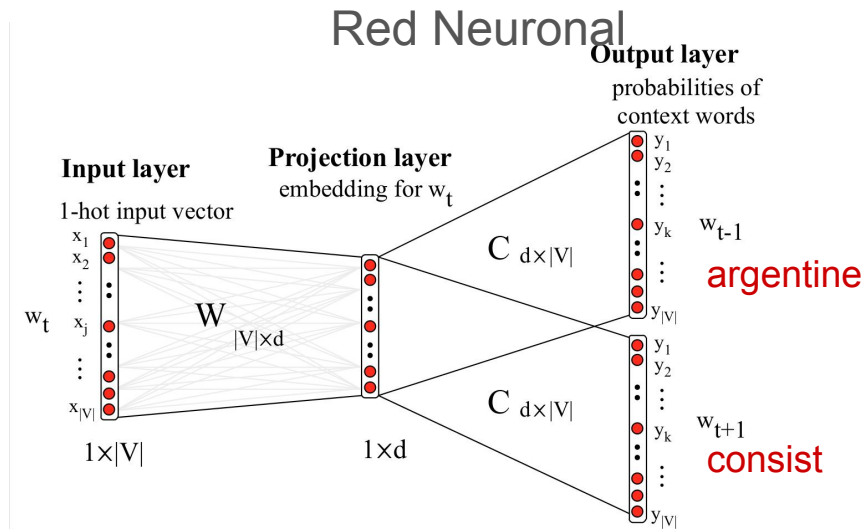
	choripan	vino	chimichurri	uva	pera	...	kiwi
choripan	47	54	23	5	2	...	1
vino	54	354	17	21	3	...	4
chimichurri	23	17	59	1	1	...	0
uva	5	21	1	203	20	...	19
pera	2	3	1	20	399	...	11
...
kiwi	1	4	0	19	11	...	61

LSA tips:

- Un número estándar de dimensiones es $k=300$ (ambiente científico)
- El k óptimo depende del número de tópicos distintos existente en los textos y de la tarea a realizar.
- Muchas veces sirve tirar la primer dimensión o hasta las primeras 50
- La similaridad coseno entre palabras no es absoluta. Cuanto mayor sea k , las palabras van a estar más distantes entre sí (menor similaridad)

Word2Vec (Skip-gram)

Choripan. The **Argentine** **choripán** **consists** of a sausage made out of beef and pork, hot off the grill, split down the middle, and served on a roll. The *chorizo* may be used whole or cut in half lengthwise, in which case it is called a *mariposa* (butterfly). It is customary to add sauces on the bread, most likely *chimichurri*.



Objective function

$$O = \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t)$$

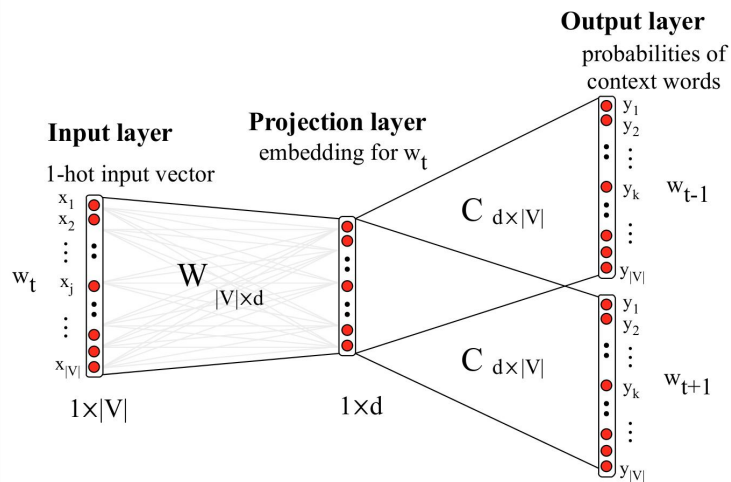
$$= \log p(\text{argentine} | \text{choripan}) + \log p(\text{consist} | \text{choripan})$$

argentine

choripan

Soft-max function

$$p(w_k | w_j) = \frac{\exp(c_k \cdot v_j)}{\sum_{i \in |V|} \exp(c_i \cdot v_j)}$$



argentina

choripan

$$p(w_k | w_j) = \frac{\exp(c_k \cdot v_j)}{\sum_{i \in |V|} \exp(c_i \cdot v_j)}$$

		1	2	...	k-1	k
costanera	v_1	0.22	0.51	...	0.12	0.11
vino	v_2	0.20	0.47	...	0.30	0.47
chimichurri	v_3	0.24	0.34	...	0.21	0.24
...	...	0.40	0.26	...	0.40	0.16
choripan	v_j	0.64	0.17	...	0.14	0.37
...
kiwi	$v_{ V }$	0.17	0.11	...	0.27	0.31

target embeddings

	costanera	...	argentina	...	kiwi
	c_1	...	c_k	...	$c_{ V }$
1	0.41	...	0.13	...	0.21
2	0.76	...	0.33	...	0.43
3	0.52	...	0.93	...	0.34
4	0.33	...	0.30	...	0.54
...
k-1	0.22	...	0.42	...	0.34
k	0.25	...	0.18	...	0.57

context embeddings

argentina choripan

$$p(w_k|w_j) = \frac{\exp(c_k \cdot v_j)}{\sum_{i \in |V|} \exp(c_i \cdot v_j)}$$

Es imposible de calcular!

Negative sampling

Sampleo k palabras (w_i)
de la distribución $p(w)$

$$\log p(w_k|w_j) \approx \log \sigma(c_k \cdot v_j) + \sum_{i=1}^k \mathbb{E}_{w_i \sim p(w)} [\log \sigma(-w_i \cdot v_j)]$$

con

$$\sigma(x) = \frac{1}{e^{-x} + 1} = \frac{e^x}{1 + e^x}$$

Con 5 Negative samples (k=3)

Ej: {el, gorro, cafetera}

$$x \gg 1 \implies \sigma(x) \approx 1$$

$$x \ll -1 \implies \sigma(x) \approx 0$$

$$\log p(w_k|w_j) \approx \log \sigma(c_k \cdot v_j) + \log \sigma(-w_1 \cdot v_j) + \log \sigma(-w_2 \cdot v_j) + \log \sigma(-w_3 \cdot v_j)$$

Función objetivo a maximizar

$$O = \sum_{k=1}^{|V|} \sum_{j=1}^{|V|} \#(w_k, w_j) \left[\log(\bar{c}_k \cdot \bar{v}_j) + \sum_{i=1}^k \mathbb{E}_{w_i \sim p(w)} [\log \sigma(-\bar{w}_i \cdot \bar{v}_j)] \right]$$

$$\text{con } \sigma(x) = \frac{1}{e^{-x} + 1} = \frac{e^x}{1 + e^x}$$

La solución óptima cumple:

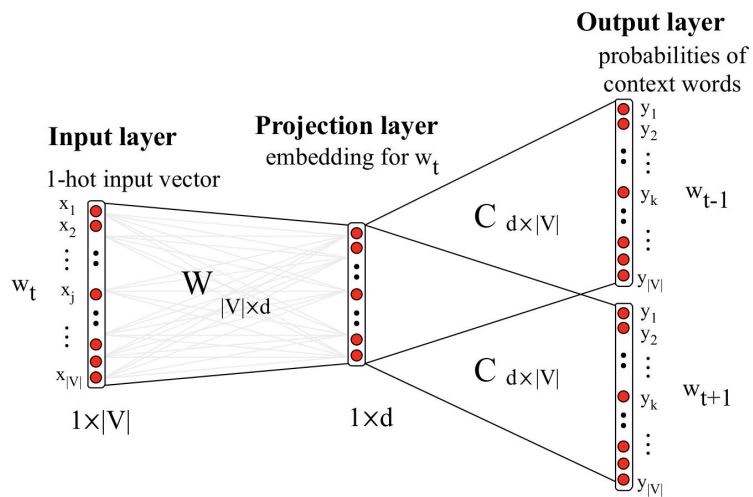
$$\bar{c}_k \cdot \bar{v}_j = PMI(w_k, w_j) - \log(k)$$

El Skip-gram factoriza el PMI en 2 matrices

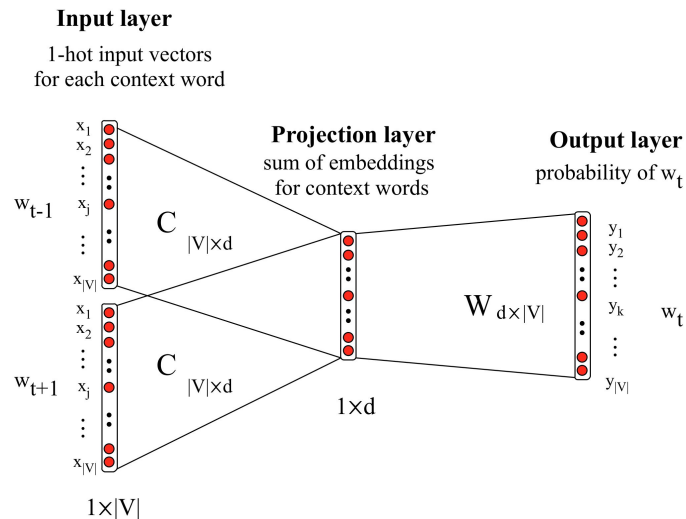
$$C.V = X^{PMI} - \log(k)$$

Word2vec

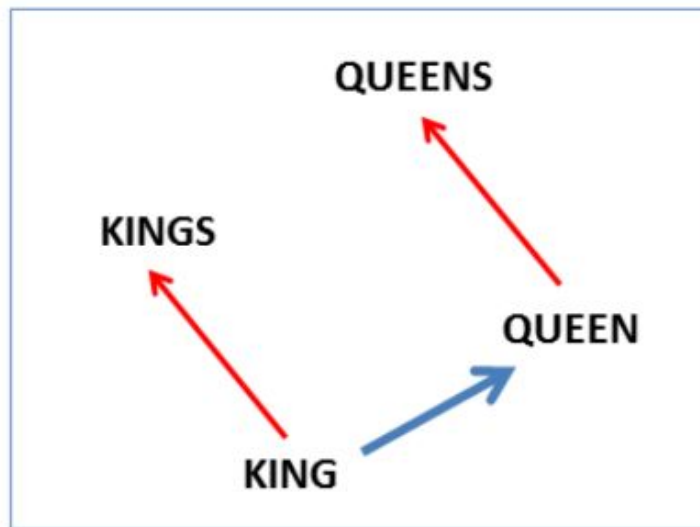
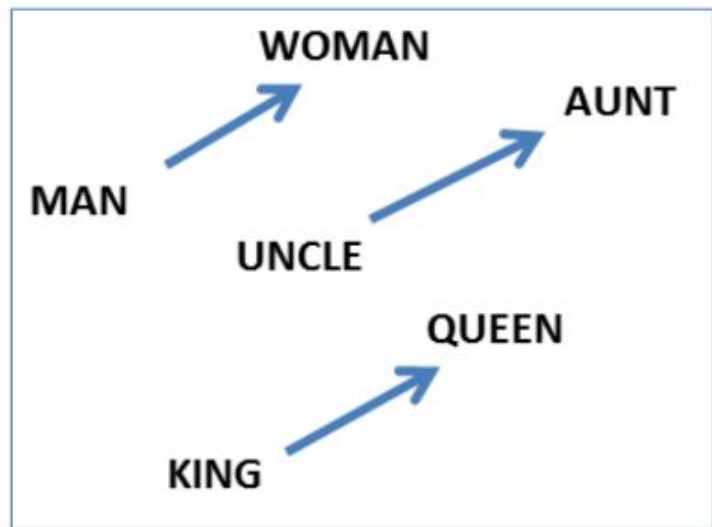
Skip-gram



CBOW



Word analogy task



King – Queen \approx Woman – Man

King \approx Woman – Man + Queen


Más modelos:

Global Vectors for Word Representation (GLOVE)

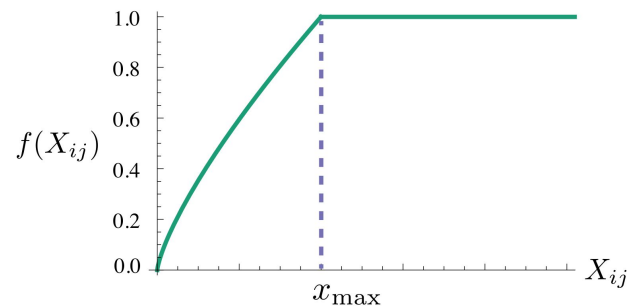
Loss function

$$J = \sum_{i,j=1}^V f(X_{ij}) \left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij} \right)^2$$

bias



The diagram shows the word bias term b_i and the context bias term \tilde{b}_j in the equation above. Arrows point from the word "bias" to both b_i and \tilde{b}_j .



$$f(x) = \begin{cases} (x/x_{\max})^\alpha & \text{if } x < x_{\max} \\ 1 & \text{otherwise} \end{cases}$$

LSA

Edad: 29 años

The Force: 14700 citas



VS

Word2vec

Edad: 6 años

The Force: 15700 citas



Evaluación de word-embeddings: TOEFL test

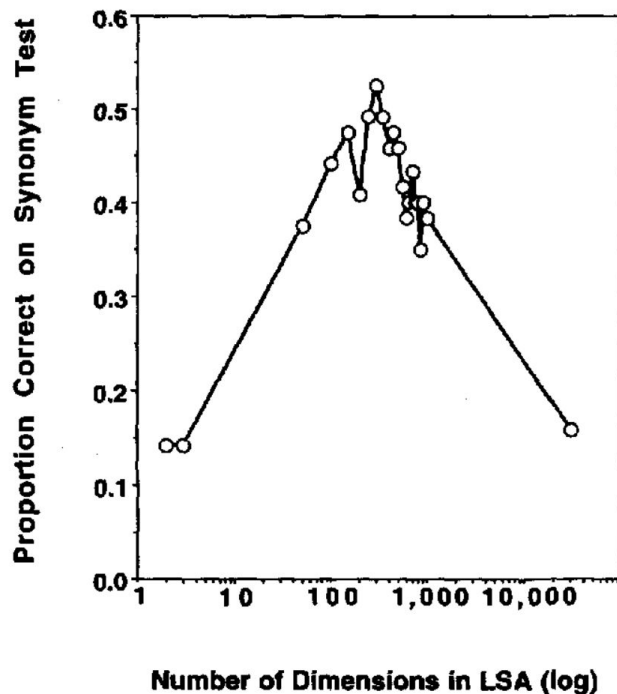
Encontrar el sinónimo entre 4 opciones

	Palabra	Opcion 1	Opcion 2	Opcion 3	Opcion 4
Pregunta 1	hue	color	scent	contrast	glare
Pregunta 2	hind	rear	curved	muscular	hairy
...
Pregunta 80	make	earn	print	trade	borrow

Busco el más cercano.

Métrica:
porcentaje de acierto

Evaluación de word-embeddings: TOEFL test



Evaluación de word-embeddings: similaridad

WordSimilarity-353 Test

Word 1	Word 2	Human
tiger	tiger	10
fuck	sex	9.44
Maradona	football	8.62
book	paper	7.46
...
professor	cucumber	0.31
king	cabbage	0.23

embedding1
1
0.52
0.42
0.45
...
0.002
0.001

embedding2
1
0.40
0.35
0.25
...
0.015
0.003

Quiero ver que embedding capture mejor las similitudes puntuadas por humanos

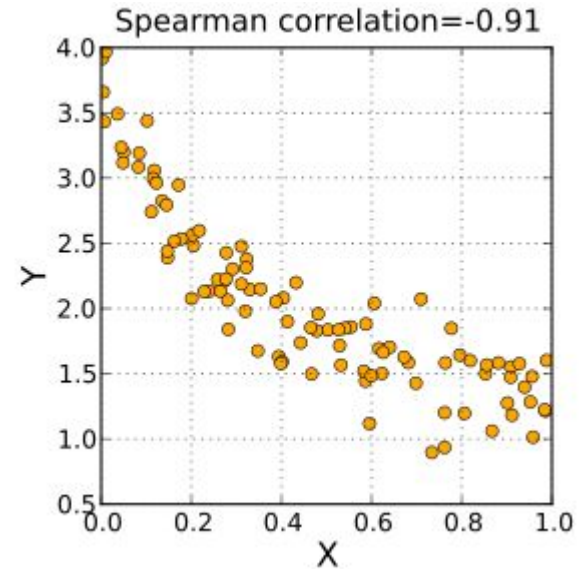
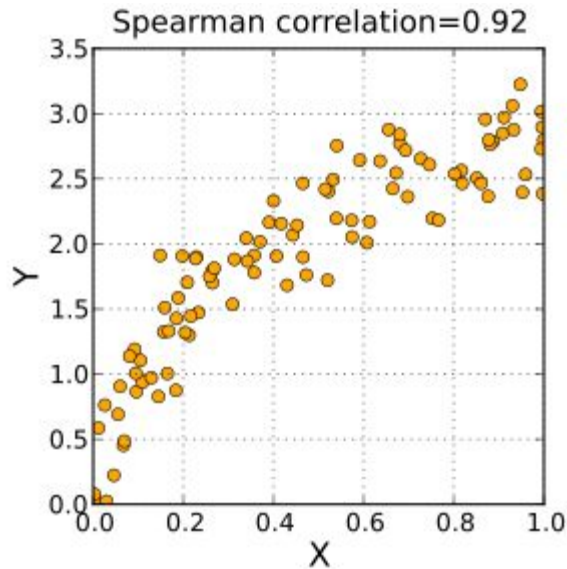
➤ Correlación

Spearman correlation

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}.$$

$$d_i = \text{rg}(X_i) - \text{rg}(Y_i)$$

↖
ranking



WordSimilarity-353 Test Collection

Human	embedding1	embedding2
10	1	1
9.44	0.52	0.40
8.62	0.42	0.35
7.46	0.45	0.25
...
0.31	0.002	0.015
0.23	0.001	0.003

Comparo correlaciones

	embedding1	embedding2
Spearman Correlation	0.82	0.74

Categorization Test



53 categorías de 10 palabras
cada una (patel 1997)

Performance: Silhouette Coefficient

Distancia promedio al
cluster más cercano

Distancia promedio al
cluster propio

$$S_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

- Probablemente no exista una técnica de word-embeddings que sea universalmente mejor en todas las tareas
- Para una tarea particular (x ej. sentiment analysis), lo óptimo es seleccionar el embedding que mejor realice esa tarea

LSA

Edad: 28 años

The Force: 13000 citas



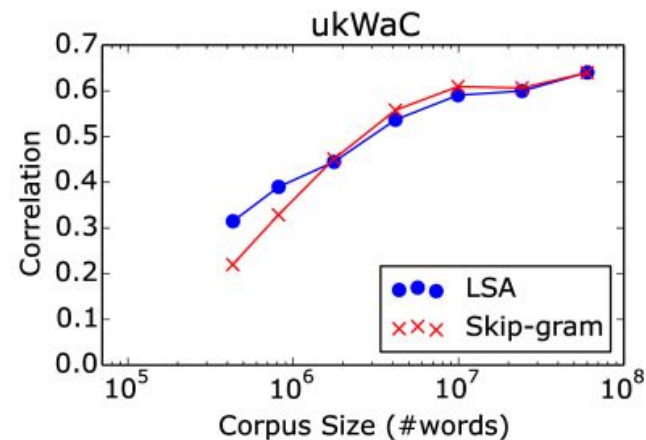
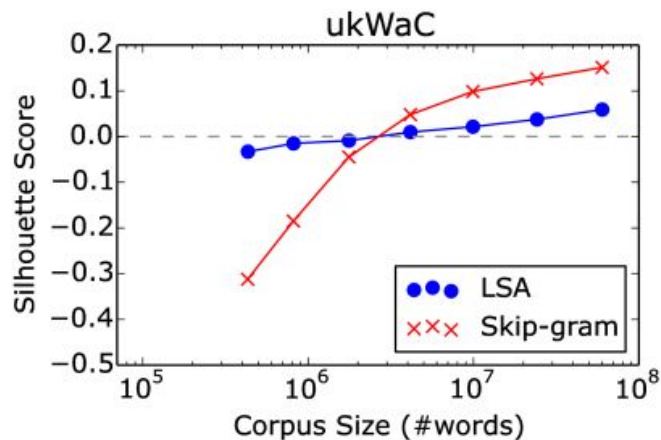
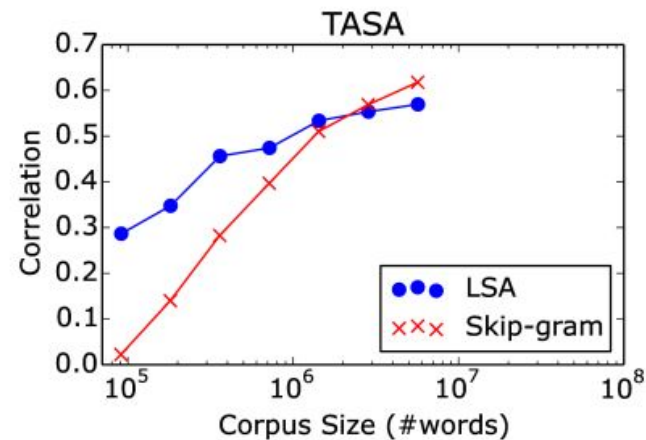
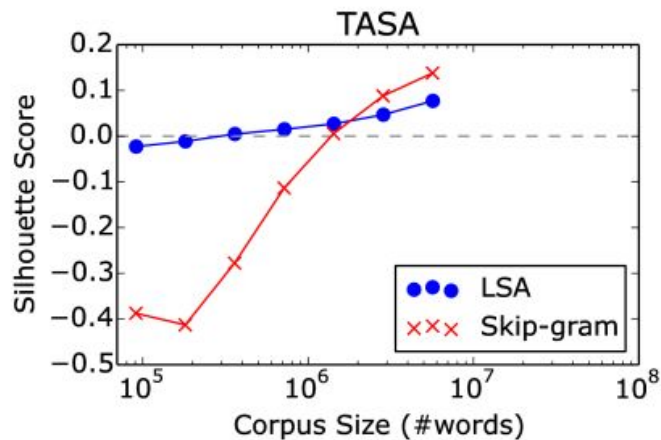
VS

Word2vec

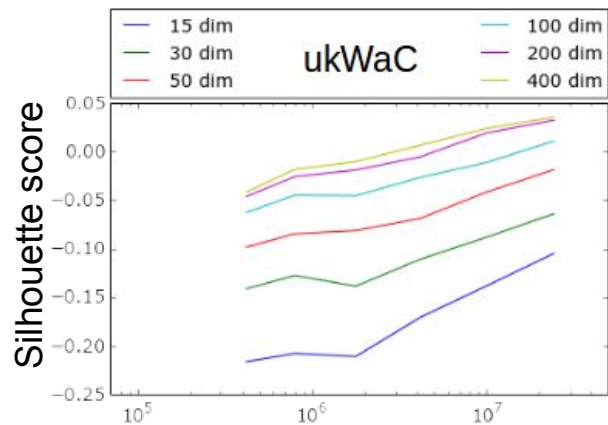
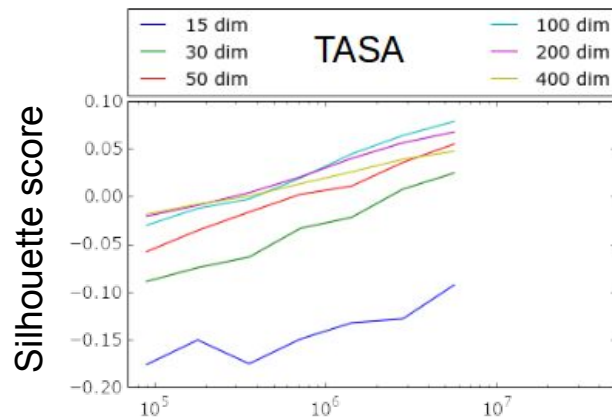
Edad: 5 años

The Force: 8000 citas

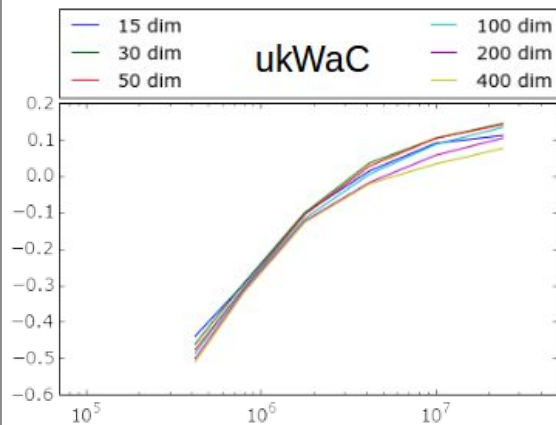
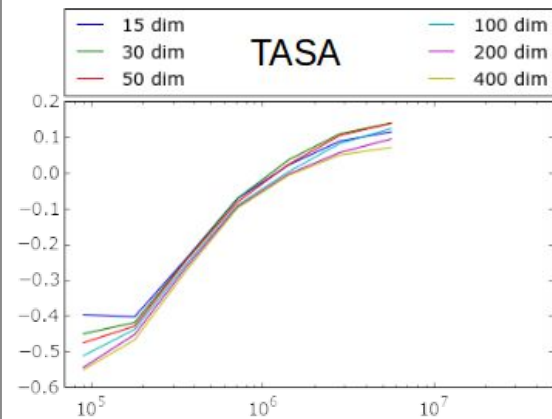




LSA



Word2vec



Pre-trained word-embeddings

- Word2vec en google news, 100B words words, (<https://code.google.com/archive/p/word2vec/>)
- Glove (<https://nlp.stanford.edu/projects/glove/>):
 - Wikipedia 2014 + Gigaword 5 (6B tokens, 400K vocab, uncased, 50d, 100d, 200d, & 300d vectors, 822 MB download): glove.6B.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
- Fasttext (<https://fasttext.cc/>):
 - crawl-300d-2M.vec.zip: 2 million word vectors trained on Common Crawl (600B tokens).
 - en wikipedia en distintos idiomas (incluido **español**)
<https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md>

Ejemplo de uso 1

Embeddings como features

Competencia: CLPsych 2017 shared task

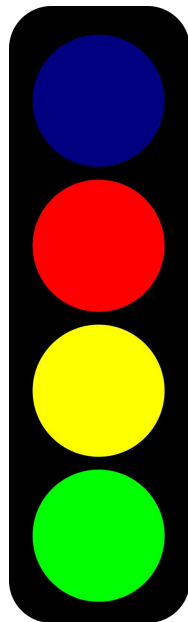


Competencia: CLPsych 2017 shared task

@Author - 06 Sep 2015, 15:56

Re: psychosis

I've had a difficult day. I got very close to self harming, but my best mates keep me safe. Now I'm at home seeing GoT.



- **Crisis:** El autor está en riesgo. Los moderadores deben atender este mensaje urgentemente
- **Rojo:** los moderadores deben atender este mensaje cuanto antes
- **Amarillo:** Los moderadores deben atender este mensaje en algún momento
- **Verde:** No requiere atención de un moderador

Dataset

	crisis	red	amber	green	total
train	40	137	296	715	1188
test	42	48	94	216	400
extra	-	-	-	-	156375

Embeddings as features

	crisis	red	amber	green	total
train	40	137	296	715	1188
test	42	48	94	216	400
extra	-	-	-	-	156375



Set de entrenamiento de un embedding

I've had a difficult day. I got very close to self harming, but my best mates keep me safe. Now I'm at home seeing GoT.

embedding

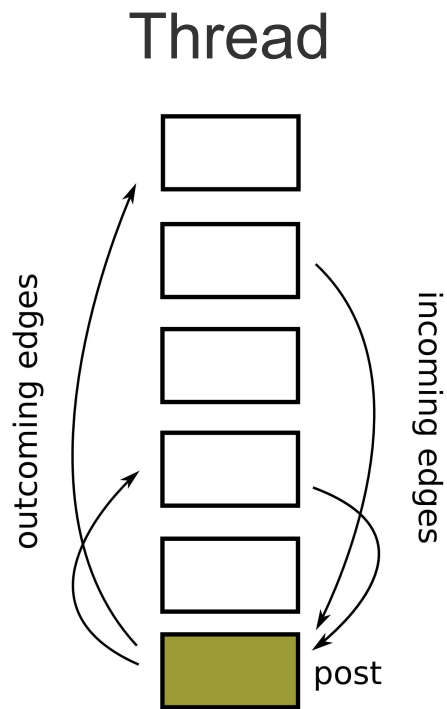


I've → (0.4,0.9,...0.3)
had → (0.5,0.8,...0.5)
... → (...)
GoT → (0.1,0.9,...0.1)

mean

→ (0.1,0.3,...0.4)

Embeddings as features



- Mean embedding incoming posts
- Mean embedding outgoing posts

Embeddings as features

I've had a difficult day. I got very close to self harming, but my best mates keep me safe. Now I'm at home seeing GoT.

embedding



I've	→ (0.4,0.9,...0.3)
had	→ (0.5,0.8,...0.5)
...	→ (...)
GoT	→ (0.1,0.9,...0.1)



Deep Neural Network

prediction

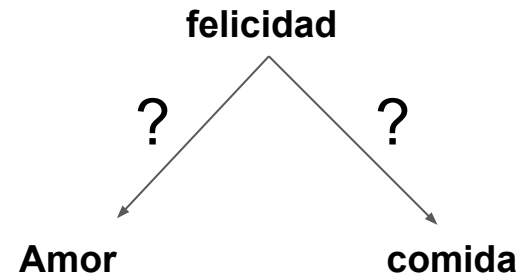
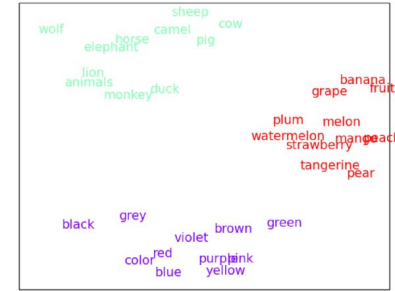


Ejemplo de uso 2

**Embeddings como método
para investigar un corpus**

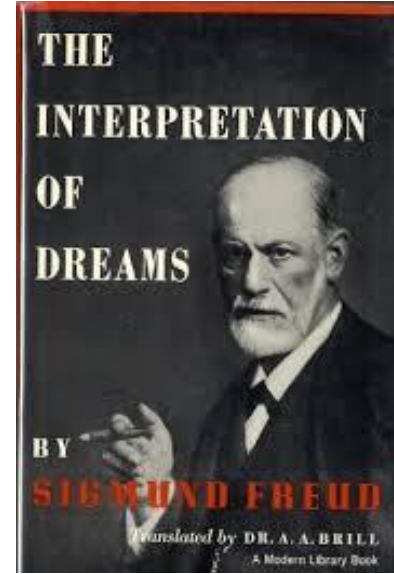
Estudio de asociaciones de palabras

Inspeccionando el dataset



Los sueños como ventana a la mente

Más de 20.000 reportes de sueños

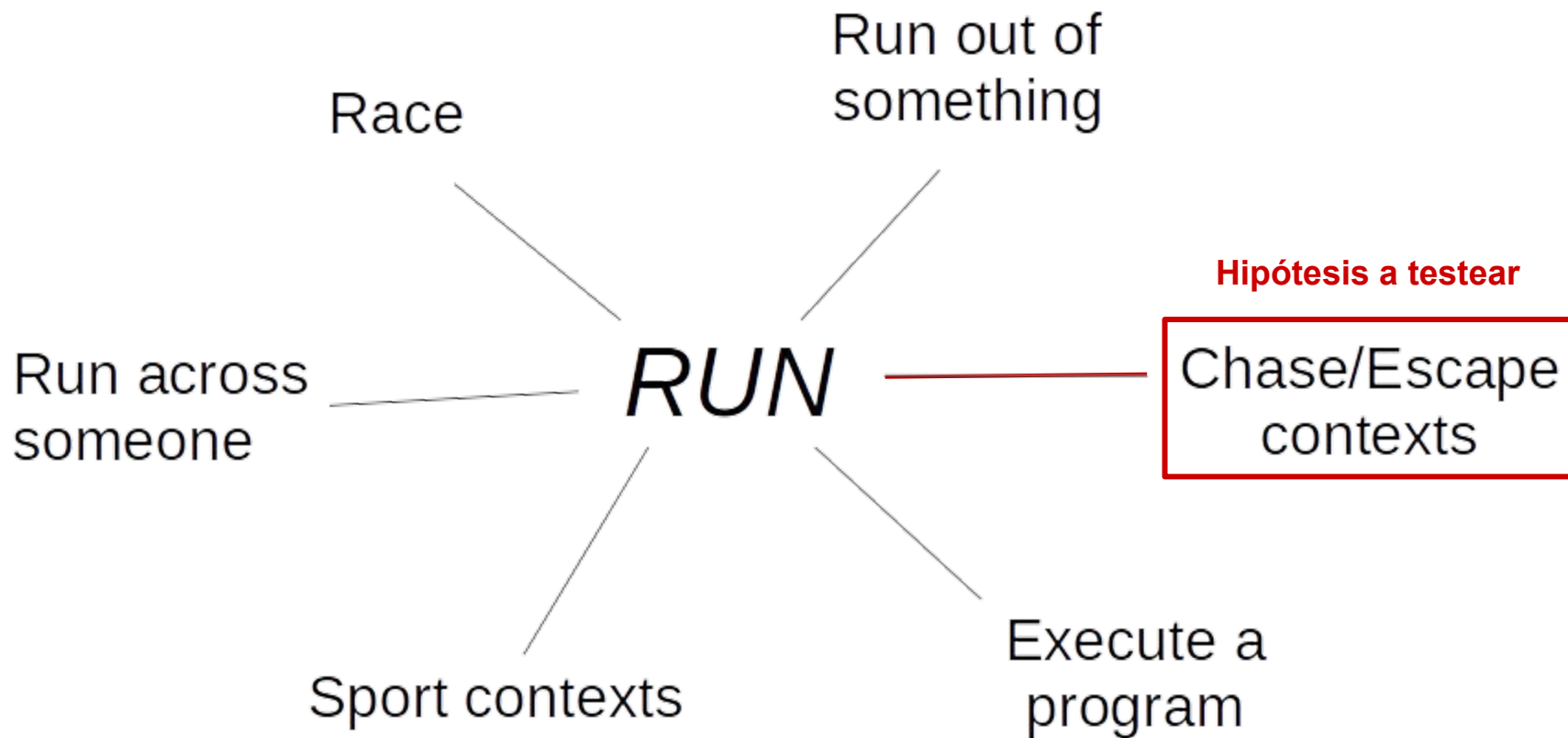


La interpretación de los
sueños, Freud 1899

Teoría de la simulación de amenazas



De que corremos en los sueños?



De que corremos en los sueños?

Palabras más cercanas a “run”

Rank	LSA Dreams	Word2vec Dreams	LSA TASA	Word2vec TASA	LSA UkWaC	Word2vec UkWaC
1st	running	chase	drive	running	running	running
2nd	escape	running	ride	runs	dash	runs
3rd	catch	scream	running	ran	jumping	marathon
4th	chase	chasing	stay	go	jump	bash
5th	chasing	escape	go	operate	yell	start
6th	follow	runs	haul	organise	workouts	rlogin
7th	ran	chases	walk	compete	kick	runners
8th	sight	grab	jump	start	jogging	starts
9th	coming	screaming	throw	break	workout	jump
10th	runs	nazi	staying	install	stretch	loaded
11th	dangerous	hide	get	operated	tiring	weekend
12th	guards	chased	carry	gone	fun	vms
13th	robbers	yells	move	move	fast	startx
14th	hide	safety	stop	set	repetitions	marathons
15th	toward	wolf	cut	managed	throw	mkdir

Ejemplo de uso 3

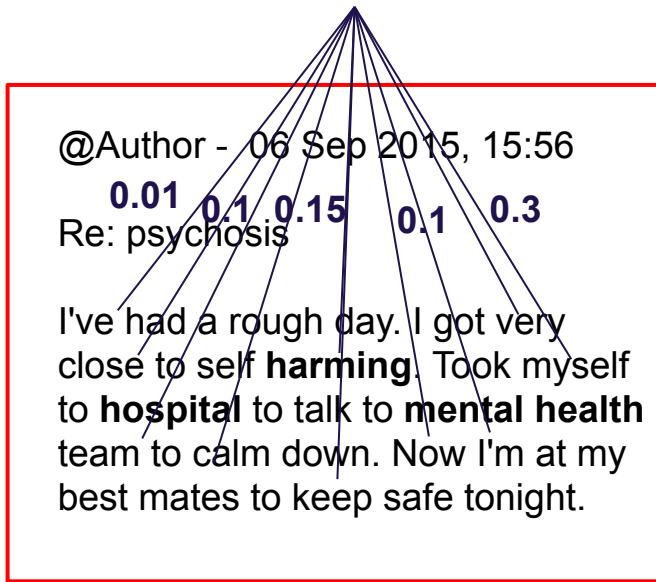
**Embeddings como métrica de
similaridad semántica externa**

Competencia: CLPsych 2017 shared task



Word-embeddings como métrica de similaridad semántica

depression



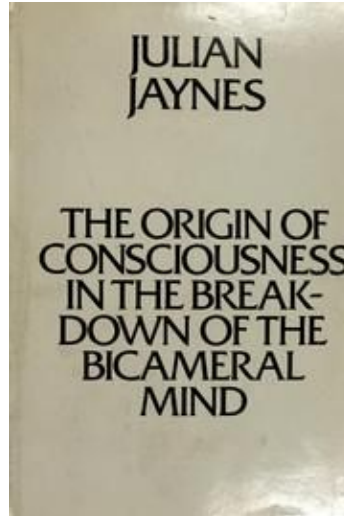
- Cercanía promedio = 0.05
- Fracción de palabras cercanas = 4/35

Usamos muchas palabras
depression - fear - anxiety -
mental_health - hopelessness -
suicide - antidepressant

Palabras cercanas a depression:

Bipolar_disorder, depression_anxiety, mental_illness, psychosis, alcoholism, depressive, suicidal_thoughts, schizophrenia, anxiety_disorders, psychological_distress, manic_depression, anxiety_disorder, mental_disorders, Depression, major_depressive_disorder, postpartum_depression, obsessive_compulsive_disorder, mood_disorders, insomnia, depressive_symptoms, psychiatric_disorders, bulimia, loneliness, PTSD, migraines, antidepressants, dementia

La historia de la introspección



Hipótesis: cambios en la introspección a través del tiempo

La historia de la introspección

Introspección

0.12

Hipnotismo de un flagelo dulce, tan dulce. cuero, piel y metal carmín y charol. Cuando el cuerpo no espera lo que llaman amor. Cada lágrima de hambre el mas puro néctar nada mas dulce que el deseo en cadenas. Mas se pide y se vive canción animal

Libro 1

Introspección

0.09

Ella tambien se canso de este sol viene a mojarse los pies a la luna
Ella tambien se canso de este sol viene a mojarse los pies a la luna
Cuando se cansa de tanto querer ella es tan clara que ya no es ninguna

Libro 2

...

Introspección

0.17

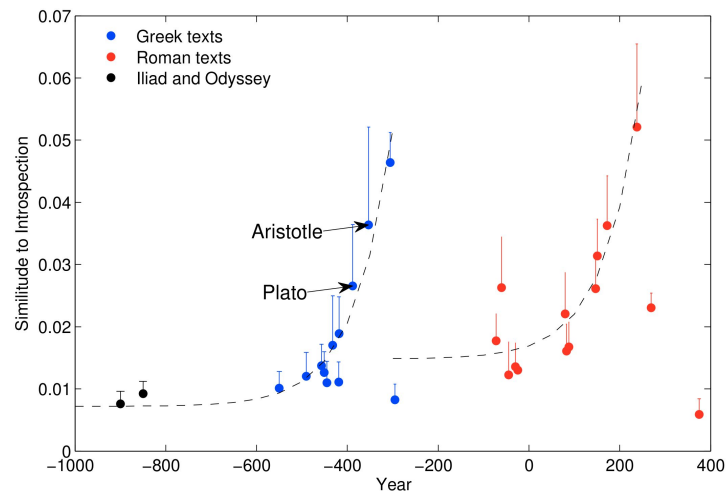
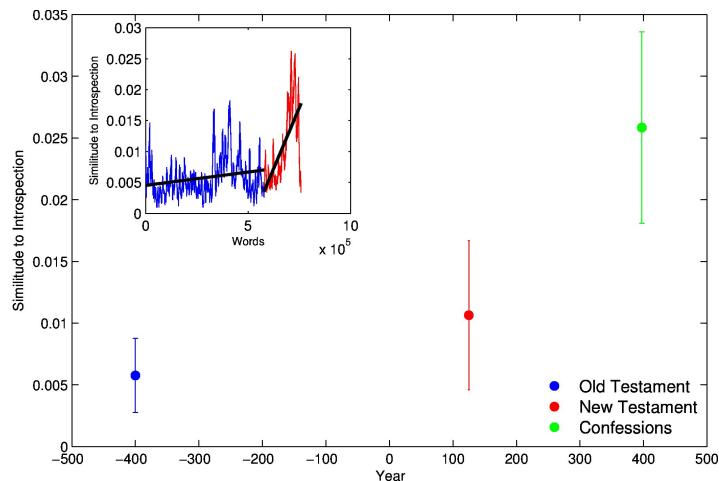
Me veras volar por la ciudad de la furia, donde nadie sabe de mi y yo soy parte de todos. Nada cambiara con un aviso de curvas, en sus caras veo el temor ya no hay fabulás en la ciudad de la furia

Libro N

Primeras 20 palabras:

Soul_searching, introspective, navel_gazing, contemplation, catharsis, contemplative, self_pity, rumination, self_indulgence, reflection, discernment, enlightenment, recrimination, repentance, self_congratulation, meditation, cynicism, reminiscence, humility, self_loathing

La historia de la introspección



"A quantitative philology of introspection." Diuk, Carlos G., D. Fernandez Slezak, I. Raskovsky, M. Sigman, and G. A. Cecchi. (2012).

Observatorio de Hollywood

¿Qué ves cuando me ves?



SubRip File

279
00:24:09,973 --> 00:24:11,665
I want you to rest well, and a
month from now...

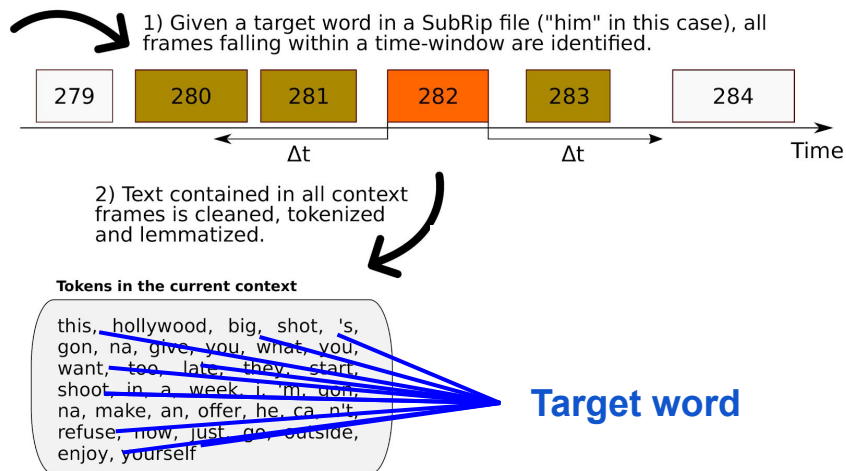
280
00:24:11,839 --> 00:24:15,203
This Hollywood big shot's
gonna give you what you want.

281
00:24:15,373 --> 00:24:18,032
Too late. They start shooting in
a week.

282
00:24:18,573 --> 00:24:21,561
I'm gonna make **him** an offer he
can't refuse.

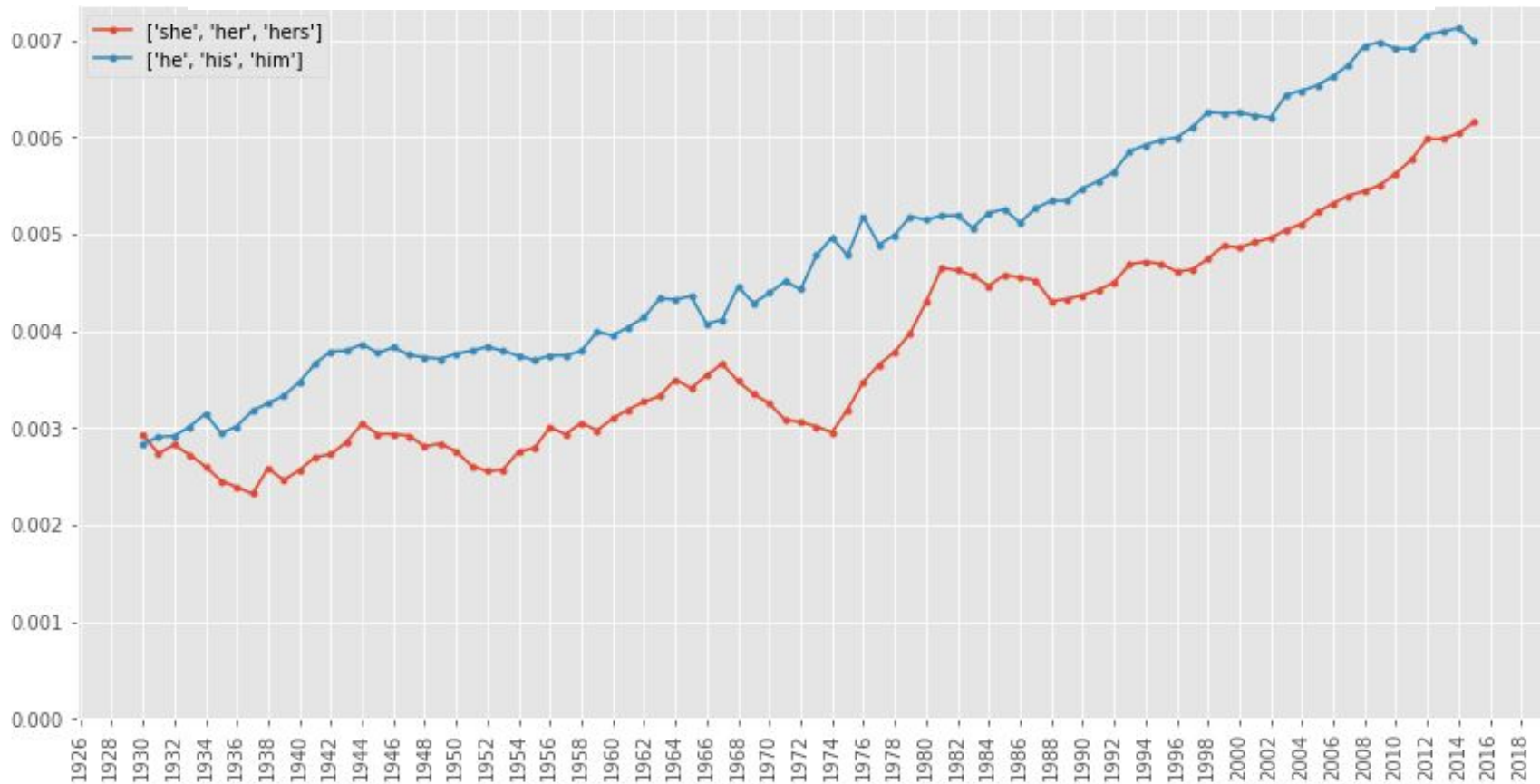
283
00:24:24,606 --> 00:24:26,334
Now just go outside, enjoy
yourself...

284
00:24:26,572 --> 00:24:30,505
and forget about all this
nonsense.



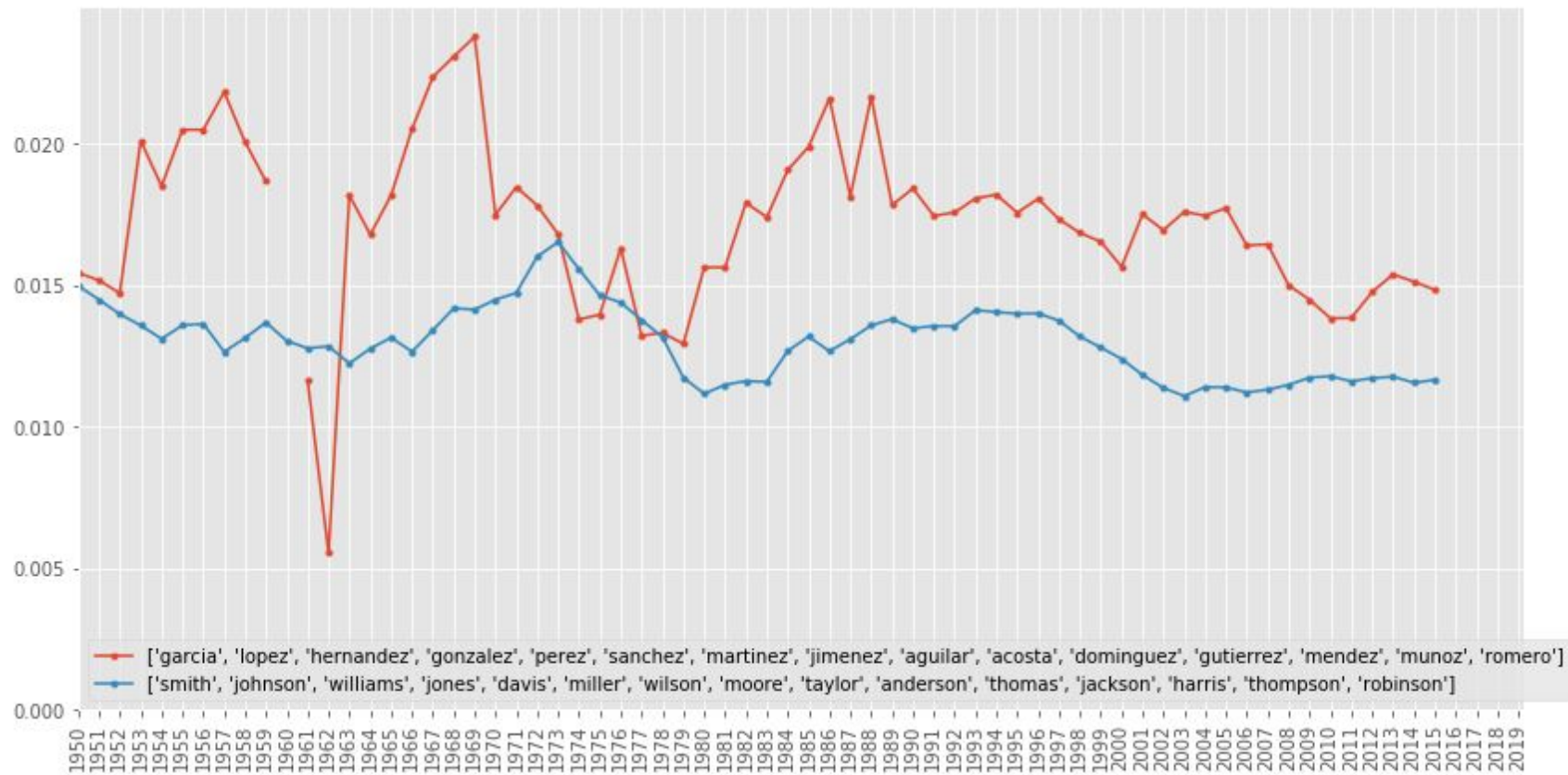
The process is repeated for every word in every subtitle under analysis.

Asociación con **technology**

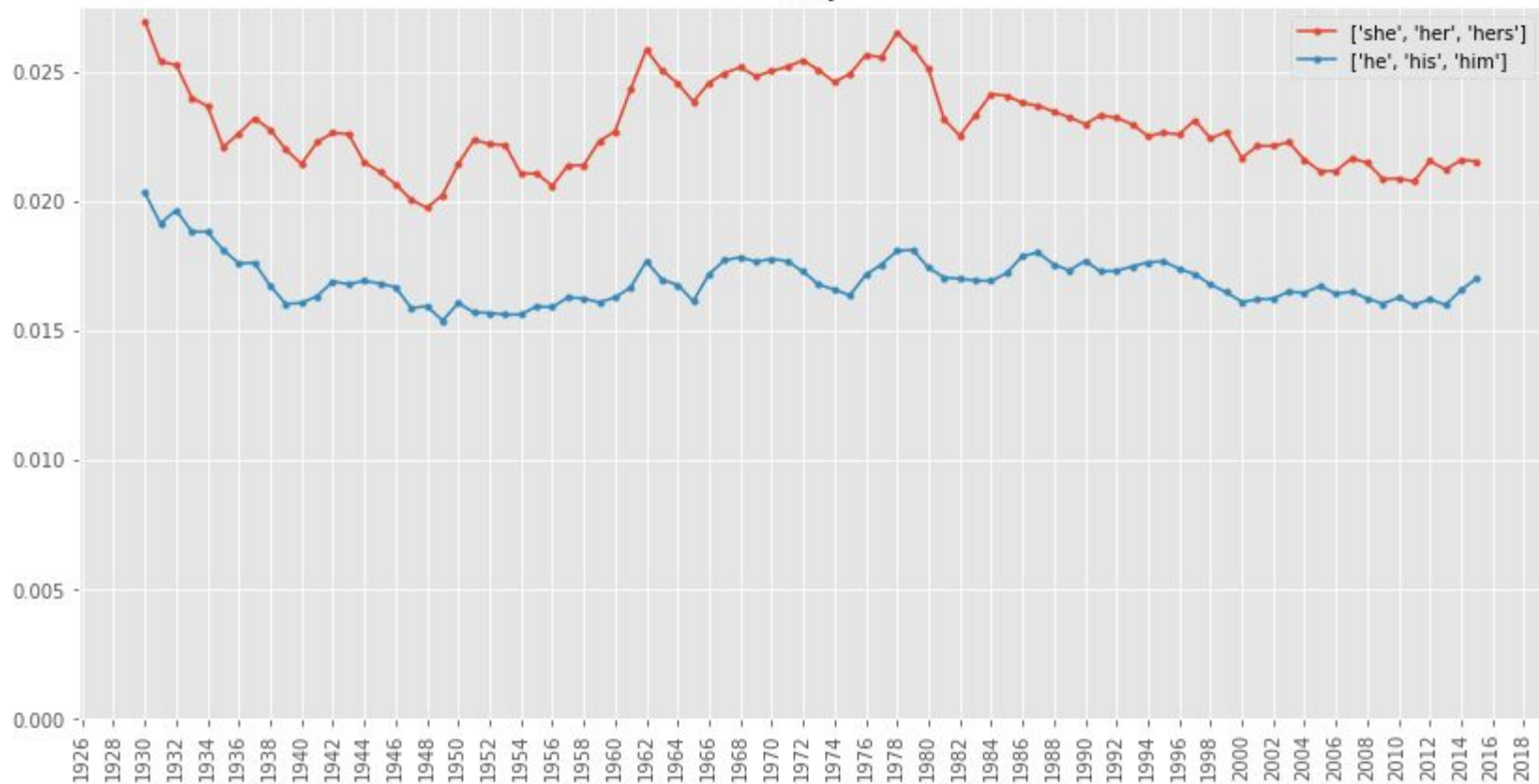


¿Quien es Smith y quién es García?

Asociación con **robbers**

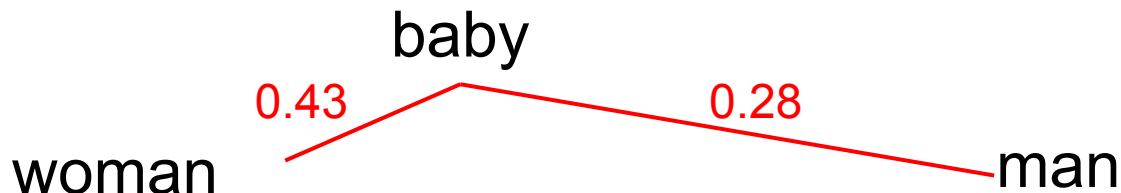


baby



Palabras más cercanas a “baby” según Word2vec

Newborn, babies, infant, newborn_baby, toddler, child, triplets, Baby, newborns, unborn_baby, twins, tot, quadruplets, newborn_babies, infants, mother, grandchild, **daughter**, kitten, pregnant, **mommy**, Babies, birth, firstborn, Newborn, fetus, puppy, puppies, **mom**, toddlers, **mama**, pups, womb, **girl**, unborn_child, birthing, crib, pup, kittens, **pregnancy**, daddy, Caesarean_section, stillborn, **boy**, premature_babies, Jayden, **expectant_mothers**, **mothers**, **twin_daughters**, pacifier, diaper_bag, **mommies**, expectant_parents, stroller, piglet, preschooler, diaper, **son**, **caesarean_section**, Ava, baby_sitter, **expectant_mother**, **daughter_Ava**, **breastfeeding**, babysitter, **momma**, stork, cuddles, Infant, children, cesarean_section, diapers, bassinet, **Mommy**, **grandmother**, tots, octuplets, childbirth, preemies, cub, piglets, **granddaughter**, **mothering**, nappy, Suri_Cruise, **surrogate_mother**, **breast_feeding**, **Mom**, born_prematurely, **maternity_ward**, **Caesarean**, cuddle, pram, chick, **mum**, Toddler, tummy, sextuplets, **midwife**, giraffe



¡Siempre chequear el significado de las palabras que uso!

FIN