Similaridad Semántica

¿25?

0 100

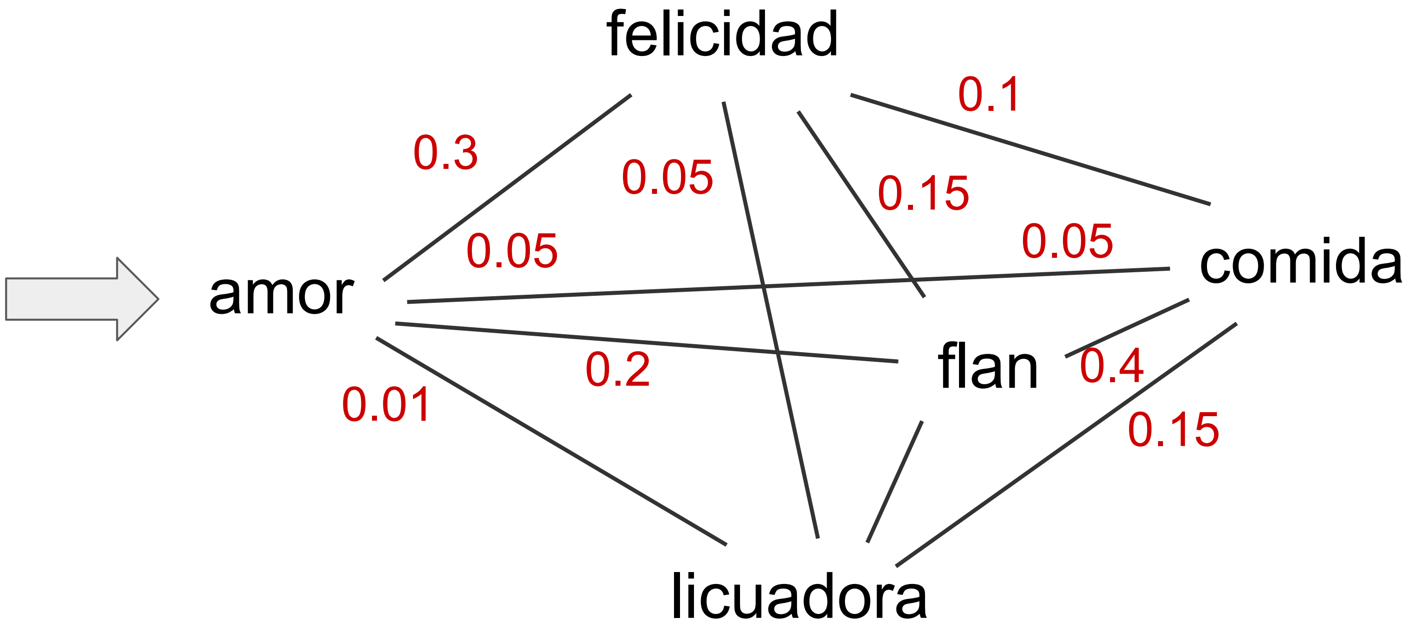
¿felicidad?

amor comida

Similaridad Semántica

Corpus de textos

WIKIPEDIA
The Free Encyclopedia



La semántica de una palabra puede deducirse de su contexto

- Sarabaraban aparece de noche
- Cuando hay un peligro aparece sarabaraban
- Sarabaraban puede ayudarte
- Sarabaraban es un científico

La semántica de una pala
deducirse de su cor

- Sarabaraban aparece de noch
- Cuando hay un peligro aparec
- Sarabaraban puede ayudarte
- Sarabaraban es un científico

J. R. Firth 1957

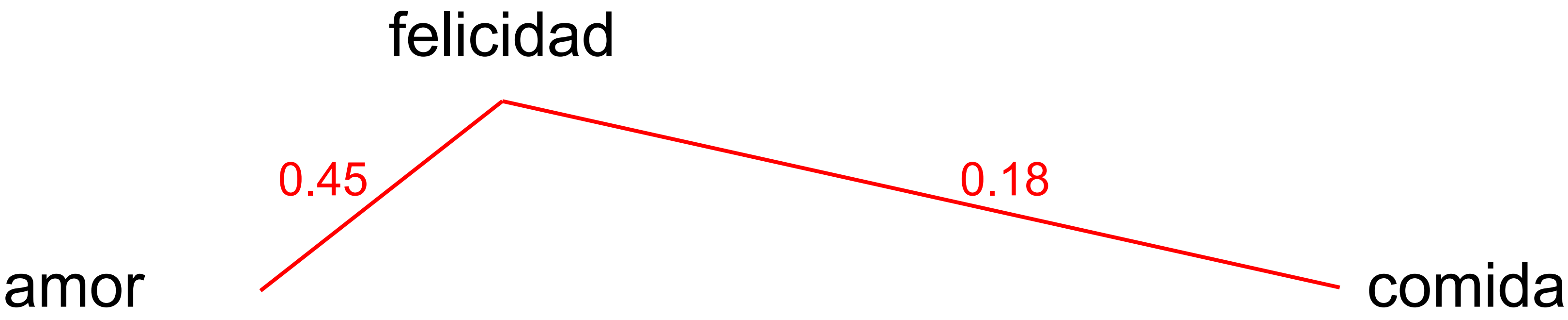
Word-embeddings

Corpus de textos

WIKIPEDIA
The Free Encyclopedia

Podemos calcular
cercanías entre palabras

Vector Space Models



Vector Space Model

Term-Document matrix

	Doc 1	Doc2	Doc 3	Doc 4	Doc 5	Doc 6	...	Doc N
choripan	1	0	3	2	0	0	...	0
vino	0	2	0	0	0	0	...	0
chimichurri	0	1	2	1	0	0	...	0
uva	1	0	0	0	0	2	...	1
pera	0	0	0	0	3	0	...	1
...	...	...	...	...	...	...	...	...
kiwi	0	0	0	0	1	2	...	0

Vector Space Model

Term-Document matrix

	Doc 1	Doc2	Doc 3	Doc 4	Doc 5	Doc 6	...	Doc N
choripan	1	0	3	2	0	0	...	0
vino	0	2	0	0	0	0	...	0
chimichurri	0	1	2	1	0	0	...	0
uva	1	0	0	0	0	2	...	1
pera	0	0	0	0	3	0	...	1
...	...	...	...	...	...	...	...	...
kiwi	0	0	0	0	1	2	...	0

Transformación TF-IDF

$$\text{tf-idf}\left(t, d\right) = \text{tf}\left(t, d\right) . \, \text{idf}\left(t\right)$$

Número de documentos en el set de entrenamiento

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

Número de documentos en los que aparece el término *t*

Vector Space Model

Term-Document matrix

	Doc 1	Doc2	Doc 3	Doc 4	Doc 5	Doc 6	...	Doc N
choripan	0.83	0	0.37	0.72	0	0	...	0
vino	0	0.02	0	0	0	0	...	0
chimichurri	0	0.91	0.22	0.31	0	0	...	0
uva	0.01	0	0	0	0	0.55	...	0.18
pera	0	0	0	0	0.13	0	...	0.11
...	...	...	...	...	...	...	...	...
kiwi	0	0	0	0	0.41	0.22	...	0

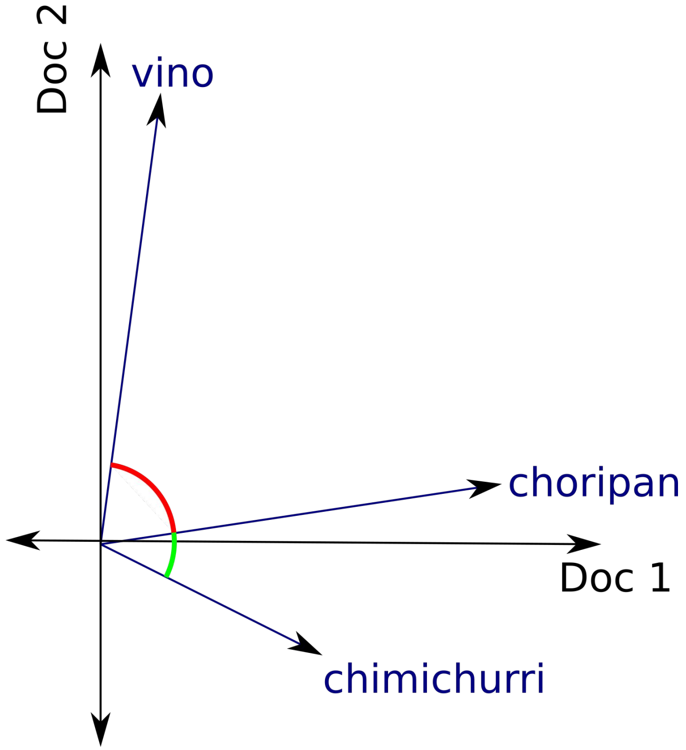
Vector Space Model

Term-Document matrix

	Doc 1	Doc2	Doc 3	Doc 4	Doc 5	Doc 6	...	Doc N
choripan	0.83	0	0.37	0.72	0	0	...	0
vino	0	0.02	0	0	0	0	...	0
chimichurri	0	0.91	0.22	0.31	0	0	...	0
uva	0.01	0	0	0	0	0.55	...	0.18
pera	0	0	0	0	0.13	0	...	0.11
...	...	...	...	...	...	...	...	...
kiwi	0	0	0	0	0.41	0.22	...	0

Similaridad coseno

$$\text{cossim}\left(\vec{v}_1, \vec{v}_2\right) = \cos(\alpha) = \frac{\vec{v}_1 . \vec{v}_2}{|\vec{v}_1| . |\vec{v}_2|}$$



Similaridad coseno

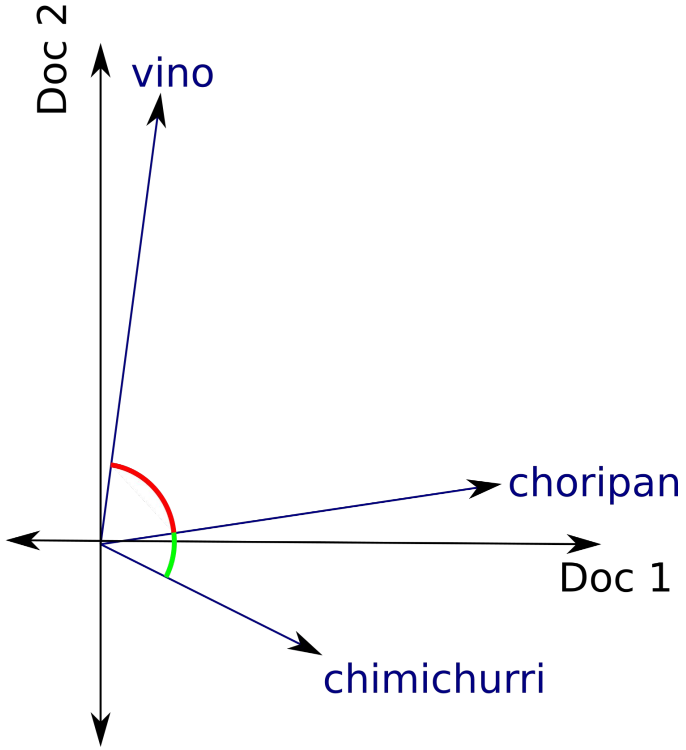
$$\text{cossim}\left(\vec{v}_1, \vec{v}_2\right) = \cos(\alpha) = \frac{\vec{v}_1 . \vec{v}_2}{|\vec{v}_1| . |\vec{v}_2|}$$

$$\text{cossim}\left(\vec{v}_1, \vec{v}_2\right) = \frac{\sum_{i=1}^{N} v_{1,i} . v_{2,i}}{\sqrt{\sum_{i=1}^{N} v_{1,i}^2} \sqrt{\sum_{i=1}^{N} v_{2,i}^2}}$$

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

Ejemplo

$$\begin{aligned}
\vec{v}_1 &= (0, 5, 1) \\
\vec{v}_2 &= (1, 0, 2) \\
\text{cossim}\left(\vec{v}_1 . \vec{v}_2\right) &= \frac{0x1 + 5x0 + 1x2}{\sqrt{0^2 + 5^2 + 1^2} \sqrt{1^2 + 0^2 + 2^2}} \approx 0.175
\end{aligned}$$



Vector Space Model

Term-Document matrix

	Doc 1	Doc2	Doc 3	Doc 4	Doc 5	Doc 6	...	Doc N
choripan	0.83	0	0.37	0.72	0	0	...	0
vino	0	0.02	0	0	0	0	...	0
chimichurri	0	0.91	0.22	0.31	0	0	...	0
uva	0.01	0	0	0	0	0.55	...	0.18
pera	0	0	0	0	0.13	0	...	0.11
...	...	...	...	...	...	...	...	...
kiwi	0	0	0	0	0.41	0.22	...	0

Information-Retrieval

Query: “El chimichurri es un condimento típico de Argentina...”

Entreno el TF-ID en el dataset

	Doc 1	Doc2	Doc 3	Doc 4	Doc 5	Doc 6	...	Doc N
choripan	0.83	0	0.37	0.72	0	0	...	0
vino	0	0.02	0	0	0	0	...	0
chimichurri	0	0.91	0.22	0.31	0	0	...	0
uva	0.01	0	0	0	0	0.55	...	0.18
pera	0	0	0	0	0.13	0	...	0.11
...	...	...	...	...	...	...	...	...
kiwi	0	0	0	0	0.41	0.22	...	0

Aplico los pesos entrenados

Query
0.15
0
0.74
0
0.09
...
0

Term-context matrix

sliding window (size=2)

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*.

Term-context matrix

	choripan	vino	chimichurri	uva	pera	...	kiwi
choripan	1	54	23	5	2	...	1
vino	54	1	17	21	3	...	4
chimichurri	23	17	0	1	1	...	0
uva	5	21	1	0	20	...	19
pera	2	3	1	20	1	...	11
...	...	...	...	...	...	...	...
kiwi	1	4	0	19	11	...	0

Term-context matrix

	choripan	vino	chimichurri	uva	pera	...	kiwi
choripan	1	54	23	5	2	...	1
vino	54	1	17	21	3	...	4
chimichurri	23	17	0	1	1	...	0
uva	5	21	1	0	20	...	19
pera	2	3	1	20	1	...	11
...	...	...	...	...	...	...	...
kiwi	1	4	0	19	11	...	0

Term-Document matrix

- First-order co-occurrence
- Syntagmatic associations
- Ejemplo:
  mostaza - hamburguesa

Term-Context matrix

- Second-order co-occurrence
- Paradigmatic associations
- Ejemplo:
  mostaza - ketchup

Similaridad semántica

Qué está más asociado al **choripan**, el **vino** o el **chimichurri**?

Term-context matrix

f =

W \ C	choripan	vino	chimichurri	en
choripan	0	10	10	30
vino	10	5	5	50
chimichurri	10	5	0	10
en	30	50	10	100

Pointwise Mutual Information (PMI)

$$\mathrm{PMI}(w, c) = \log\left(\frac{P(w,c)}{P(w)\,.\,P(c)}\right)$$

W \ C	choripan	vino	chimichurri	en
choripan	0	10	10	30
vino	10	5	5	50
chimichurri	10	5	0	10
en	30	50	10	100

Church and Hanks (1990) Word Association Norms, Mutual Information, and Lexicography

Ejemplo

$f =$

W \ C	choripan	vino	chimichurri	en	N_C
choripan	0	10	10	30	50
vino	10	5	5	50	70
chimichurri	10	5	0	10	25
en	30	50	10	100	180
N_W	50	70	25	190	335

$$p_{ij} = \frac{\boxed{f_{ij}}}{\boxed{\sum_{i=1}^{W} \sum_{j=1}^{C} f_{ij}}}$$

$$p_{i*} = \frac{\boxed{\sum_{j=1}^{C} f_{ij}}}{\boxed{\sum_{i=1}^{W} \sum_{j=1}^{C} f_{ij}}}$$

$$p_{*j} = \frac{\boxed{\sum_{i=1}^{W} f_{ij}}}{\boxed{\sum_{i=1}^{W} \sum_{j=1}^{C} f_{ij}}}$$

$$\mathrm{PMI}\,(w, c) = \log\left(\frac{p_{ij}}{p_{i*}p_{*j}}\right)$$

$$\mathrm{PMI}\,(w = chimi, c = chori) = \log\left(\frac{\frac{10}{335}}{\frac{25}{335}\,\frac{50}{335}}\right) \approx 1.422$$

Ejemplo

$$\mathrm{PMI}\,(w,c) = \log\left(\frac{P(w,c)}{P(w)\;.\;P(c)}\right)$$

PMI =

W \ C	choripan	vino	chimichurri	en
choripan	-	-0.044	**1.422**	**0.081**
vino	-0.044	-1.549	-0.063	**0.333**
chimichurri	**1.422**	-0.063	-	-0.504
en	**0.081**	**0.333**	-0.504	-0.108

Otra forma de verlo

$$\mathrm{PMI}(w,c) = \log\left(\frac{P(w,c)}{P(w)\,.\,P(c)}\right)$$

$$\mathrm{PMI}(w,c) = \log\left(\frac{P(c|w)P(w)}{P(w)\,.\,P(c)}\right)$$

$$= \log\left(\frac{P(c|w)}{P(c)}\right)$$

W \ C	choripan	vino	chimichurri	en	N_C
choripan	0	10	10	30	50
vino	10	5	5	50	70
chimichurri	10	5	0	10	25
en	30	50	10	100	190
N_W	50	70	25	190	335

$$= \log\left(\frac{P(c{=}chori|w{=}chimi)}{P(c{=}chori)}\right) = \log\left(\frac{10/25}{50/335}\right) \approx 1.422$$

¿Que significa un PMI negativo?

$$\mathrm{PMI}\,(w,c) = \log\left(\frac{P(w,c)}{P(w)\;.\;P(c)}\right)$$

Si $P(w)=P(c)=10^{-6}$

Para que $PMI<0 \longrightarrow P(w,c)<10^{-12}$

Se suele usar el Positive-PMI (PPMI)

$$\mathrm{PPMI}(w,c) = \max\left(0, \log\left(\frac{P(w,c)}{P(w)\,.\,P(c)}\right)\right)$$

PPMI =

W \ C	choripan	vino	chimichurri	en
choripan	0	0	**1.422**	**0.081**
vino	0	0	0	**0.333**
chimichurri	**1.422**	0	0	0
en	**0.081**	**0.333**	0	0

Cuidado!

$$\mathrm{PPMI}(w,c) = \max\left(0, \log\left(\frac{P(w,c)}{P(w)\,.\,P(c)}\right)\right)$$

En el caso límite donde 2 palabras están totalmente correlacionadas como por ejemplo (hocus pocus): P(w,c)=P(w)=P(c):

$$\mathrm{PMI}(w,c) = \log\left(\frac{P(w,c)}{P(w)\,.\,P(c)}\right) = \log\left(\frac{1}{P(w,c)}\right) = -\log(P(w,c))$$

P(w,c) chicos dan PPMIs mas grandes!

Jurafsky and Martin (2017) Speech and Language Processing, 3rd editions
Bouma (2009) Normalized Pointwise Mutual Information in Collocation Extraction

Soluciones

$$\text{PPMI}(w,c) = \max\left(0, \log\left(\frac{P(w,c)}{P(w)\,.\,P(c)}\right)\right)$$

- ➢ Filtrar $f_{ij} < k$

Levy et al (2015) Improving distributional similarity with lessons learned from word-embeddings
Jurafsky and Martin (2017) Speech and Language Processing, 3rd editions
Bouma (2009) Normalized Pointwise Mutual Information in Collocation Extraction

Soluciones

$$\mathrm{PPMI}(w,c) = \max\left(0, \log\left(\frac{P(w,c)}{P(w)\,.\,P(c)}\right)\right)$$

- ➢ Filtrar f_{ij} <k
- ➢ Aumentar P(c), Levy et al. (2015)

$$P_\alpha(c) = \frac{count(c)^\alpha}{\sum_c counts(c)^\alpha}$$

Levy et al (2015) Improving distributional similarity with lessons learned from word-embeddings
Jurafsky and Martin (2017) Speech and Language Processing, 3rd editions
Bouma (2009) Normalized Pointwise Mutual Information in Collocation Extraction

Soluciones

$$\mathrm{PPMI}(w,c) = \max\left(0, \log\left(\frac{P(w,c)}{P(w)\,.\,P(c)}\right)\right)$$

- ➢ Filtrar f_{ij} <k
- ➢ Aumentar P(c), Levy et al. (2015) $$P_\alpha(c) = \frac{count(c)^\alpha}{\sum_c counts(c)^\alpha}$$
- ➢ Add-k smoothing

	choripan	vino	chimichurri	en
choripan	30+k	10+k	10+k	30+k
vino	10+k	45+k	5+k	50+k
chimichurri	10+k	5+k	20+k	10+k
en	30+k	50+k	10+k	500+k

Levy et al (2015) Improving distributional similarity with lessons learned from word-embeddings
Jurafsky and Martin (2017) Speech and Language Processing, 3rd editions
Bouma (2009) Normalized Pointwise Mutual Information in Collocation Extraction

Soluciones

$$\mathrm{PPMI}\,(w,c) = \max\left(0, \log\left(\frac{P(w,c)}{P(w)\,.\,P(c)}\right)\right)$$

- ➢ Filtrar f_{ij} <k
- ➢ Aumentar P(c), Levy et al. (2015) $P_\alpha(c) = \frac{count(c)^\alpha}{\sum_c counts(c)^\alpha}$
- ➢ Add-k smoothing
- ➢ Normalized-PMI, Bouma (2009)

	choripan	vino	chimichurri	en
choripan	30+k	10+k	10+k	30+k
vino	10+k	45+k	5+k	50+k
chimichurri	10+k	5+k	20+k	10+k
en	30+k	50+k	10+k	500+k

$$\mathrm{NPMI}\,(w,c) = \frac{\log\left(\frac{P(w,c)}{P(w)\,.\,P(c)}\right)}{-\log(P(w,c))}$$

Levy et al (2015) Improving distributional similarity with lessons learned from word-embeddings
Jurafsky and Martin (2017) Speech and Language Processing, 3rd editions
Bouma (2009) Normalized Pointwise Mutual Information in Collocation Extraction

Aplicaciones: Collocations (expresiones)

Tokenización:

- ➢ “Las tardecitas de **Buenos Aires** tiene ese qué sé yo, viste?
- ➢ “Spinetta era de **villa urquiza**”

Aplicaciones: Collocations (expresiones)

Tokenización:

➢ “Las tardecitas de **Buenos Aires** tiene ese qué sé yo, viste?

➢ “Spinetta era de **villa urquiza**”

Más ejemplos:

- martin fierro
- salud mental
- ping pong
- susana gimenez

Identificación de collocations

Opciones:

- ➢ Usar listas de collocation
- ➢ Buscar bigramas en un corpus con un alto PPMI

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*.

Otras opciones:

- Implementación de Gensim: NPMI

- Implementación de NLTK: PMI

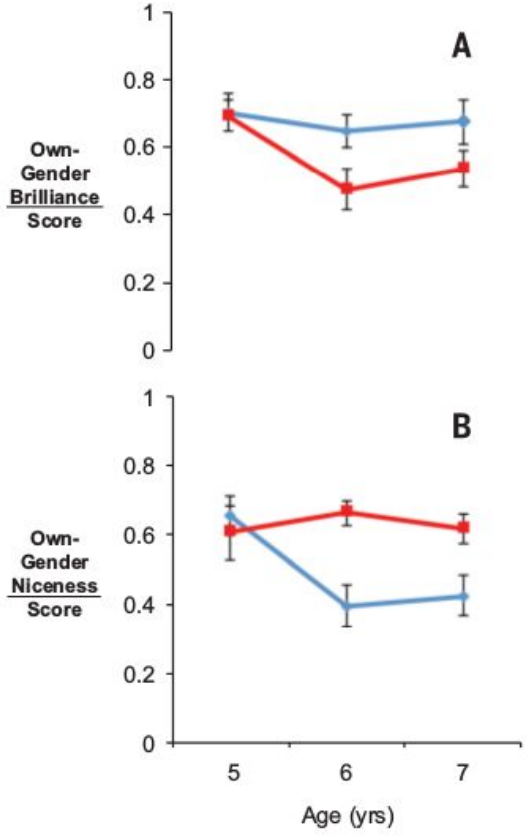
Choripan. The Argentine *choripán* consists of a sausage made out of beef and pork, hot off the grill, split down the middle, and served on a roll. The *chorizo* may be used whole or cut in half lengthwise, in which case it is called a *mariposa* (butterfly). It is customary to add sauces on the bread,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

Observatorio del cine

“Brilliance = males” stereotype

A) Story about a really really **smart** person

B) Story about a really really **nice** person



Bian et al. (2017) Gender stereotypes about intellectual ability emerge early and influence children’s interests

Cuales son las fuentes de estereotipos?

Hipótesis:

- ➢ Tratamiento diferencial de padres y maestras/os
- ➢ Falta de roles modelos
- ➢ Exposición a productos culturales que refuerzan el estereotipo

“Brilliance = males” stereotype in films subtitles

opensubtitles.org

Más de 11.000 subtítulos

NLP

Análisis

"Brilliance" related words:

ingenious, genius, ingeniousness, ingeniously, bright, brightness, brightly, brilliant, brilliance, brilliantly, clever, cleverness, cleverly, intelligent, intelligence, intelligently.

Pronombres femeninos:

she, hers, her, herself.

Pronombres masculinos:

he, his, he, himself.

Quantifying “brilliance=male” stereotype in films

$$PMI(w,c) = log\left(\frac{p(w,c)}{p(w)p(c)}\right)$$

w = pronombres

c = “brilliance” related words

Quantifying “brilliance=male” stereotype in films

$$PMI(w,c) = log\left(\frac{p(w,c)}{p(w)p(c)}\right)$$

w = pronombres
c = “brilliance” related words

- Asociación sintagmática: 1^{st} order co-occurrences
- Facil interpretacion

Quantifying “brilliance=male” stereotype in films

$$PMI(w, c) = log\left(\frac{p(w, c)}{p(w)p(c)}\right)$$

w = pronombres

c = “brilliance” related words

- Asociación sintagmática: 1st order co-occurrences
- Facil interpretacion

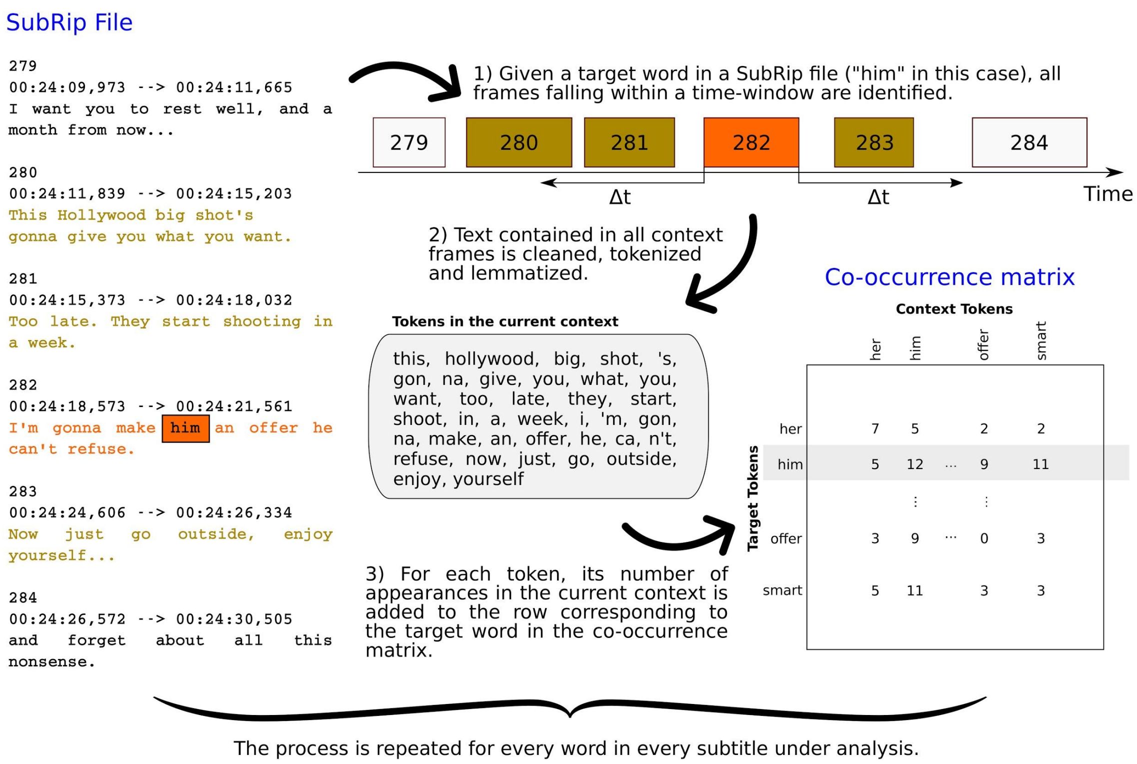
Gender bias

brilliance-male association

brilliance-female association

$$\Delta PMI \quad = \quad PMI(w_m, c) - PMI(w_f, c)$$

Sliding time-window to compute co-occurrence



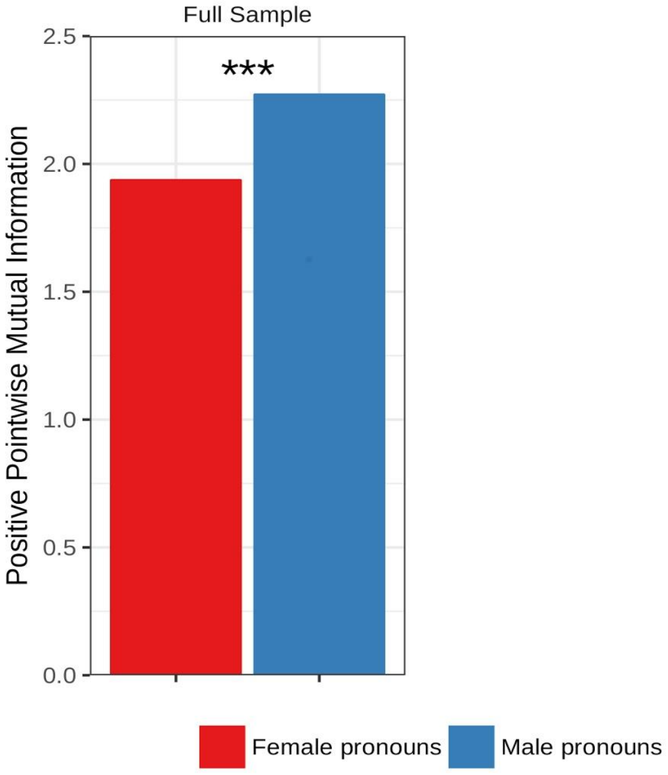
Unifico filas y columnas de pronombres y estereotipos

- ➢ Female pronouns = {she, hers, her, herself }
- ➢ Male pronouns = {he, his, him, himself }

	he	dancer	pilot	...	table
she	100	40	15	...	50
he	200	10	300	...	50
...	...	...	...	...	...
bus	40	5	25	...	5
cup	20	9	5	...	45

	he	dancer	estereo	...	table
F pron	200	125	50	...	80
M pron	400	30	500	...	90
...	...	...	...	...	...
bus	40	5	1	...	5
cup	20	9	2	...	45

"brilliance = male" stereotype



$$\Delta PMI = log\left(\frac{p(c|w_m)}{p(c|w_f)}\right) = 0.33$$

$$\frac{p(c|w_m)}{p(c|w_f)} = 1.26$$

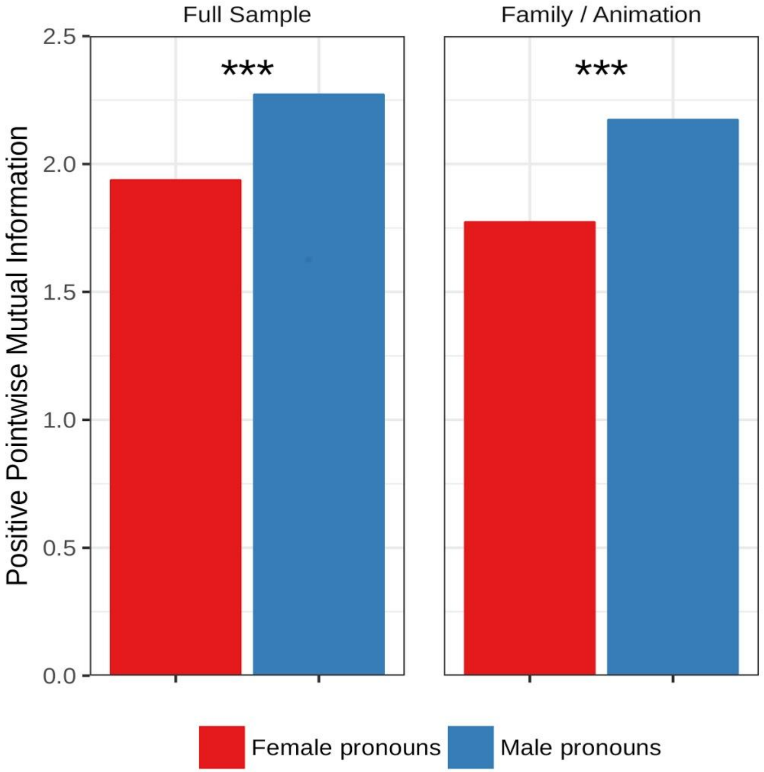
Test de significancia

Ejemplo: Odd ratio

Contingency table

	c	not c	total
w_f	c_f	nc_f	c_f+nc_f
w_m	c_m	nc_m	c_m+nc_m

“*brilliance = male*” stereotype



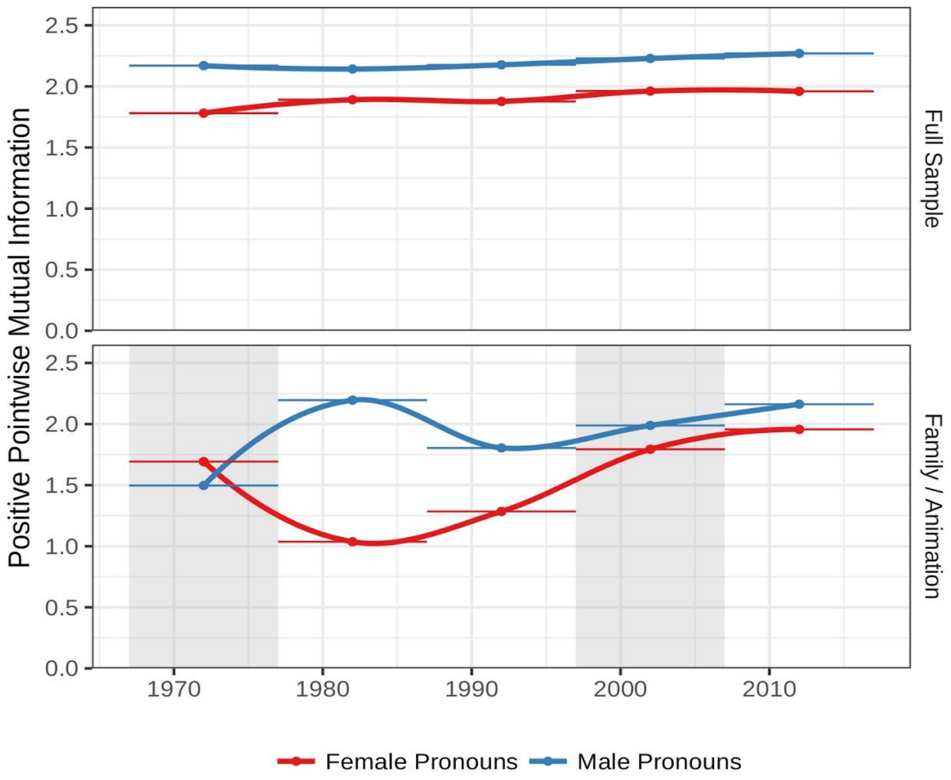
$$\Delta PMI = log\left(\frac{p(c|w_m)}{p(c|w_f)}\right) = 0.33$$

$$\frac{p(c|w_m)}{p(c|w_f)} = 1.26$$

$$\Delta PMI = 0.4$$

$$\frac{p(c|w_m)}{p(c|w_f)} = 1.32$$

“brilliance = male” stereotype



Roles estereotipados

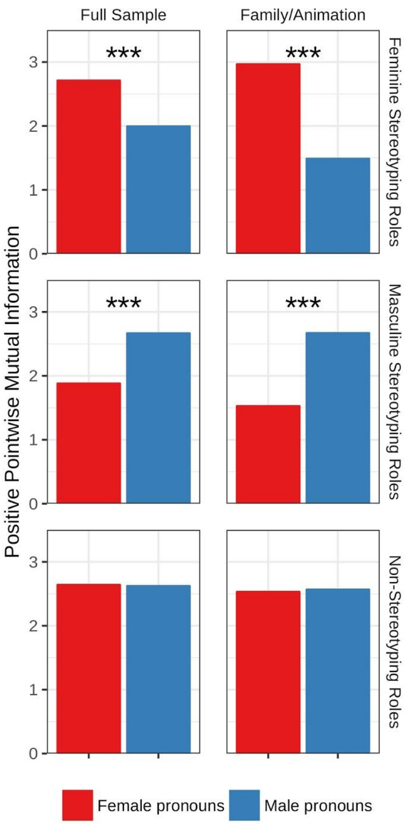
- **Feminine** stereotyping roles
  dancer, decorator, designer, dietician, florist, homemaker, housekeeper, model, nanny, typist...

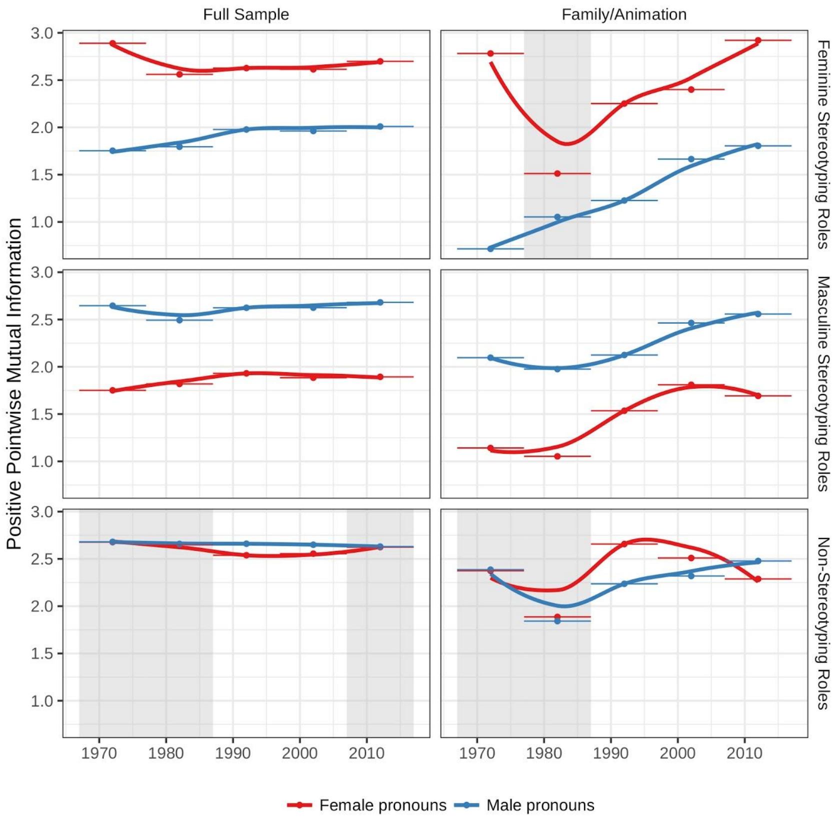
- **Masculine** stereotyping roles
  engineer, programmer, physicist, architect, detective, pilot, firefighter, inventor, mechanic, officer…

- **non-**stereotyping roles
  assistant, cashier, editor, poet, reporter, worker, doctor, lawyer, servant...

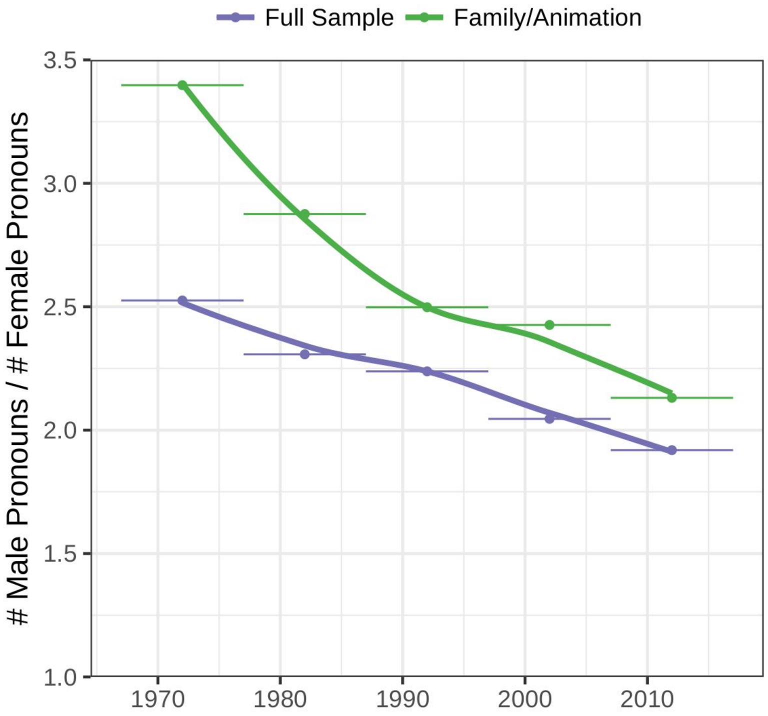
A. P. Lenton, C. Sedikides, M. Bruder, Sex Roles 60, 269 (2009).

Stereotyping roles

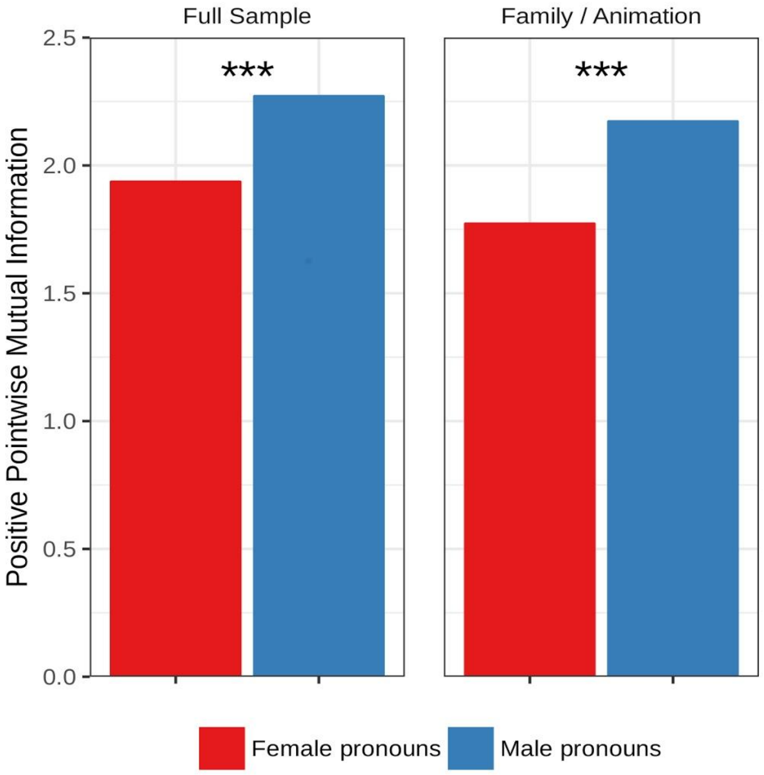


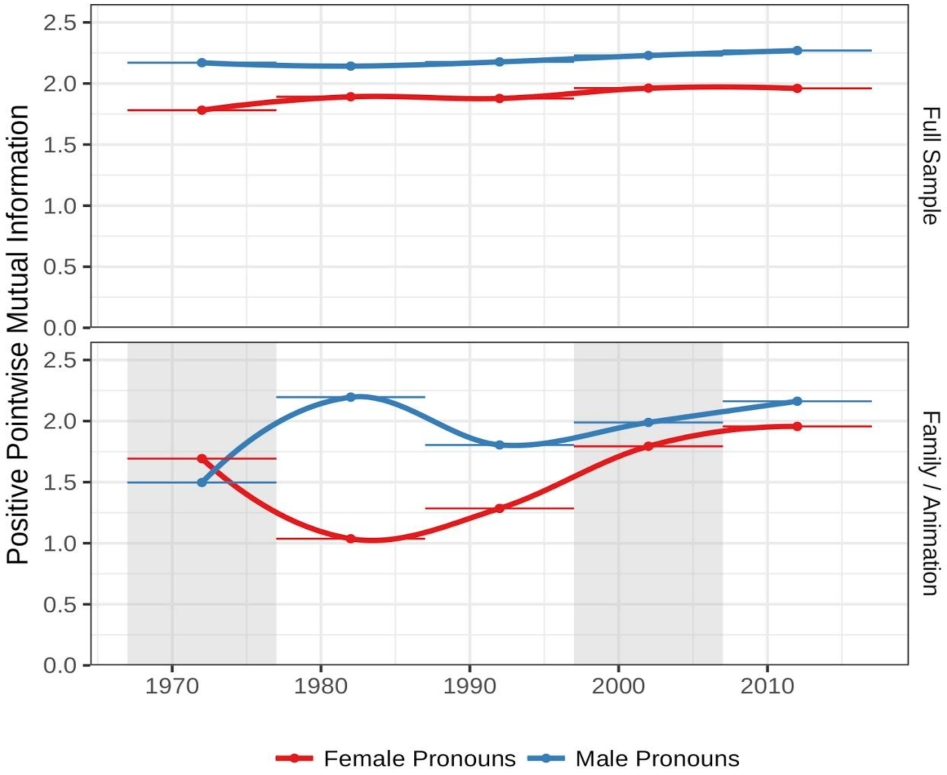


Frecuencias:



Gálvez R., Tiffemberg V. and Altszyler E. (2018) Half a century of stereotyping associations between gender and intellectual ability in films


Full Sample
Family / Animation
***
***
Positive Pointwise Mutual Information
2.5
2.0
1.5
1.0
0.5
0.0
Female pronouns
Male pronouns


Positive Pointwise Mutual Information
2.5
2.0
1.5
1.0
0.5
0.0
1970
1980
1990
2000
2010
Full Sample
Family / Animation
Female Pronouns
Male Pronouns

Son las películas las que realmente han estado moviendo todo en Estados Unidos desde que fueron inventadas. Te muestran qué hacer, cómo hacerlo, cuándo hacerlo y cómo sentirse al respecto.

Andy Warhol