



Introducción al análisis computacional de texto en estudios organizacionales

Francisco Olivos

Profesor Asistente

Departamento de Sociología y Política Social

Lingnan University, Hong Kong

10 de Junio, 2023

Debes analizar el contenido de 100 correos electrónicos enviados por la Facultad de RRLL a los profesores en el último año.

Cómo lo harías?



Debes analizar el contenido de 100 correos electrónicos enviados por la Facultad de RRLL a los profesores en el último año.

Cómo lo harías?

Análisis de contenido

Codificación manual

Grounded theory

Análisis temático



Ahora, debes analizar el contenido de 1,100,000 correos electrónicos enviados a los profesores en los últimos 10 años.

Cómo lo harías?



Ahora, debes analizar el contenido de 100,000 correos electrónicos enviados a los profesores en los últimos 10 años.

Cómo lo harías?

Lo mismo... Pero seleccionamos una muestra aleatoria de 100.






Métodos computacionales nos pueden asistir en analizar este gran volumen de información sin tener que seleccionar una muestra

```
elif _operation == "MIRROR_X":
    mirror_mod.use_x = False
    mirror_mod.use_y = True
    mirror_mod.use_z = False
elif _operation == "MIRROR_Z":
    mirror_mod.use_x = False
    mirror_mod.use_y = False
    mirror_mod.use_z = True

#selection at the end -add back the deselected mirror modifier object
mirror_ob.select= 1
modifier_ob.select=1
bpy.context.scene.objects.active = modifier_ob
print("Selected" + str(modifier_ob)) # modifier ob is the active ob
#mirror_ob.select = 0
name = bpy.context.selected_objects[0]
#bpy.data.objects[name].select = 1
```



Métodos computacionales nos pueden asistir en analizar este gran volumen de información sin tener que seleccionar una muestra

Pero no solo eso, también:

- ✓ **Manteniendo reproducibilidad**
- ✓ **Reduciendo errores humanos**
- ✓ **Menos efecto investigador**
- ✓ **Abriendo posibilidades de análisis particulares a cada técnica empleada**

```
elif _operation == "MIRROR_X":
    mirror_mod.use_x = False
    mirror_mod.use_y = True
    mirror_mod.use_z = False
elif _operation == "MIRROR_Z":
    mirror_mod.use_x = False
    mirror_mod.use_y = False
    mirror_mod.use_z = True

#selection at the end -add back the deselected mirror modifier object
mirror_ob.select= 1
modifier_ob.select=1
bpy.context.scene.objects.active = modifier_ob
print("Selected" + str(modifier_ob)) # modifier ob is the active ob
#mirror_ob.select = 0
name = bpy.context.selected_objects[0]
#bpy.data.objects[name].select = 1
```

Para tener en mente...

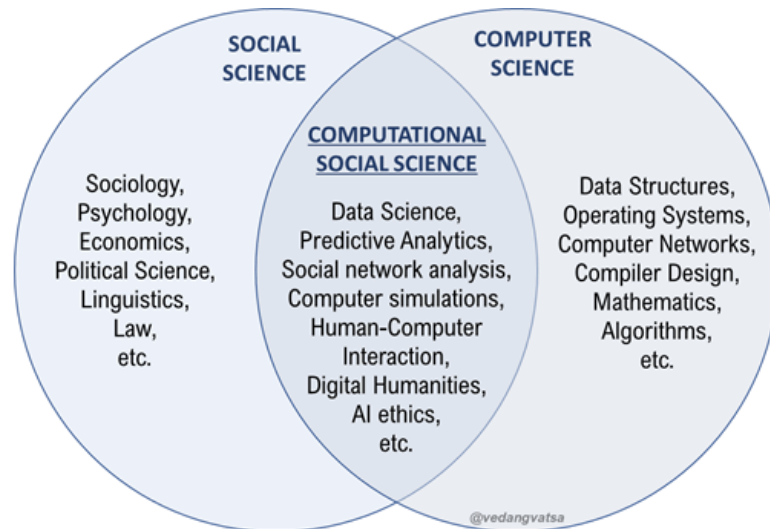
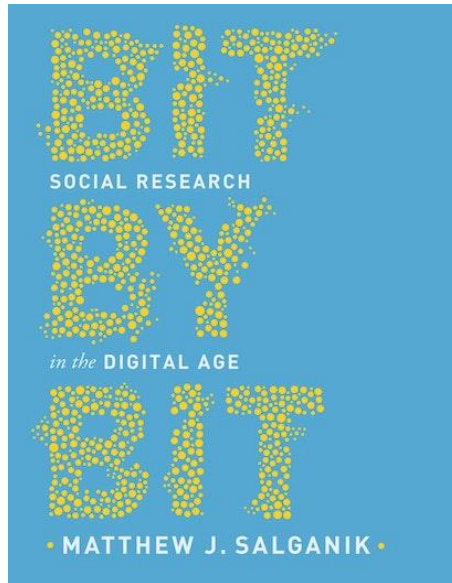
- Los métodos computacionales **no reemplazan ni al investigador ni a métodos cualitativos** tradicionales en el análisis de texto, sino que más bien son complementos y abren posibilidades de análisis.
- Siempre requerirá de la validación e interpretación de los investigadores.



Contenido

- Qué son los métodos computacionales?
- Cómo se aplican al análisis de texto?
- Ejemplo en estudios organizacionales
- Métodos supervisados
 - ✓ Análisis basados en diccionarios
- Métodos no supervisados
 - ✓ Modelamiento de tópicos



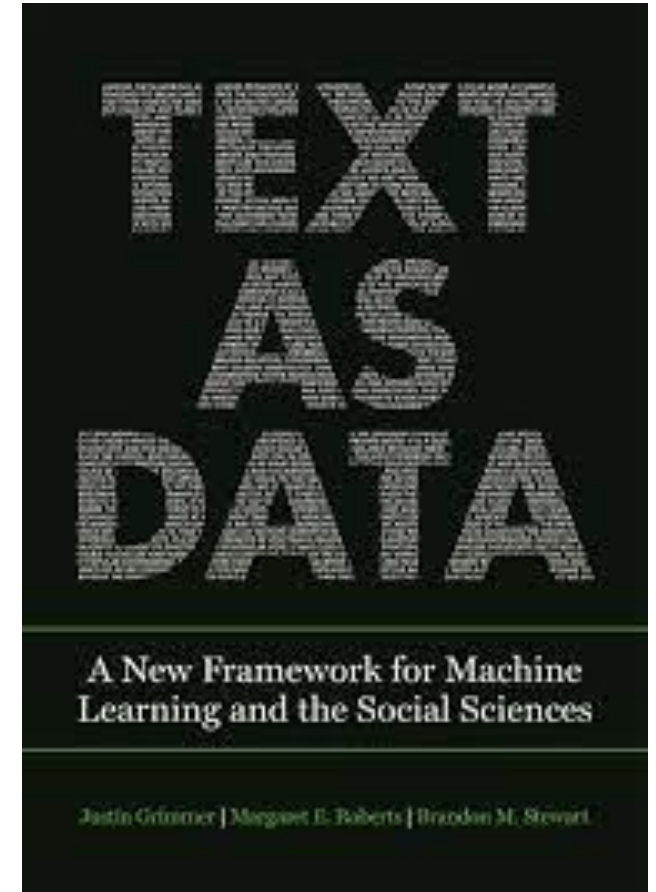


Qué son las ciencias sociales computacionales?

- Usar **ciencias de la computación** para “aproximarnos” a fenómenos sociales.
- “**Aproximarnos**” incluye generar datos, analizarlos, simular computacionalmente, visualizar resultados, etc.
- **El concepto no es una novedad.** Softwares de análisis estadístico o cualitativos ya son aproximaciones computacionales.
- **Lo nuevo:** complejidad, automatización y disponibilidad de datos.

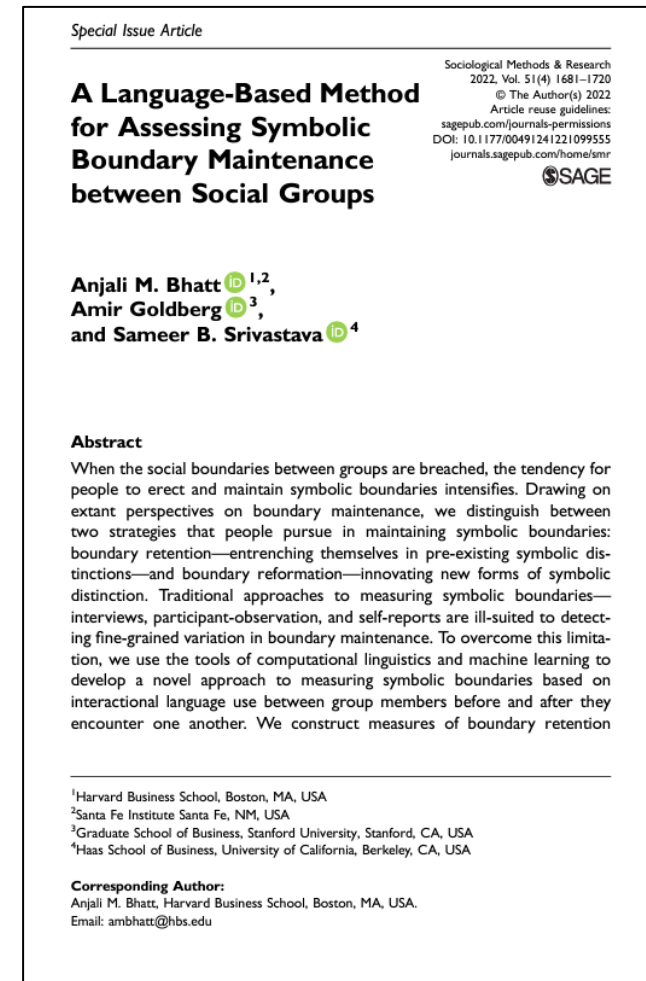
Cómo se aplican al análisis de texto?

- El **análisis de texto computacional** aplica técnicas del procesamiento de lenguaje natural y machine learning.
- En general, el análisis parte de los textos crudos y se **transforman en datos cuantitativos**, los que se analizan luego con técnicas estadísticas.



Ejemplo en estudios organizacionales

1. Se enfocan en el contexto de **fusions y adquisiciones organizacionales** (M & A).
2. El estudio usa datos de un **banco de tamaño medio** en EEUU (n1 = 306) que compra e integra dos **bancos regionales más** (n2 = 247, n3 = 51) en dos años.
3. La empresa les entregó todos los **correos internos** (alrededor de 1.5 millones de correos), incluyendo metadata y contenido adjunto.}
4. Los autores usan **un diccionario virtual** para evaluar en qué extensión las comunicaciones se ajustan al estilo del grupo, de un grupo particular y en un tiempo específico.
5. **Los resultados muestran que:**
 - a) Los límites simbólicos persisten hasta 18 meses después de la fusión;
 - b) Empleados adquiridos tienen mayor redefinición de límites y menos retención que los empleados que adquieren la empresa;



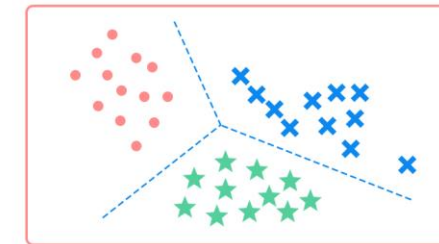
Métodos supervisados y no supervisados

- Son las dos aproximaciones en las que se **clasifican las técnicas**.
- Un **método supervisado** parte de un input dado por el investigador (i.e., texto pre-clasificado), mientras que en los **no supervisado** se aplican algoritmos de aprendizaje para detectar patrones inductivamente.
- En este workshop **introduciremos dos técnicas**:
 - Supervisado: análisis de sentimientos.
 - No supervisado: modelamiento de tópicos.



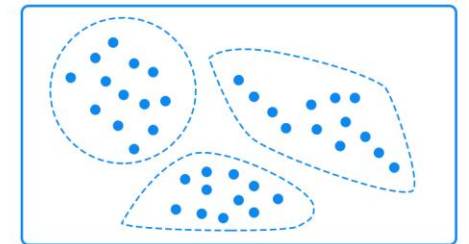
Supervised vs. Unsupervised Learning

Classification



Supervised learning

Clustering

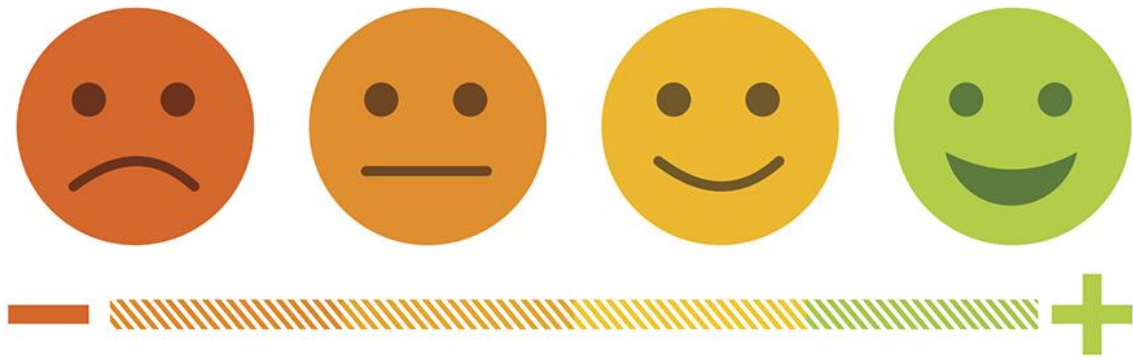


Unsupervised learning

Análisis de sentimientos

- <https://www.social-searcher.com/>





Análisis de sentimientos

- Es un tipo de **análisis supervisado** de una categoría más amplia llamada análisis basado en diccionarios.
- Hay “**diccionarios**” **disponibles** internet que nos permiten clasificar las palabras de un texto.
- El proceso de **creación de los diccionarios** es ya una contribución científica.
- Los diccionarios de sentimientos son un set de **palabras asociadas a distintos sentimientos**.
- Así podemos obtener una métrica de los sentimientos en nuestros textos **basados en una preclasificación**.

Ejemplo de diccionario

- Linguistic Inquiry Word Count (LIWC)
- Malas noticias (pero una oportunidad): **pocos diccionarios en español.**
- Posibilidad de **traducción** automatizada también.

LIWC			LIWC Cont.		
Category	Example	T-statistics	Category	Example	T-statistics
Linguistics Processes			Negative emotion	hurt, ugly, nasty	6.49***
Words > 6 letters		-3.41**	Anxiety	fearful, nervous	2.37
Dictionary words		9.60****	Anger	hate, kill, annoy	5.30***
Total function words		8.98****	Sadness	cry, grief, sad	3.54***
Personal pron.	I, them, her	7.07****	Cognitive process	cause, ought	6.09***
1st pers singular	I, me, mine	9.83****	Insight	think, know	0.11
1st pers plural	we, us, our	-2.38	Causation	effect, hence	0.93
2nd person	you, your, thou	-0.91	Discrepancy	should, would	5.53***
3rd pers singular	she, her, him	3.63**	Tentative	maybe, perhaps	5.95***
3rd pers plural	their, they'd	2.47	Certainty	always, never	4.02***
Impersonal pron.	it, it's, those	7.07****	Inhibition	block, constrain	0.32
Articles	a, an, the	4.13***	Inclusive	with, include	4.74 ***
Common verbs	walk, went, see	6.27***	Exclusive	but, without	7.53 ****
Auxiliary verbs	am, will, have	5.76***	Perceptual process		1.93
Past tense	went, ran, had	8.70****	See	view, saw, seen	1.68
Present tense	is, does, hear	4.00***	Hear	listen, hearing	-0.88
Future tense	will, gonna	5.84***	Feel	feels, touch	1.94
Adverbs	very, really	7.92****	Biological process		4.22***
Prepositions	to, with, above	7.62****	Body	cheek, spit	5.02***
Conjunctions	and, whereas	4.59***	Health	clinic, flu, pill	1.51
Negations	no, not, never	1.71	Sexual	horny, incest	-0.61
Quantifiers	few, many, much	2.98*	Ingestion	dish, eat, pizza	4.37***
Numbers	second, thousand	-3.68**	Relativity	area, bend, exit	9.52 ****
Swear words	damn, piss, fuck	5.53***	Motion	arrive, car	3.07*
Spoken Categories			Space	down, in, thin	8.87****
Assent	agree, OK, yes	7.05****	Time	end, until	5.87***
Nonfluency	er, hm, umm	1.41	Personal Concerns		
Filters	blah, imean		Work	job, majors	0.05
Psychological			Leisure	chat, movie	2.97*
Social process	mate, talk, child	0.10	Achievement	earn, win	-1.22
Family	son, mom, aunt	2.24	Home	family, kitchen	3.37**
Friends	buddy, neighbor	2.10	Money	audit, cash	0.23
Humans	adult, baby, boy	0.89	Religion	church, altar	-0.77
Affective process	happy, cry	3.55**	Death	bury, coffin	0.49
Positive emotion	love, nice, sweet	0.08			

Table 1. Two-sample T-test statistics of linguistic variables between geo-locator and non-locators. Significant differences of each LIWC attribute are indicated in the third column. (*p < 0.01, **p < 0.001, ***p < 0.0001, ****p < 1e-10)

Hagámoslo!

RECLAMOS

CL

Comentarios en línea

Site de no consentimiento

Buscar

Buscar

Ingresar Reclamo

Ranking

Cielo e Inferno

Directorio

Ingresar Reclamo

Ranking

Uber Eats Banco Falabella

Rappi Pullman Bus Cupoquick Blue Express Uber Isabella Moda Falabella Lider Walmart Daffi Mercadolibre Wom Falabella.com Mundo Shein Entel Comiquem Pedidos Ya Lippi Cusimilco Costa Best Evertrip Movistar Sencillo Odeon KFC Ripley Fransa BancoEstado Meryda Tomsa Widesp Absorbible OpenEnglish Gimnasticaoncl Naga Kupacul Did Mero Lippas MercadPago Minorista Santiago MallAmericas Hise

MAPFRE SEGUROS

Me obligan a aceptar la Indemnización Directa

En La Florida, Santiago, Chile siendo el 9 de July del 2023

El 30.11.2022 tuve un accidente en mi vehículo, SINISTRO 8012200005793 por lo que activé el seguro de la compañía Mapfre. Realicé todos los procedimientos habituales y me costó casi un mes conseguir hora para que

STAFF CHILE

No cumplimiento de contrato

En Renca, Santiago, Chile siendo el 13 de June del 2023

contrate los servicios de Staff Chile para que pudieran resolver jurídicamente el endeudamiento que tuve en plena pandemia, el contrato se realizó el 05 de noviembre de 2020 y en el especificaba la representación y

Siendo el 24 de June del 2023

Tuve que ingresar un reclamo a sams y aún no hay resultado y la empresa no se ha comunicado conmigo.

FALABELLA LA DEHESA

Incumplimiento en entrega y pésima atención a cliente

En Lo Barnechea, Santiago, Chile siendo el 13 de June del 2023

El día 30 de mayo del presente año, compré un televisor Old marca Samsung, por una suma de más de un millón de pesos, en la tienda Falabella La Dehesa, con despacho para el día sábado 10 de junio, artículo que fi

Siendo el 24 de June del 2023

Después de mas de 20 días su respuesta fue devolución del dinero, y el reclamo en el SERIAC ni lo peccaron.

AUTOMOTORA ROSSELOT

Vehículo nuevo

En La Gracia, Santiago, Chile siendo el 4 de July del 2023

Estimados señores Tiggo Espero que esta carta les encuentre bien. Me dirijo a ustedes con respecto a la unidad Tiggo 9 modelo 1999 que compré el 1999 de la cual me encontré, lamentablemente me

MÁS RECLAMADOS

Uber Exp 41%

Banco Falabella 21%

Rappi 22%

Pullman Bus 21%

Cupoquick 17%

Blue Express 14%

Uber 14%

Isabella Moda 14%

Falabella 13%

Lider Walmart 13%

Daffi 13%

LO QUE LA GENTE ESTÁ MIRANDO

Widesp 24%

Uber Exp 24%

Uber 19%

Lider Exp 18%

Blue Express 14%

Forti 14%

Isabella Moda 14%

MercadPago 14%

Hogar de Cristo 13%

CompuNet 12%

Lipasa 10%

Falabella.com 10%

LOS QUE RESUELVEN

Mapfre 81.2%

Seguros 81.2%

Comerciales 81.2%

Ultramar 74.1%

Walmart 64.7%

Servicios 62.7%

Financ 61.8%

Dominio 61.8%

Proveedores 61.2%

Chile 61.7%

Ripley 61.8%

Casa 60.8%

PCBomero 61.4%

Bucanet 61.4%

Corporación 61.7%

Pay 61.8%

Tiggo 61.8%

Hogar 61.8%

Farmacia 61.8%

Antes 61.8%


Rappi 61.8%

Original Research Article

Journal of Research in Crime and Delinquency
1–41
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00224278221101119
journals.sagepub.com/home/jrc

SAGE

Citizen Complaints as an Accountability Mechanism: Uncovering Patterns Using Topic Modeling

Francisco Olivos¹ ,
Patricio Saavedra²,
and Lucia Dammert³

Abstract
Objectives: Citizen complaints are considered by policing researchers as an indicator of police misconduct, and a proxy of police-community relations.

and EU-based studies tend to focus on sustained com- by official agencies and officer-based correlates. Using nerros, the Chilean militarized police force, this study topics contained in a large set of complaints against tal platform, and (b) the change of those topics across mplainants' educational level. **Methods:** We use novel ural language processing techniques to identify latent corpus of complaints ($N=1,623$), hosted on an online o 2020. **Results:** Our findings show eight latent themes

gy and Social Policy, Lingnan University, Tuen Mun, Hong Kong
ces, University of O'Higgins, Rancagua, Chile
o de Chile, Santiago de Chile, Chile

or:
rothy Y L Wong Building, Department of Sociology and Social Policy,
Mun, Hong Kong.
e@ln.edu.hk.

Ruido por fiesta

A

Este reclamo tiene más de seis meses de antigüedad

CARABINEROS DE CHILE

RUIDO POR FIESTA

En Estación Central, Santiago, Chile

Lunes 15, January 2018

Número de Reclamo:

626 observadores

C

B

SEÑORES CARABINEROS DE CHILE

PRESENTE

Estimados, vivo en , comuna de estación central, la madrugada de esta noche el apto tuvo Ruido por fiesta hasta las 7:00 am, conserjería les reclamo y les paso multa, pero hicieron caso omiso, incluso yo personalmente llame a Carabineros pero al final nunca fueron a revisar

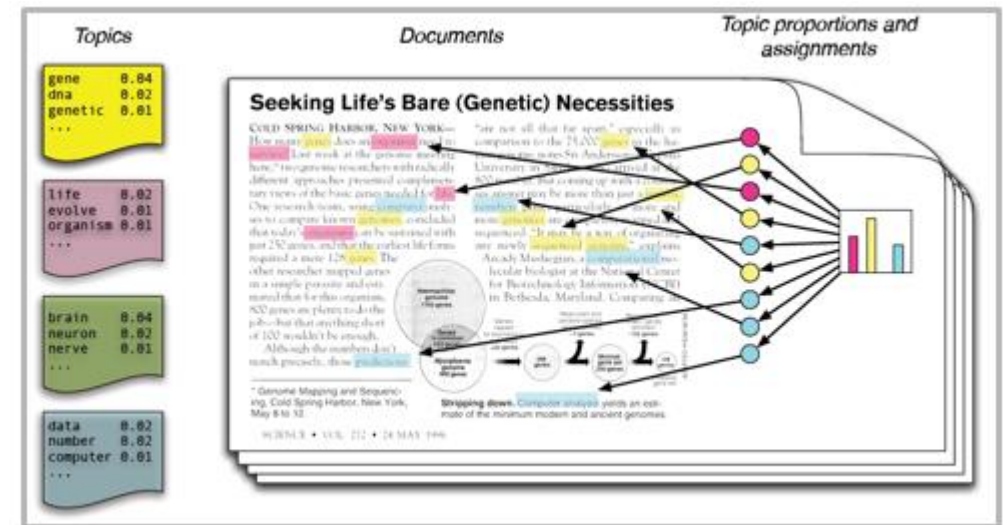
D

Download the code 01 from the repository sent by email

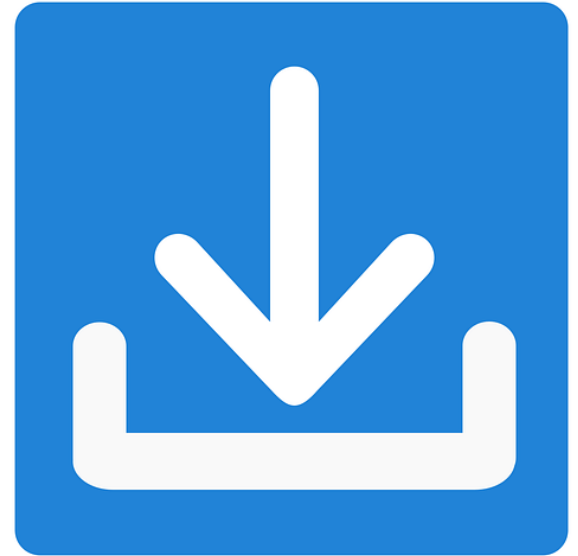


Modelamiento de tópicos

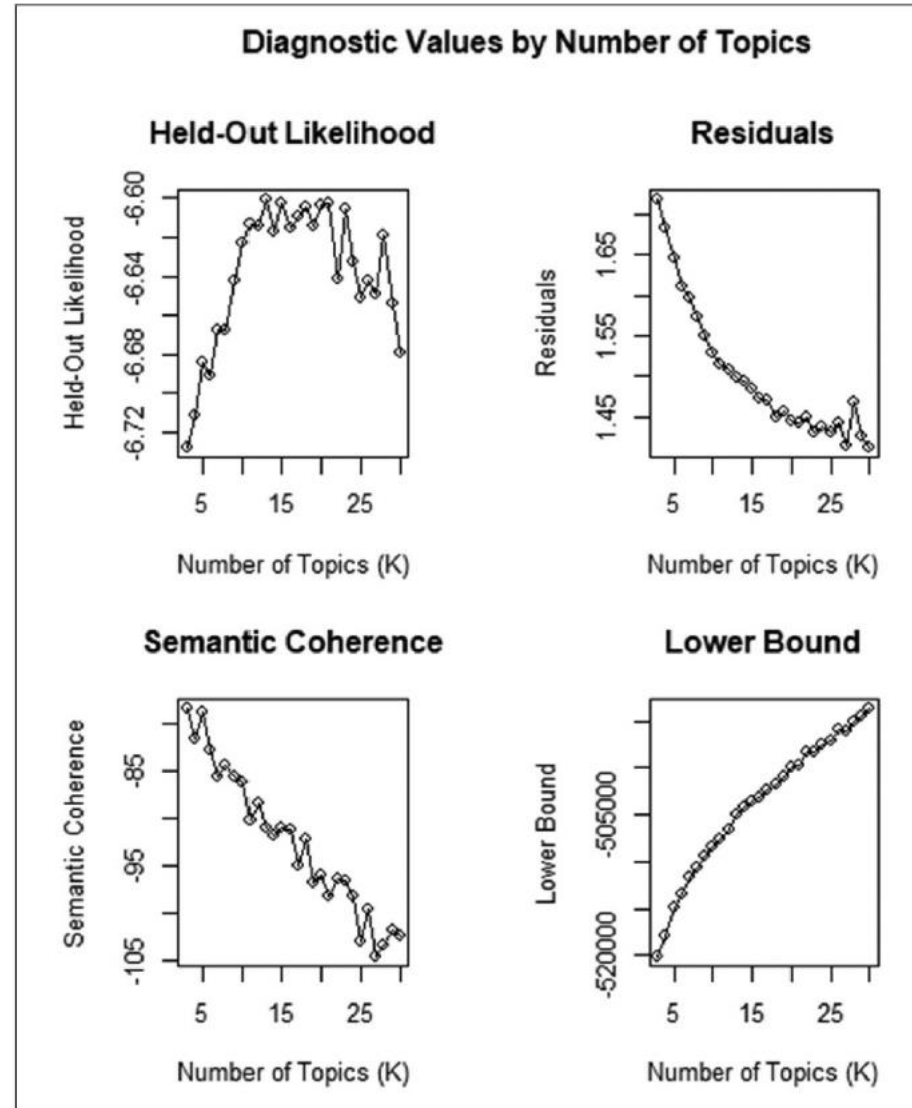
- ✓ Método **no supervisado** para identificar tópicos latentes en un conjunto de documentos.
- ✓ Se basa en la **co-ocurrencia de palabras**.
- ✓ Una palabra puede pertenecer a **distintos tópicos**.
- ✓ Podemos obtener la **prevalencia de cada tópico** en cada documento y explicarlas por características del documento (meta-data)



Download the code 02 from the repository sent by email



Métricas



Palabras más probables

Table 2. Most Probable Words per Topic.

Topic Label # Words	1 Employer	2 Traffic	3 Control	4 Noise	5 Household	6 Parking	7 Certificate	8 Generalized
1	officer	vehicle	day*	house	house	street	work	chile
2	personnel	car	precinct*	neighbors	complaint	place	years	institution
3	day	says	vehicle*	week	then	bad	thanks	persons
4	form	offense	had*	night	residency	law	answer	to be
5	present	documents	said	noises	daughter	in front	money	person
6	officers	license	corporal	does	son	cars	year	today
7	sergeant	speed	power	district	partner	to park	course	good
8	police (adjective)	day	then*	tomorrow	father	vehicles	does	respect
9	situation	going	son	alley	place	parked	santiago	type
10	precinct	moment	daughter	neighbor	mother	people	company	worst
11	procedure	stop [sign]	after*	call	while	big	[I] can	people
12	service	driver	moment*	[I] live	name	exit	[I] need	form
13	commissioned officer	court	had*	day	[they] said	[it] seems	good	day
14	chief	license	[I] said	disturbing	door	to be	hello	justice
15	corporal	then	[he/she] says	[they] do	day	pass	to have	precinct

Note: *words with spelling mistakes.

Predicción por año

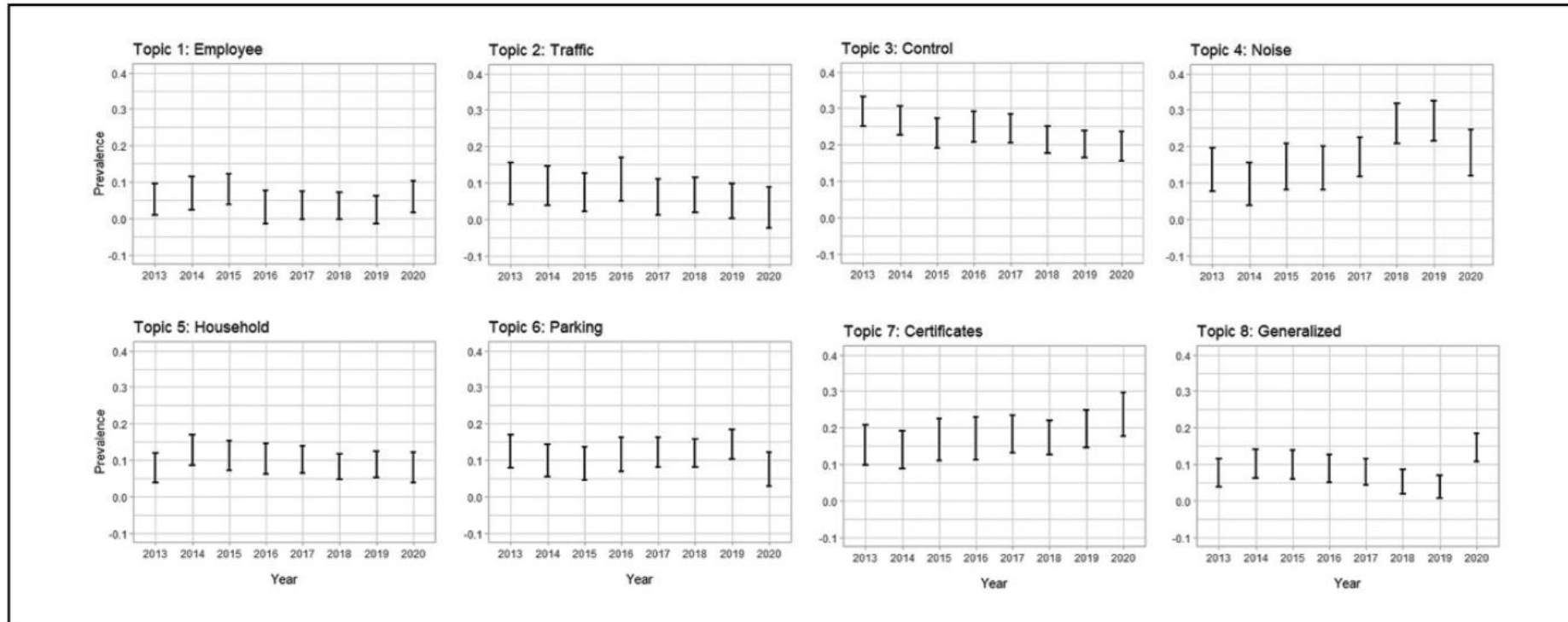


Figure 6. Prevalence associated with each topic by year. Note: Intervals of confidence at 95 percent.

Predicción por nivel educativo en base a la proporción de palabras con acento

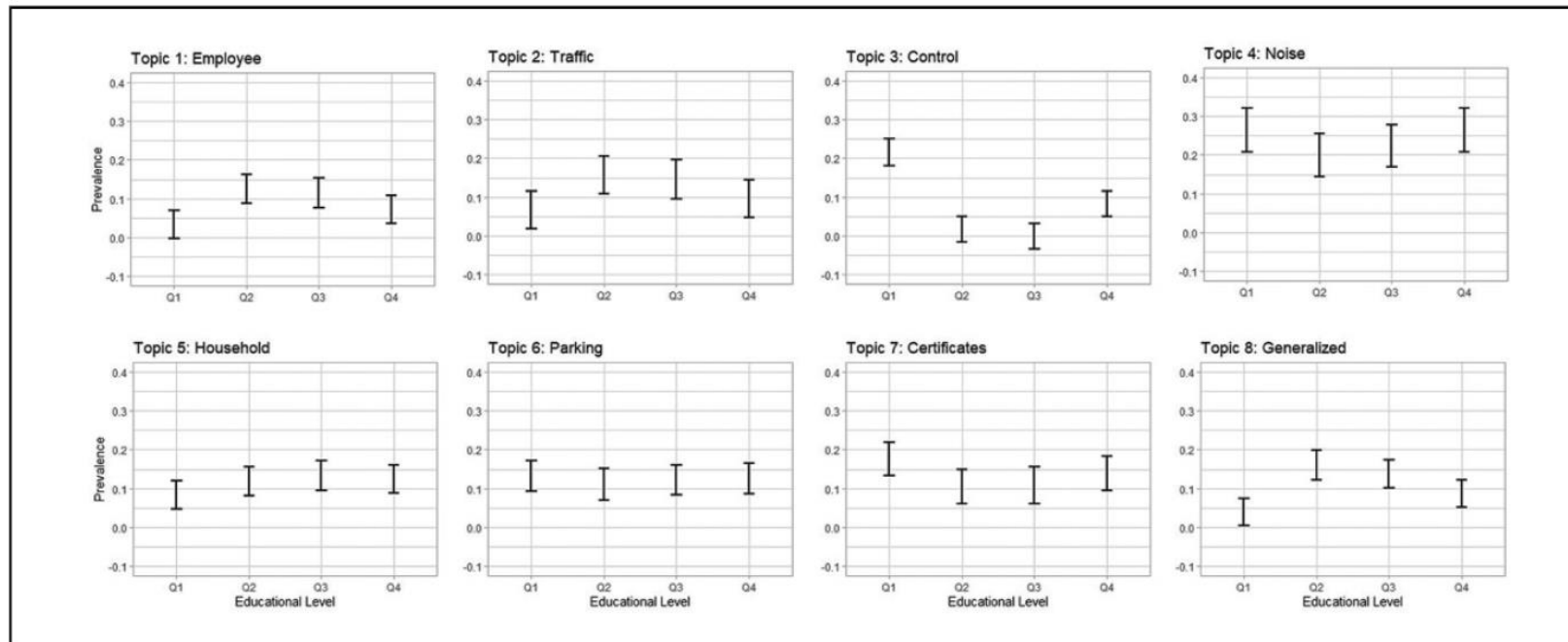


Figure 7. Prevalence associated with each topic by complainants' educational level.
Note: Intervals of confidence at 95 percent.

Gracias

franciscoolivosrave@LN.edu.hk

ResearchGate profile

www.fjolivos.com

Fjolivos en Twitter



Introducción al análisis computacional de texto en estudios organizacionales

Francisco Olivos

Profesor Asistente

Departamento de Sociología y Política Social

Lingnan University, Hong Kong

10 de Junio, 2023