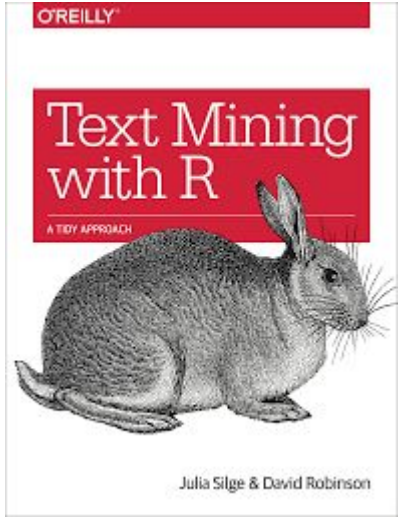
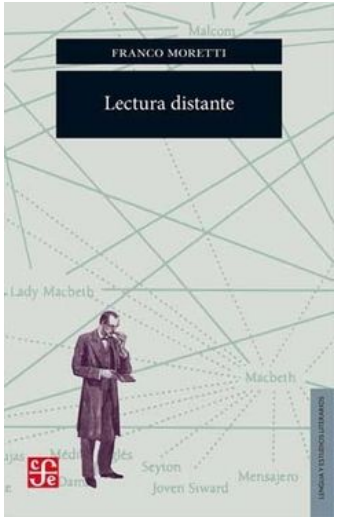
Laboratorio de datos: web scraping y Procesamiento de Lenguaje Natural

Clase 1. Fundamentos conceptuales

factor~data
EIDAES_UNSAM

Herramientas

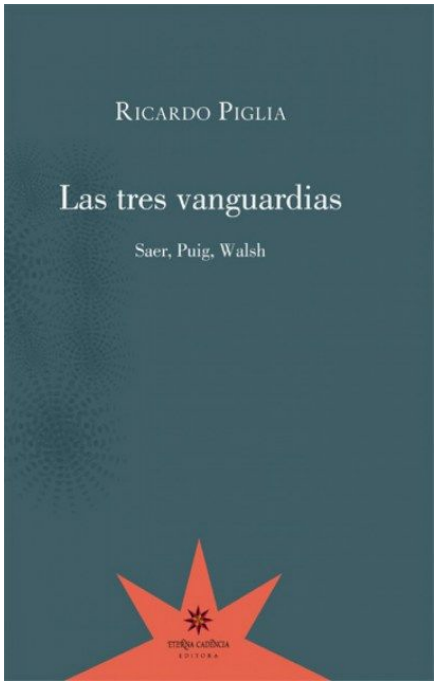




factor-data
EIDAES_UNSAM

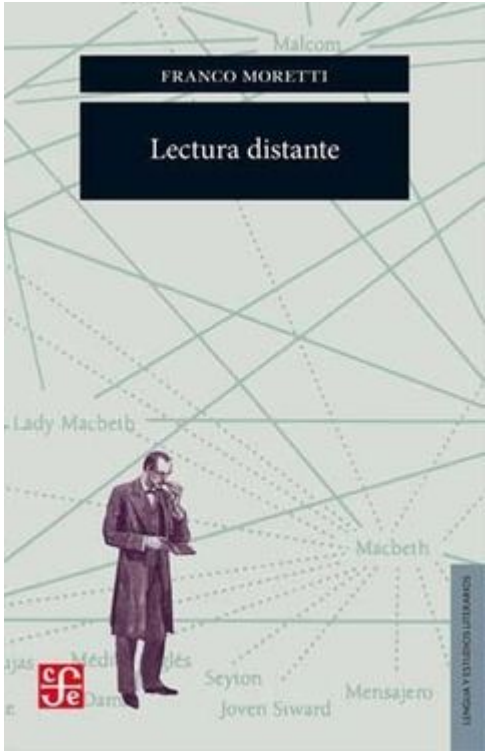
¿Qué es NLP?

“Podemos imaginar que si conociéramos el conjunto de narraciones que circulan en la Ciudad de Buenos Aires en un día conoceríamos un tipo particular de funcionamiento de esta ciudad con bastante precisión. (...) Si tuviéramos la posibilidad, fantástica, de disponer de todas esas narraciones, podríamos detectar las grandes formas, los grandes núcleos formales a partir de los cuales se construyen **los grandes relatos sociales**.” (p. 60-61)



- Todo el tiempo estamos produciendo textos
    - Charlas
    - Entrevistas
    - Posts
    - Redes sociales

- ¿Cómo podemos aprovechar esos textos en la investigación?



“Mucha gente ha leído más y mejor que yo, por supuesto, pero eso tampoco basta: aquí hablamos de cientos de lenguas y literaturas. Todo indica que leer ‘más’ no es la solución. En especial, porque hemos comenzado a descubrir (...) la enormidad de lo no leído (...) pero el punto es que existen treinta mil novelas británicas del siglo XIX o cuarenta o cincuenta, sesenta mil: nadie lo sabe a ciencia cierta, nadie las ha leído, nadie las leerá. Y, además, hay novelas francesas, chinas, argentinas, estadounidenses.

Leer ‘más’ siempre es bueno, pero no es la solución” (p. 59)

El problema de los datos

Lejos de querer ofrecer una respuesta general y rigurosa. Creo que en algunas áreas se usa como grados de precisión. Cualitativo "si esto sube tal cosa baja", semicuantitativo "alrededor de 5" o un "cambio de un orden de magnitud". Cuantitativo, "4.2 con una desviación de 0.3".

Por el propósito de su uso: buscás la comprensión de sentido (?)

1 1 3

La definición más general me parece que sería "una realización de una variable aleatoria discreta cuyo recorrido tiene una cardinalidad pequeña". Lo bueno es que cuando me preguntes cuánto es pequeño te puedo responder "y... es complejo"

No se.. . La verdad tendría que pensarlo un poco. En física me da la sensación que unx siempre tiende a querer ponerle número a las cosas poder hacer predicciones (al menos decir si algo es grande o chico respecto a otra cosa o si crece o no, si ta lejos o cerca)...

Una manifestación verbal o escrita, dependiendo de la herramienta utilizada, sobre un fenómeno determinado. Pero ese dato, así en crudo, solo es un relato de una pareció de la realidad social, por eso es fundamental determinar en contexto.

Sin releer nada, diría que un dato es una construcción teórico-empírica en el marco de una investigación científica. Es cuali si es narrativo, "natural", a veces inductivo (Burawoy es etnografo + tirando a deductivo, x ej.)

para mi es sinonimo de variable categorica, no-ordinal.

Te diría que lo cuali no es el dato sino la metodología. Un mismo dato puede ser cuali o cuanti en función del método de investigación que sigas. Decir "Juan es varón" parece a priori cuali, sin embargo usar ese dato para contar varones y mujeres en un grupo sería cuanti.

Me quedé pensando. La definición que te dije seguramente "funciona", pero probablemente le falta rigurosidad. Dejame intentar algo más formal. Una VA es una función de Omega a R, no? Sea D: Omega^2 -> R tal que D(w_1,w_2)=X(w_1)-X(w_2) (sigue)

Me sumo, aunque difícil responder sin aclarar a que nivel / momento de construcción de dato te referis. Si es en términos generales: dato construido a través de un análisis que tiene a la interpretación por centro.

El problema de los datos

Lejos de querer ofrecer una respuesta general y rigurosa. Creo que en algunas áreas se usa como grados de precisión. Cualitativo "si esto sube tal cosa baja", semicuantitativo "alrededor de 5" o un "cambio de un orden de magnitud". Cuantitativo, "4.2 con una desviación de 0.3".

No se.. . La verdad tendría que pensarlo un poco. En física me da la sensación que unx siempre tiende a querer ponerle número a las cosas poder hacer predicciones (al menos decir si algo es grande o chico respecto a otra cosa o si crece o no, si ta lejos o cerca)...

La definición más general me parece que sería "una realización de una variable aleatoria discreta cuyo recorrido tiene una cardinalidad pequeña". Lo bueno es que cuando me preguntes cuánto es pequeño te puedo responder "y... es complejo"

para mi es sinonimo de variable categorica, no-ordinal.

Me quedé pensando. La definición que te dije seguramente "funciona", pero probablemente le falta rigurosidad. Dejame intentar algo más formal. Una VA es una función de Omega a R, no? Sea D: Omega^2 -> R tal que D(w_1,w_2)=X(w_1)-X(w_2) (sigue)

El problema de los datos

Por el propósito de su uso: buscás la comprensión de sentido (?)

1 1 3

Una manifestación verbal o escrita, dependiendo de la herramienta utilizada, sobre un fenómeno determinado. Pero ese dato, así en crudo, solo es un relato de una pareció de la realidad social, por eso es fundamental determinar en contexto.

Sin releer nada, diría que un dato es una construcción teórico-empírica en el marco de una investigación científica. Es cuali si es narrativo, "natural", a veces inductivo (Burawoy es etnografo + tirando a deductivo, x ej.)

Te diría que lo cuali no es el dato sino la metodología. Un mismo dato puede ser cuali o cuanti en función del método de investigación que sigas. Decir "Juan es varón" parece a priori cuali, sin embargo usar ese dato para contar varones y mujeres en un grupo sería cuanti.

Me sumo, aunque difícil responder sin aclarar a que nivel / momento de construcción de dato te referis. Si es en términos generales: dato construido a través de un análisis que tiene a la interpretación por centro.

factor-data
EIDAES_UNSAM

El problema de los datos (la grieta)

- Lo “cualitativo”
  - Constructivista/ relativista
  - Atención al detalle
  - Lectura “cercana”
  - Escala pequeña
  - Fuerte presencia de lo subjetivo
  - Poco escalable
  - Centrada en el “sentido”
- Lo “cuantitativo”
  - Positivista (?)
  - Importancia de la generalidad
  - Lectura “distante”
  - Gran escala
  - Intento de reproducibilidad
  - Escalable
  - Centrada en la “cantidad”

El problema de los datos (la grieta)

- Lo “cualitativo”
  - Constructivista/ relativista
  - Atención al detalle
  - Lectura “cercana”
  - Escala pequeña
  - Fuerte presencia de lo subjetivo
  - Poco escalable
  - Centrada en el “sentido”

Procesamiento de Lenguaje Natural

- Lo “cuantitativo”
  - Positivista (?)
  - Importancia de la generalidad
  - Lectura “distante”
  - Gran escala
  - Intento de reproducibilidad
  - Escalable
  - Centrada en la “cantidad”

Estructura tripartita del dato (Galtung, 1973)

- Un datos está constituida por la Intersección de tres elementos:

  1. caso (unidad de análisis),

  2. un atributo (variable) y

  3. un valor (categoría) de esa variable.

i	V1	V2	...	Vp
1				
2				
3				
...				
n				

El problema de los datos

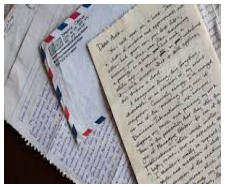
```
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 3

[1] a bailar a bailar | que la orquesta se va | sobre el fino garabato | de un tango nervioso y lerdo | se ira borrando el recuerdo | a ba
ilar a bailar | que la orquesta se va | el ultimo tango perfuma la noche | un tango dulce que dice adios | la frase callada se asoma a los
 labios | y canta el tango la despedida! | vamos! a bailar! | tal vez no vuelvas a verla nunca | y el ultimo tango perfuma la noche | y es
te es el tango que dice el adios | a bailar a bailar | que la orquesta se va! | quedara el salon vacio | con un monton de esperanzas | que
 iran camino al olvido | a bailar a bailar | que la orquesta se va!

[2] este tango nacio para bailarse | y asi hamacarse muy suavemente | oigan ustedes este compas… | es muy sencillo bailar el tango | un do
ble paso despues descanso | la media vuelta la vuelta entera | y siempre junto a la compañera | este tango nacio para bailarse | no hay qu
e quedarse mirandolo



[3] nacio en la calle quito | entre boedo y colombres | barrio de tauras de hombres | de timbas y de garitos | mi recuerdo es muy estricto
 | de proscenio un corralon | modesto fue su blason | y la dulce purretita | se lavaba la carita | en el viejo pileton | amante del variet
e | soñaba con ser artista | comenzo como corista | hasta llegar a vedette | piernas tipo mistinguette | cintura bien contorneada | anatom
ia envidiada | y un rostro angelical | para que plumas y percal | lucieran como hermanadas | siempre causo sensacion | en cine radio y tea
tro; | se volco al dos por cuatro | con sentida emocion | triunfo en television | y nadie podra dudar | fue figura consular | en todos los
 escenarios | recogio aplausos a diario | se llamaba beba bidart
```

El problema de los datos

MAS_500 Aglomerados segun tamanio	AGLOMERADO Codigo de Aglomerado	PONDERA Ponderacion	CH03 Relacion de parentesco	CH04 Sexo	CH05 Fecha de nacimiento (dia, mes y anio)
N	8	108	2	2	03/06/1990
N	8	108	3	2	29/12/2005
N	8	108	3	1	26/01/2018
N	8	108	1	2	30/03/1978
N	8	108	3	2	20/09/2009
N	8	141	1	1	26/04/1967
N	8	221	1	1	15/03/1955
N	8	221	2	2	25/04/1956
N	8	221	3	2	10/06/1994
N	8	221	1	1	22/07/1944
N	8	221	3	1	23/08/1985
N	8	309	1	1	14/06/1976
N	8	309	2	2	17/06/1978
N	8	309	3	2	20/07/1997
N	8	309	3	1	19/10/2001
N	8	309	1	2	02/01/1967
N	8	309	3	2	29/06/1982
N	8	88	1	1	15/08/1974

14/06/1976

El problema de los datos



```
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 3

[1] a bailar a bailar | que la orquesta se va | sobre el fino garabato | de un tango nervioso y lerdo | se ira borrando el recuerdo | a bailar a bailar | que la orquesta se va | el ultimo tango perfuma la noche | un tango dulce que dice adios | la frase callada se asoma a los labios | y canta el tango la despedida! | vamos! a bailar! | tal vez no vuelvas a verla nunca | y el ultimo tango perfuma la noche | y este es el tango que dice el adios | a bailar a bailar | que la orquesta se va! | quedara el salon vacio | con un monton de esperanzas | que iran camino al olvido | a bailar a bailar | que la orquesta se va!

[2] este tango nacio para bailarse | y asi hamacarse muy suavemente | oigan ustedes este compas… | es muy sencillo bailar el tango | un doble paso despues descanso | la media vuelta la vuelta entera | y siempre junto a la compañera | este tango nacio para bailarse | no hay que quedarse mirandolo


[3] nacio en la calle quito | entre boedo y colombres | barrio de tauras de hombres | de timbas y de garitos | mi recuerdo es muy estricto | de proscenio un corralon | modesto fue su blason | y la dulce purretita | se lavaba la carita | en el viejo pileton | amante del variete | soñaba con ser artista | comenzo como corista | hasta llegar a vedette | piernas tipo mistinguette | cintura bien contorneada | anatomia envidiada | y un rostro angelical | para que plumas y percal | lucieran como hermanadas | siempre causo sensacion | en cine radio y teatro; | se volco al dos por cuatro | con sentida emocion | triunfo en television | y nadie podra dudar | fue figura consular | en todos los escenarios | recogio aplausos a diario | se llamaba beba bidart
```

MAS_500 Aglomerados segun tamanio	AGLOMERADO Codigo de Aglomerado	PONDERA Ponderacion	CH03 Relacion de parentesco	CH04 Sexo	CH05 Fecha de nacimiento (dia, mes y anio)
N	8	108	2	2	03/06/1990
N	8	108	3	2	29/12/2005
N	8	108	3	1	26/01/2018
N	8	108	1	2	30/03/1978
N	8	108	3	2	20/09/2009
N	8	141	1	1	26/04/1967
N	8	221	1	1	15/03/1955
N	8	221	2	2	25/04/1956
N	8	221	3	2	10/06/1994
N	8	221	1	1	22/07/1944
N	8	221	3	1	23/08/1985
N	8	309	1	1	14/06/1976
N	8	309	2	2	17/06/1978
N	8	309	3	2	20/07/1997
N	8	309	3	1	19/10/2001
N	8	309	1	2	02/01/1967
N	8	309	3	2	29/06/1982
N	8	88	1	1	15/08/1974

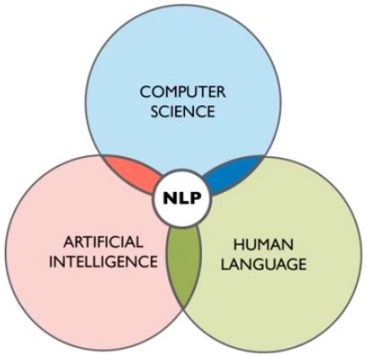
14/06/1976

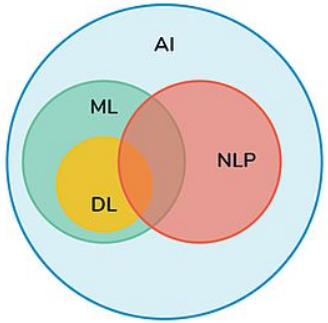
Grado de estructuración

El problema de los datos

- No estructurados
- No hay modelo predefinido
- No hay orden

- NLP => tratar de detectar patrones en estos datos no estructurados
- Área de investigación científica llamada Natural Language Processing, una subdisciplina de machine learning/ciencias de la computación que trata de emular la interpretación humana de textos.





Problemas de aplicación

- Supervisado: Clasificación de textos en categorías definidas anteriormente

- No supervisado: no hay variable dependiente. Técnicas exploratorias. Detección de temas, entrenamiento de word embeddings, etc.

Aplicaciones usuales

Clasificación - Aplicaciones

Automatización de procesos para la construcción de bases de datos de protestas

[Hanna, 2017]

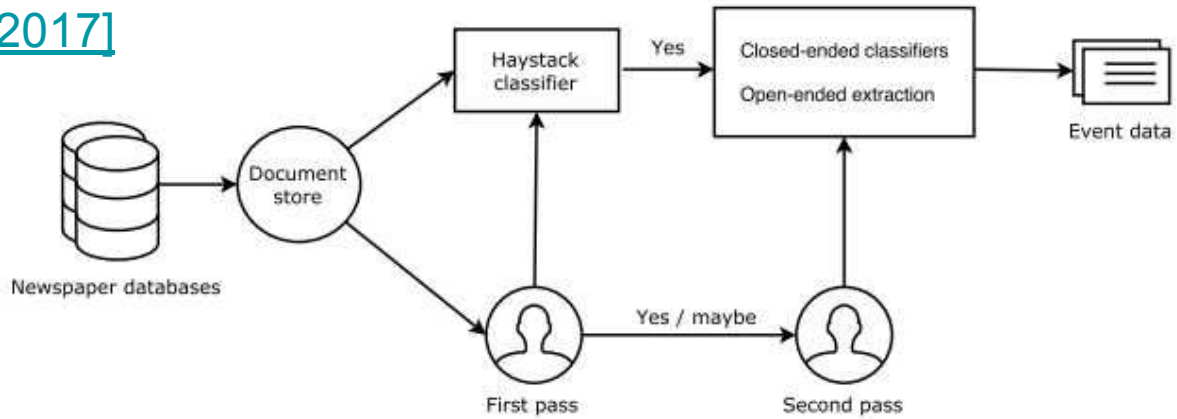


Figure 1: MPEDS pipeline with training.

Clasificación - Aplicaciones

Predicción de enfermedades mentales mediante análisis de texto

[Corcoran, Carrillo, Fernández Slezak et al, 2018]

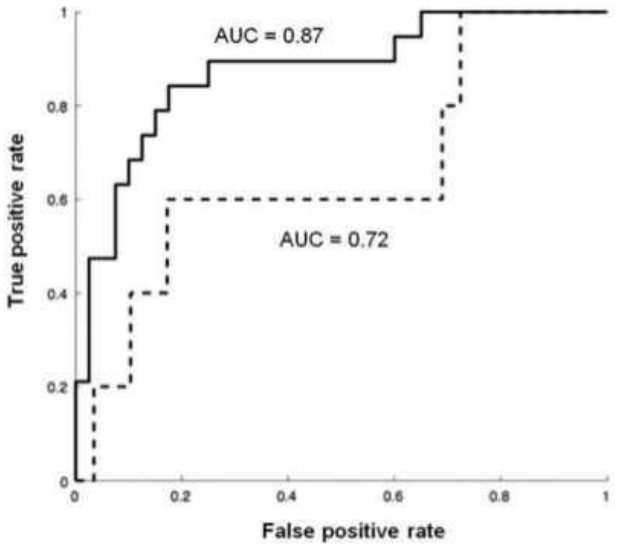


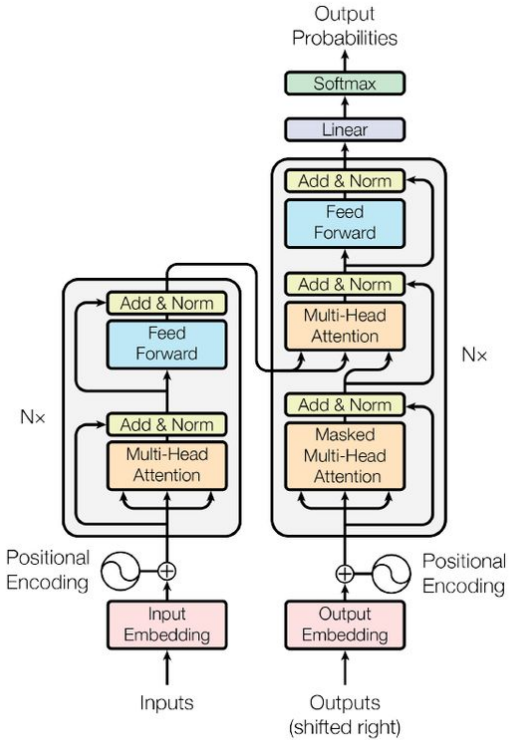
Figure 2 Receiver operating characteristics (ROC) for the University of California Los Angeles (UCLA) clinical high-risk (CHR) classifier of psychosis outcome as applied to the UCLA dataset (solid line) and to the realigned New York City (NYC) dataset (dotted line). AUC – area under the curve.

Clasificación - Aplicaciones

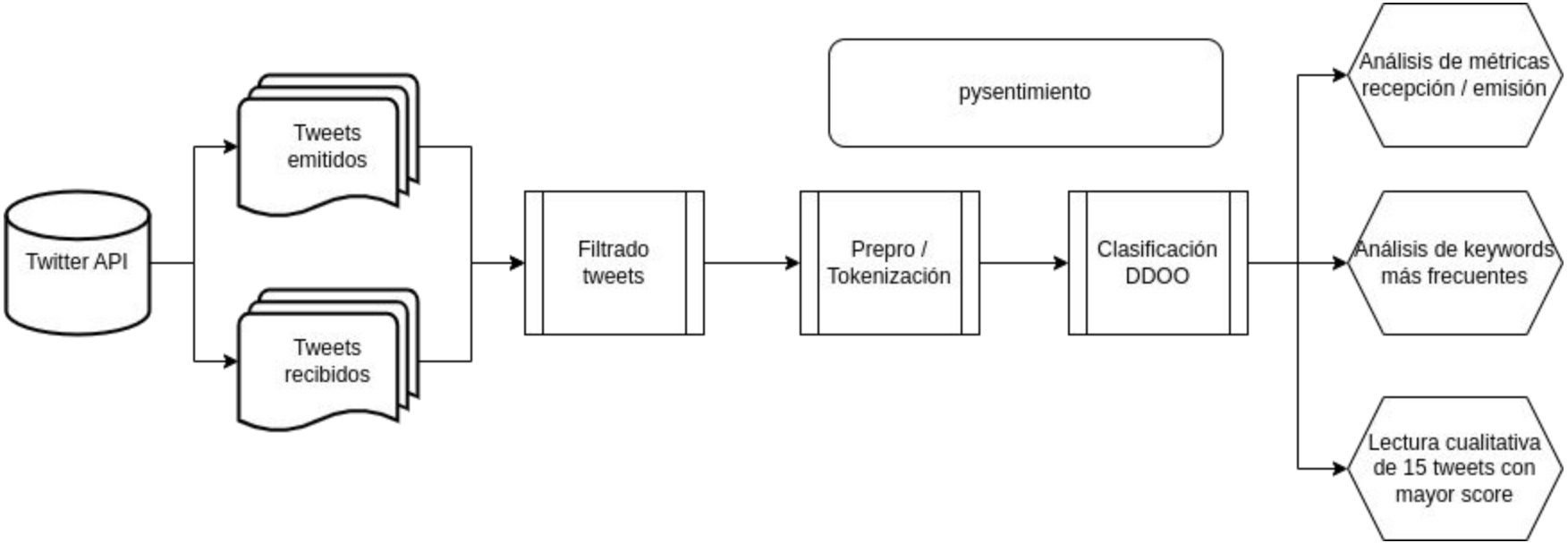
- Analizar patrones en la recepción y emisión de discursos violentos/de odio y su relación con género y orientación política

- API Twitter + Transformers pre-entrenados + LLMs

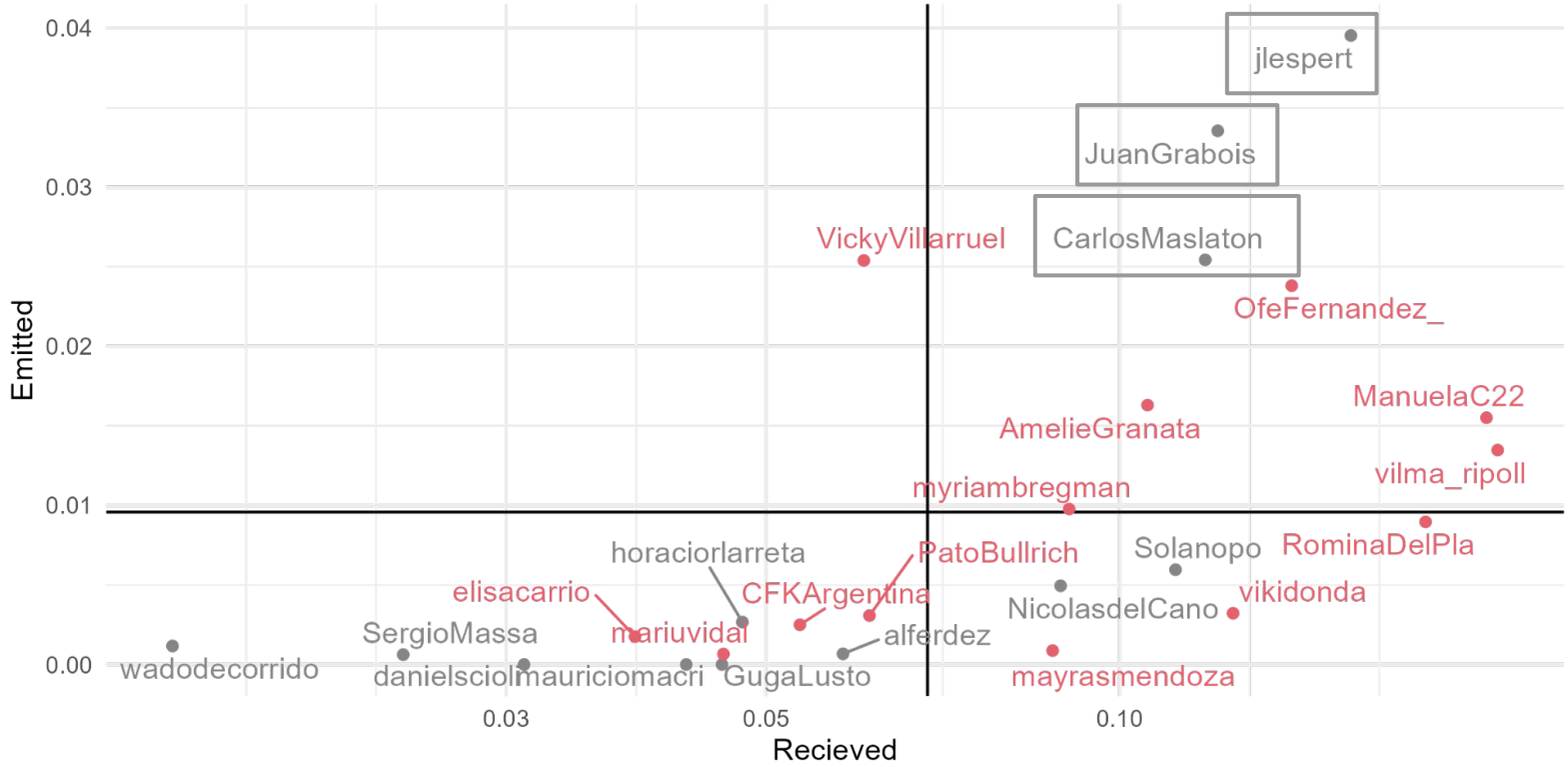
- Laia Domenech (factor~data -EIDAES-UNSAM)
- Juan |Manuel Pérez (CONICET / LIIA-UBA)
- Germán Rosati (CONICET / factor~data-EIDAES-UNSAM)
- Magalí Rodrigues Pires (FSOC-UBA)
- María Nanton (FSOC-UBA)
- Diego Kozlowski (École de bibliothéconomie et des sciences de l'information, Université de Montréal)



Clasificación - Aplicaciones



Proportion of hateful tweets received and emitted by political figures, and weighted average by total number of tweets of each political figure.



factor-data
EIDAES_UNSAM

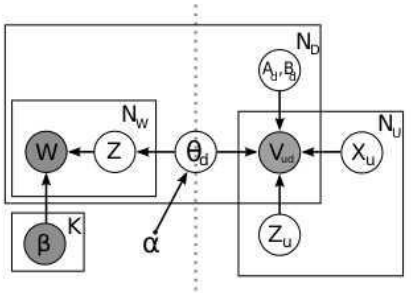
No supervisado - Aplicaciones

Posiciones ideológicas en proyectos de ley

[Gerrish y Blei, 2012]

Terrorism	Commemorations	Transportation
terrorist	nation	transportation
september	people	minor
attack	life	print
nation	world	tax
york	serve	land
terrorist attack	percent	guard
hezbolah	community	coast guard
national guard	family	substitute

Labeled topics



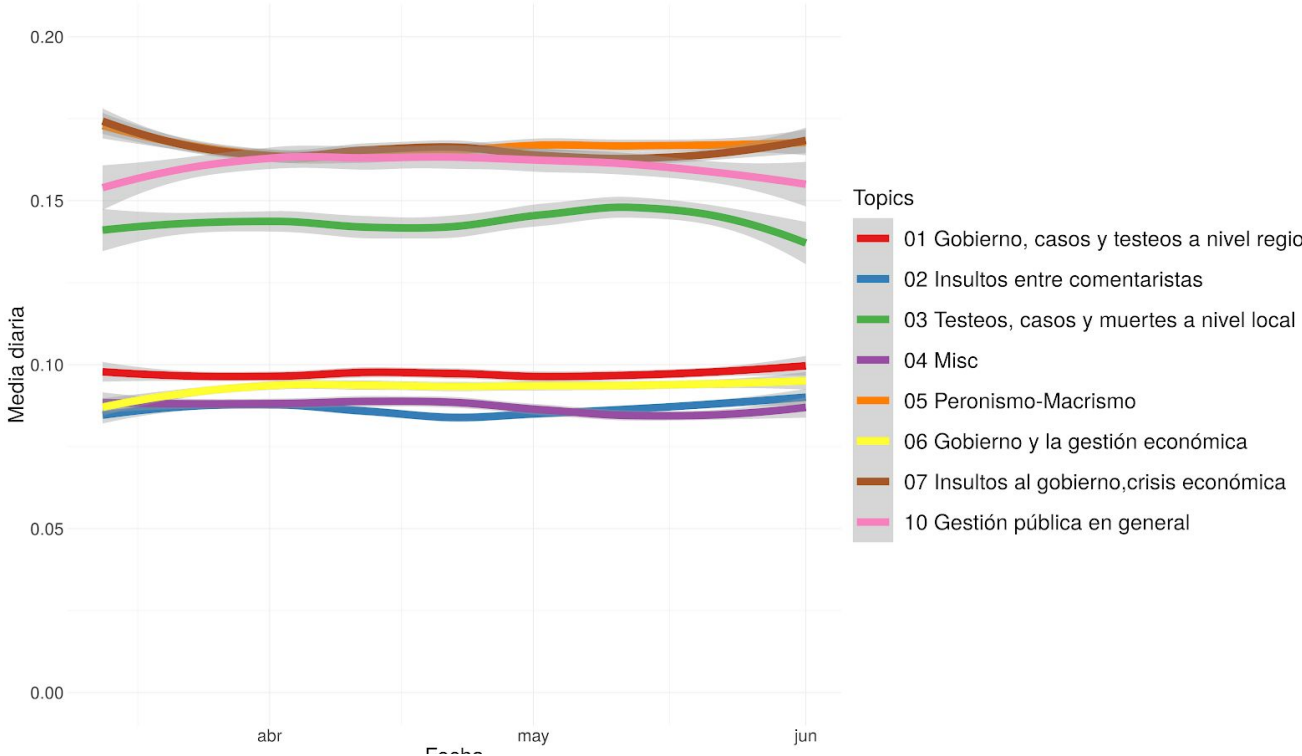
The issue-adjusted ideal point model

Figure 3: Left: Top words from topics fit using labeled LDA [6]. Right: the issue-adjusted ideal point model, which models votes v_{ud} from lawmakers and legislative items. Classic item response theory models votes v using x_u and a_d, b_d. For our work, documents' issue vectors $\boldsymbol{\theta}$ were estimated fit with a topic model (left of dashed line) using bills' words w and labeled topics β. Expected issue vectors $\mathbb{E}_q\left[\boldsymbol{\theta}|\boldsymbol{w}\right]$ are then treated as constants in the issue model (right of dashed line).

No supervisado - Aplicaciones

Detección de temas en comentarios a noticias sobre COVID-19

[Rosati, Chazarreta, Domenech y Maguire, 2023]



Un flujo de trabajo “típico” en NLP

Flujo de trabajo en NLP - Preprocesamiento

- Limpieza del texto (texto característico de los formatos)
- Cambiar mayúsculas por minúsculas
- Eliminar signos de puntuación y caracteres extraños (#$%&?’!.,)
- Eliminar números (1,2,3,4...)
- Eliminar “stopwords”

Flujo de trabajo en NLP - Preprocesamiento

- Exclusión de palabras muy comunes con poco valor para recuperar información del documento o corpus

- La cantidad de ocurrencias de una palabra en el texto determina si es o no una “stopword” cuanto más ocurrencias existan menos relevancia tiene en el texto.

- Artículos, pronombres, preposiciones, y conjunciones.

- Reducir el tamaño del texto para analizar, eliminando aproximadamente el 30 % o 40 % de dichas palabras.

Flujo de trabajo en NLP - Preprocesamiento

- Limpieza del texto (texto característico de los formatos)
- Cambiar mayúsculas por minúsculas
- Eliminar signos de puntuación y caracteres extraños (#$%&?'!.,)
- Eliminar números (1,2,3,4...)
- Eliminar "stopwords"
- Tokenización...

Flujo de trabajo en NLP - Preprocesamiento

- Tokenización: proceso que divide una secuencia (por ejemplo,una oración) en *tokens*
- Un *token* puede ser pensada como una unidad útil para el procesamiento semántico (oraciones, párrafos, documentos, etc.)

- Sistemas de escritura occidental: los espacios en blanco y ciertas formas de puntuación (puntos, comas, etc.) son delimitadores útiles para identificar tokens

Flujo de trabajo en NLP - Preprocesamiento

- Input:
  - [No es la conciencia (...) la que determina su ser sino (...) el ser social lo que determina su conciencia.]

- Output:
  - [No], [es], [la], [conciencia], [la], [que], [determina], [su], [ser], [sino], [el], [ser], [social], [lo], [que], [determina], [su], [conciencia]

Flujo de trabajo en NLP - Preprocesamiento

Reducir las palabras a su raíz

Y poder “reducir” la complejidad del dataset

Flujo de trabajo en NLP - Preprocesamiento

Reducir las palabras a su raíz

Y poder “reducir” la complejidad del dataset

“Affectation” “Affects” “Affections” “Affected” “Affection” “Affecting”

Flujo de trabajo en NLP - Preprocesamiento

Reducir las palabras a su raíz

Y poder “reducir” la complejidad del dataset

“Affectation” “Affects” “Affections” “Affected” “Affection” “Affecting”

Flujo de trabajo en NLP - Preprocesamiento

Reducir las palabras a su raíz

Y poder “reducir” la complejidad del dataset

“Affectation”	“Affects”	“Affections”	“Affected”	“Affection”	“Affecting”
“Affect”	“Affect”	“Affect”	“Affect”	“Affect”	“Affect”

Flujo de trabajo en NLP - Preprocesamiento

Reducir las palabras a su raíz

Y poder “reducir” la complejidad del dataset

“Affectation”	“Affects”	“Affections”	“Affected”	“Affection”	“Affecting”
“Affect”	“Affect”	“Affect”	“Affect”	“Affect”	“Affect”

Inconveniente: No funciona siempre. Hay palabras que **su raíz depende del contexto** de la oración. Se requiere un **análisis morfológico**.

Flujo de trabajo en NLP - Preprocesamiento

Lemmatization

En vez de cortar a la raíz podemos buscar su “lema” (también llamada “forma canónica”)

Flujo de trabajo en NLP - Preprocesamiento

Lemmatization

En vez de cortar a la raíz podemos buscar su “lema” (también llamada “forma canónica”)

El lema es la palabra que nos encontraríamos en el diccionario tradicional:

Flujo de trabajo en NLP - Preprocesamiento

Lemmatization

En vez de cortar a la raíz podemos buscar su “lema” (también llamada “forma canónica”)

El lema es la palabra que nos encontraríamos en el diccionario tradicional:

-singular para sustantivos (“Mesa” -> “Mesas”)
-masculino singular para adjetivos (“guapas” -> “guapo”)
-infinitivo para verbos (“dije”, “diré”, “dijéramos” -> “decir”)

Flujo de trabajo en NLP - Preprocesamiento

Lemmatization

En vez de cortar a la raíz podemos buscar su “lema” (también llamada “forma canónica”)

El lema es la palabra que nos encontraríamos en el diccionario tradicional:

-singular para sustantivos (“Mesa” -> “Mesas”)
-masculino singular para adjetivos (“guapas” -> “guapo”)
-infinitivo para verbos (“dije”, “diré”, “dijéramos” -> “decir”)

Similar a **stemming** ya que mapea muchas palabras a una sola pero el resultado de **lemmatization** es una palabra mientras que en stemming puede no serlo

Vamos al Notebook