

# **Introducción a la minería de textos y procesamiento de lenguaje natural para ciencias sociales**

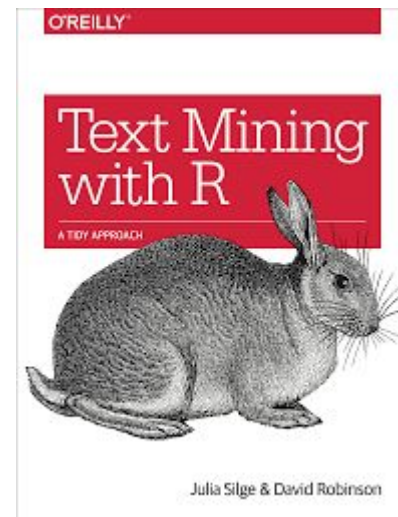
## **Clase 1. Fundamentos conceptuales**



# Dinámica de clases

- Bloques de 50-55 minutos
- Cortes de 15 minutos
- Actividades independientes

# Herramientas



# ¿Qué es NLP?



- Todo el tiempo estamos produciendo textos
  - Charlas
  - Entrevistas
  - Posts
  - Redes sociales
  - Etc.
- ¿Cómo podemos aprovechar esos textos en la investigación?



# El problema de los datos

MAS_500 Agglomerados segun tamaño	AGLOMERADO Codigo de Aglomerado	PONDERA Ponderacion	CH03 Relacion de parentesco	CH04 Sexo	CH05 Fecha de nacimiento (dia, mes y año)
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!

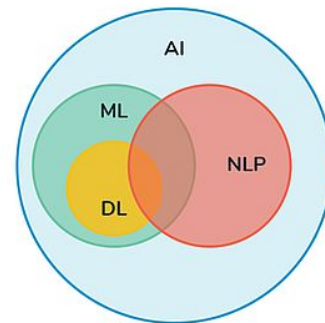
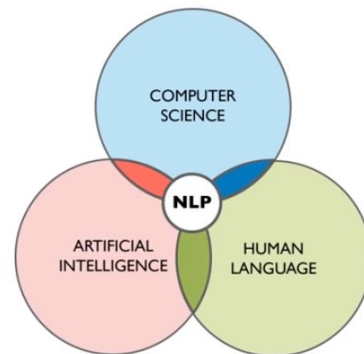
[2] este tango nacio para bailarse | y asi hamacarse muy suavemente | oigan ustedes este compas... | es muy sencillo bailar el tango | un doble paso despues descanso | la media vuelta la vuelta entera | y siempre junto a la compaÑera | este tango nacio para bailarse | no hay que quedarse mirandolo

[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 prosenio un corralon | modesto fue su blason | y la dulce purretita | se lavaba la carita | en el viejo pileton | amante del varietal | 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 cause 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



# 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

SENTIMENT ANALYSIS



factor-data  
EIDAES\_UNSAM

# Clasificación - Aplicaciones

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

[\[Hanna, 2017\]](#)

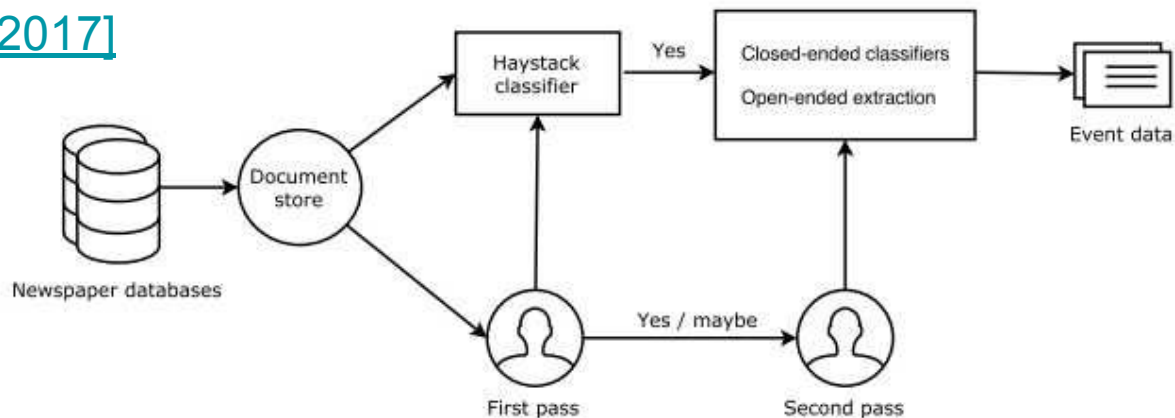
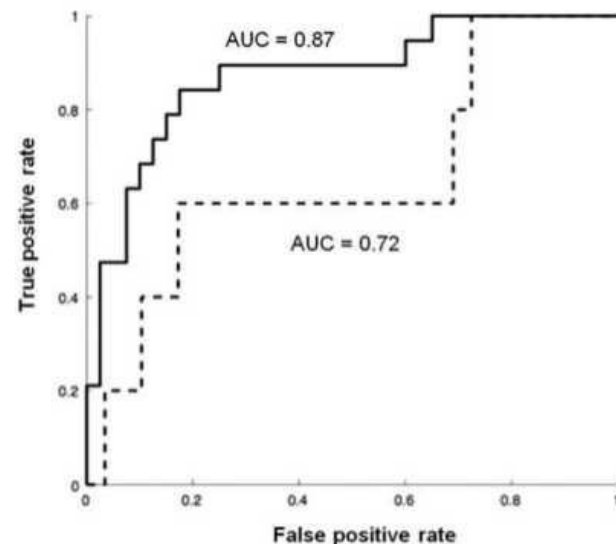


Figure 1: MPEDS pipeline with training.

# 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.

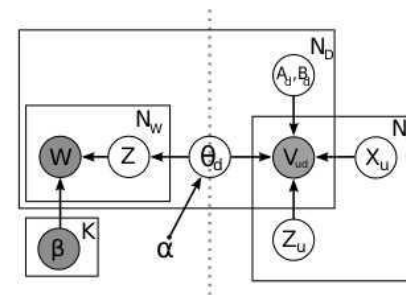
# No supervisado - Aplicaciones

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
hezbollah	community	coast guard
national guard	family	substitute

Labeled topics



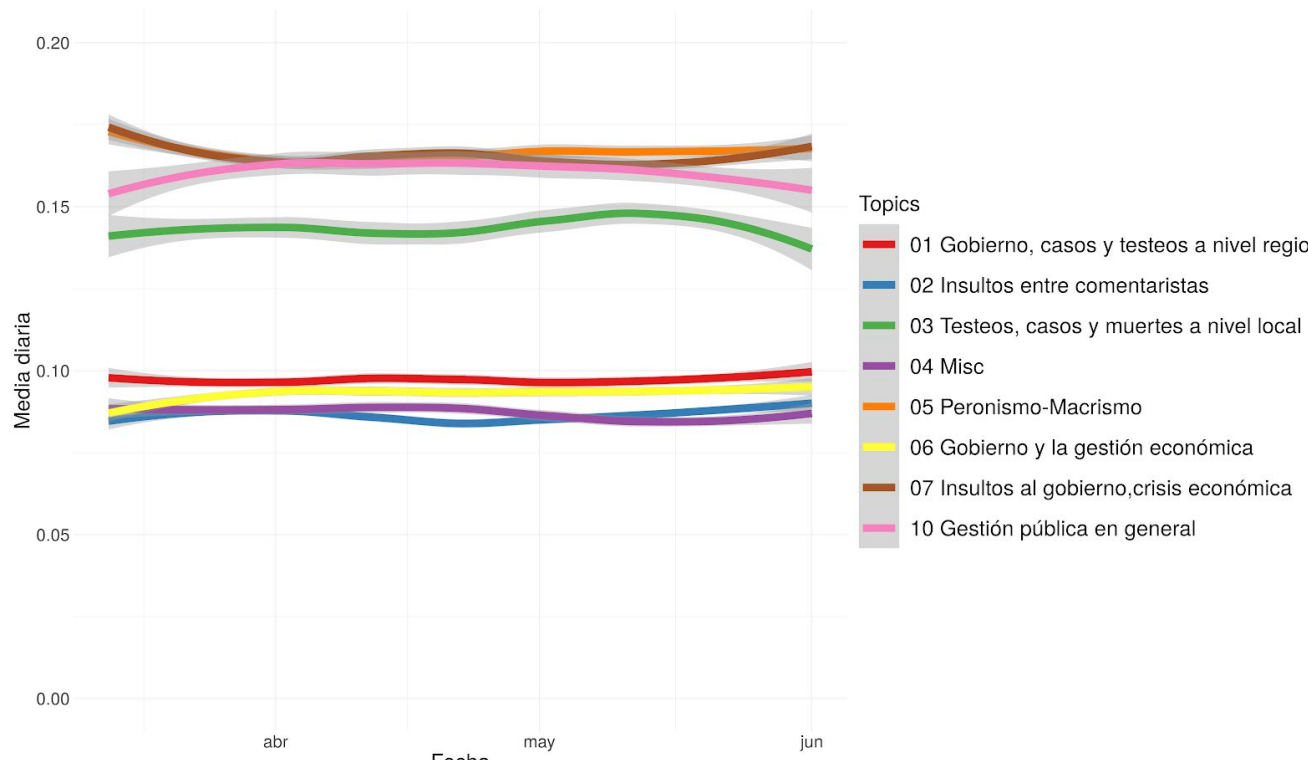
The issue-adjusted ideal point model

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  $\theta$  were estimated fit with a topic model (left of dashed line) using bills' words  $w$  and labeled topics  $\beta$ . Expected issue vectors  $\mathbb{E}_q[\theta|w]$  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

[en prensa...]



# 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

The logo for the stemming process, featuring the word "Stemming" in a blue box. The word "Stem" is in white and "ming" is in red.

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



factor-data  
EIDAES\_UNSAM