

# Introducción al Procesamiento Natural del Lenguaje

Fernanda Sobrino

6/19/2021

Por qué queremos usar texto como data?

Analís is cuantitativo de texto

Procesamiento natural del lenguaje

Proceso de análisis de texto

Definiciones

Organizar el texto

Definir características

Por qué queremos usar texto como data?

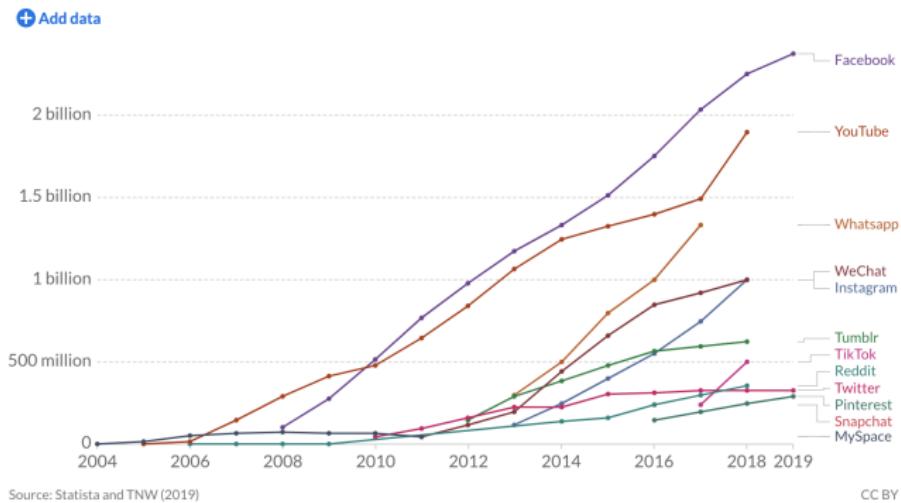
## Historicamente (pre 2000)

- ▶ las interacciones sociales se dan por medio de texto
- ▶ las ciencias sociales han evitado estudiar este tipo de datos
- ▶ Por qué?
  - ▶ difíciles de conseguir
  - ▶ pérdida de tiempo
  - ▶ no generalizables (formatos distintos, etc)
  - ▶ difícil de guardar y alcanzar (archivos históricos)
  - ▶ intensivos computacionalmente

# Post 2000: Redes Sociales

Number of people using social media platforms, 2004 to 2019  
Estimates correspond to monthly active users (MAUs). Facebook, for example, measures MAUs as users that have logged in during the past 30 days. See source for more details.

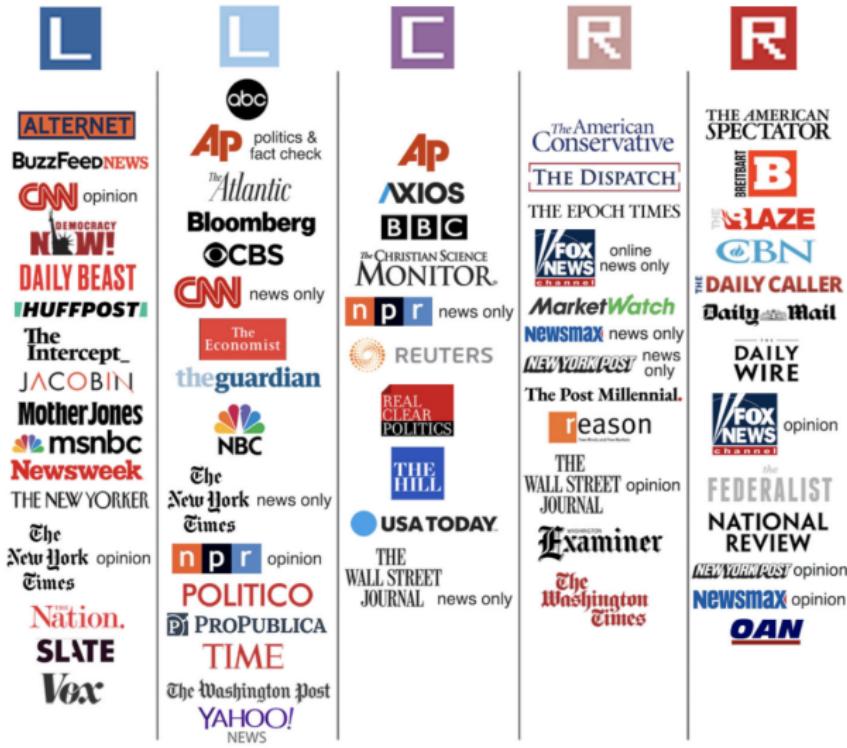
Our World  
in Data



# Post 2000: Medios tradicionales

## AllSides™ Media Bias Chart

All ratings are based on online content only — not TV, print, or radio content.  
Ratings do not reflect accuracy or credibility; they reflect perspective only.



L LEFT

L LEAN LEFT

C CENTER

R LEAN RIGHT

R RIGHT

## Post 2000: Archivos históricos

Page No. 2

**85\*** Inquiries numbered 7, 16, and 17 are not to be asked in respect to infants. Inquiries numbered 11, 12, 15, 16, 17, 19, and 29 are to be answered (if at all) merely by an affirmative mark, as /.

SCHEDULE 1.—Inhabitants in ~~the Eleventh District~~, <sup>16th Ward</sup> in the County of New York, State of New York, enumerated by me on the 20 day of June, 1870.

Post Office: New York City

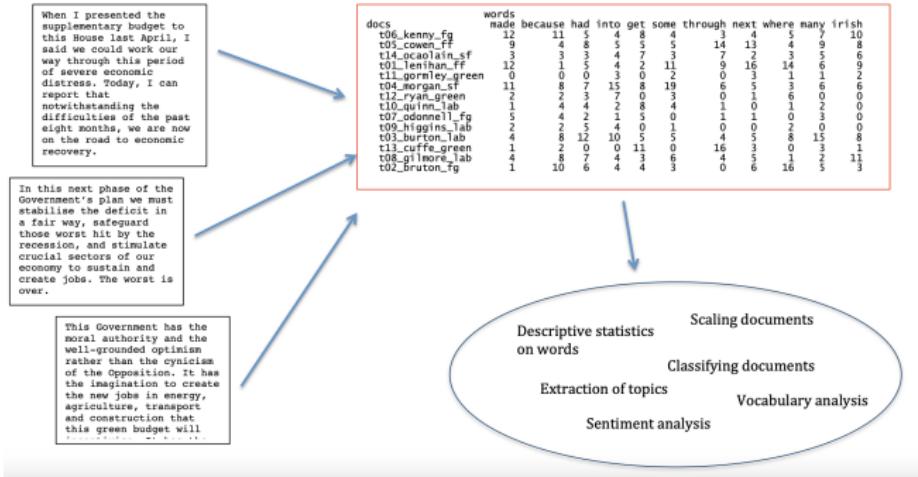
John Miller, Ass't Marshal.

## Por qué paso esto?

- ▶ incremento en la disponibilidad de texto no estructurado
- ▶ reducción en los costos de almacenamiento 1956 \$10,000 por megabyte, 2014 menos de 0.0001 por magabyte
- ▶ explosión en métodos, lenguajes de programación y librerías capaces de lidear con texto
  - ▶ generalizable: un mismo método puede ser usado entre bases de datos
  - ▶ sistemático: parámetros y estadísticas comparables entre datos
  - ▶ barato: R, Python, Julia, etc pueden lidear con estos datos y cualquiera puede hacerlo desde su casa

## Analís is cuantitativo de texto

# Procesamiento del texto a datos estructurados



# Métafora del pajar de Grimmer

El análisis automatizado mejoran la lectura

- ▶ analizar una pajita de heno ≡ entender el significado de una oración
  - ▶ Humanos: **Muy buenos**
  - ▶ Máquina: **Bastante malas**
- ▶ organizar todo el pajar ≡ describir, clasificar y escalar textos
  - ▶ Humanos: **Bastante malos**
  - ▶ Máquina: **Muy buenas**

# Principios del análisis cuantitativo de texto (Grimmer & Stuart 2013)

1. Todos los modelos están mal pero algunos son útiles
  - ▶ proceso de generación del texto desconocido
  - ▶ complejidad del lenguaje: temporalidad, sinónimos, ironía, sarcasmo, etc.
  - ▶ los modelos fallan al intentar capturar el lenguaje pero son buenos para tareas específicas
2. Los métodos cuantitativos aumentan las capacidades humanas pero no remplazan al humano
  - ▶ **algoritmo:** organizan, dirigen, sugieren
  - ▶ **humano:** lee e interpreta

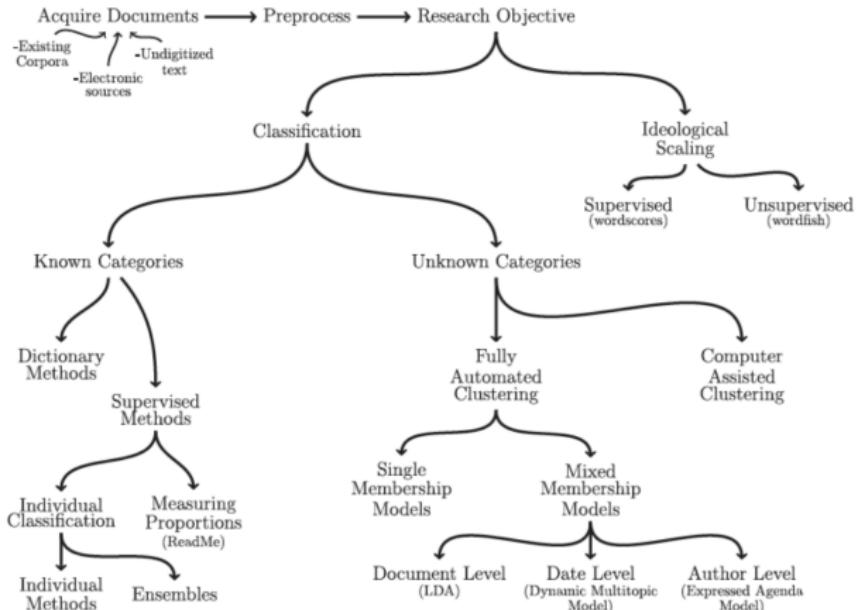
## Principios del análisis cuantitativo de texto (Grimmer & Stuart 2013)

3. No existe un método mejor que otro para automatizar texto
  - ▶ aprendizaje supervisado → categorías pre determinadas
  - ▶ aprendizaje no supervisado → categorías por clasificar
4. ⇒ validar, validar, validar
  - ▶ distinto rendimiento dependiendo de la tarea en cuestión
  - ▶ poca teoría que sustente que método escoger
  - ▶ **Evitar** aplicar métodos ciegamente

## Supuestos del análisis cuantitativo del texto

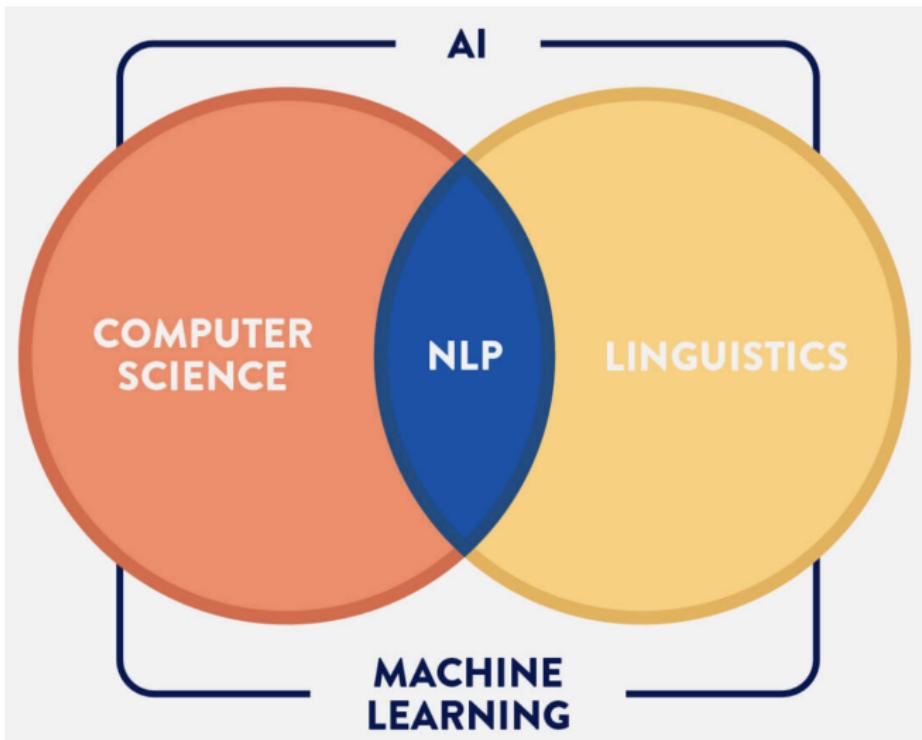
1. El texto representa alguna característica de interés
  - ▶ un atributo del autor
  - ▶ sentimiento/emoción
  - ▶ relevancia de alguna cuestión política, etc
2. El texto se puede representar extrayendo sus características
  - ▶ bolsa de palabras (bag of words)
  - ▶ word embeddings, etc
3. Una matriz de características del documento puede analizarse usando métodos cuantitativos para producir estimados validos y significativos de la característica de interés

# Métodos de análisis de texto

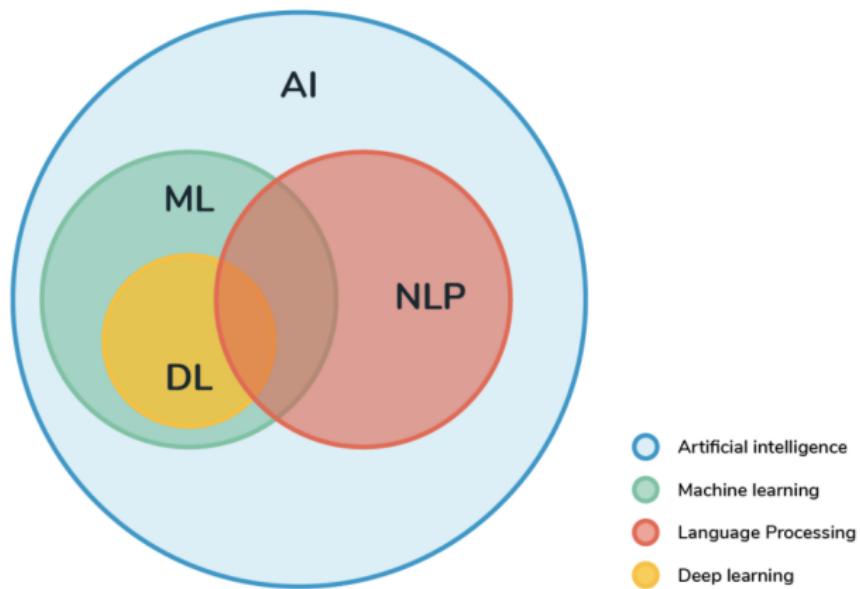


## Procesamiento natural del lenguaje

Qué es?



# Procesamiento Natural del Lenguaje



# Lenguaje Natural

- ▶ manera en que los humanos nos comunicamos
- ▶ voz y texto
- ▶ dada la cantidad de información transmitida de esta manera necesitamos:
  - ▶ métodos para entender el lenguaje natural

## Retos del lenguaje natural

- ▶ el lenguaje es ambiguo
- ▶ cambia y evoluciona constantemente
- ▶ somos muy buenos expresando, percibiendo e interpretando significados muy elaborados y matizados
- ▶ pero bastante malos para describir formalmente las reglas que gobiernan el lenguaje

## Retos del lenguaje natural: ejemplos

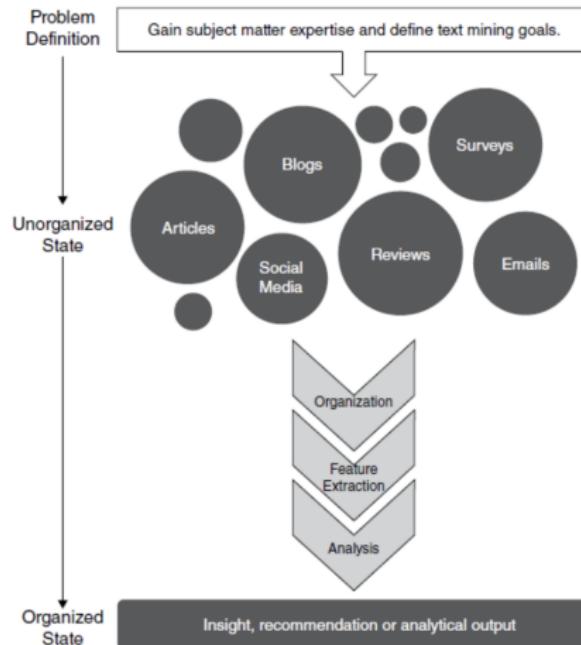
- ▶ Deja la comida que sobre sobre la mesa de la cocina, dijo llevando el sobre en la mano.
- ▶ Se moría de risa.
- ▶ Ella le dijo que los pusiera debajo
- ▶ coche/vehículo/automóvil/carro/etc

# Retos del lenguaje natural: ejemplos



## Proceso de análisis de texto

# Proceso de análisis de texto



# Flujo de trabajo

1. definir el problema y los objetivos específicos
2. identificar los datos de texto que deben ser recolectados (mining)
3. organizar el texto
4. extraer características
5. análisis
6. conclusiones

## Ejemplo: Hoberg and Phillips 2016

1. problema/pregunta de investigación:
  - ▶ cómo definir correctamente las fronteras entre industrias para poder medir mejor la competencia y la oferta?
  - ▶ nueva manera de clasificar industrias usando la descripción de sus productos
  - ▶ esto permite clasificaciones de industria mas flexibles y capaces de evolucionar con el tiempo

## Ejemplo: Hoberg and Phillips 2016

### 2. datos:

- ▶ todas las empresas que cotizan en bolsa tienen reportes 10-K
- ▶ estos reportes describen los productos que cada empresa provee
- ▶ para cada empresa  $i$  cuentan cuantas veces aparece cada palabra y lo representan como  $c_i$
- ▶ calculan que tan diferentes son dos productos usando el angulo entre los vectores  $c_i$  y  $c_j$

## Ejemplo: Hoberg and Phillips 2016

### 3. organizar texto:

- ▶ definen las industrias agrupando las empresas con los  $c$ 's mas cercanas
- ▶ 300 grupos distintos para que coincida con la clasificación de industrias pre existentes

### 4. extraer características:

- ▶ asignan una industria a cada una de las empresas por año  $\hat{v}_{it}$

### 5. análisis:

- ▶ examinan el efecto de shocks a la industria militar y de software en la competencia y la oferta de productos

# Ejemplo: Hoberg and Phillips 2016

## SubMarket 1 Entertainment (Sample Focal Firm: Wanderlust Interactive)

43 rivals: Maxis, Piranha Interactive Publishing, Brilliant Digital Entertainment, Midway Games, Take Two Interactive Software, THQ, 3DO, New Frontier Media, ...

SIC codes of rivals: computer programming and data processing [sic3=737] (24 rivals), motion picture production and allied services [sic3=781] (4 rivals), misc other (13 rivals)

Core words: entertainment (42), video (42), television (38), royalties (35), internet (34), content (33), creative (31), promotional (31), copyright (31), game (30), sound (29), publishing (29), ...

## SubMarket 2: Medical services (Sample Focal Firm: Quadramed Corp)

66 rivals: IDX Systems, Medicus Systems, Hpr, Simione Central Holdings, National Wireless Holdings, HCIA, Apache Medical Systems, ...

SIC codes of rivals: computer programming and data processing [sic3=737] (45 rivals), insurance agents, brokers, and service [sic3=641] (5 rivals), miscellaneous health services [sic3=809] (4 rivals), management and public relations services [sic3=874] (3 rivals), misc other (9 rivals)

Core words: client (59), database (54), solution (49), patient (47), copyright (47), secret (47), physician (47), hospital (46), healthcare (46), server (45), resource (44), functionality (44), billing (44), ...

## SubMarket 3: Information Transmission (Sample Focal Firm: FAXSAV)

259 rivals: Omtool Ltd, Concentric Network, Premiere Technologies, International Telecommunication Data Systems, IDT Corp, Axent Technologies, SoloPoint, Precision Systems, Netrix Corp, ...

SIC codes of rivals: computer programming and data processing [sic3=737] (112 rivals), communications equipment [sic3=366] (45 rivals), telephone communications [sic3=481] (38 rivals), computer and office equipment [sic3=357] (29 rivals), communications services, other [sic3=489] (7 rivals), miscellaneous business services [sic3=738] (7 rivals), misc other (15 rivals)

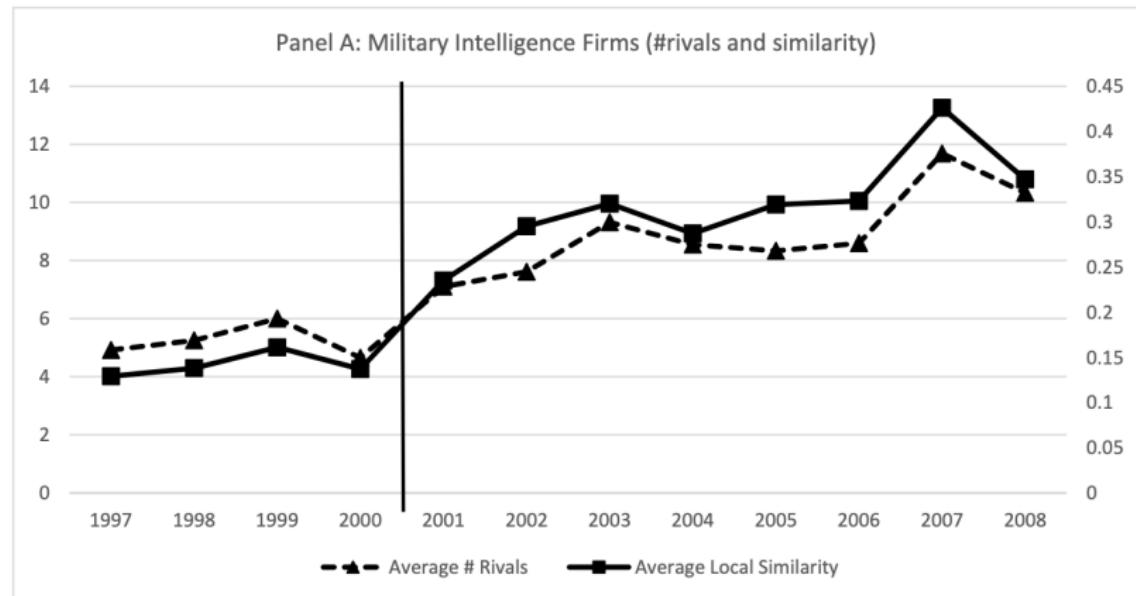
Core words: internet (236), telecommunications (211), interface (194), communication (188), solution (187), platform (184), architecture (182), call (177), infrastructure (173), voice (173), functionality (173), server (173), ...

# Ejemplo: Hoberg and Phillips 2016

## 6. Conclusiones:

- después del 11 de Septiembre del 2001 hay entrada de empresas al sector militar de inteligencia

Military Intelligence Firms: Competitor Changes



## Definiciones

## Algunas definiciones

- ▶ **corpus:** conjunto amplio y estructurado de texto para analizar
- ▶ **documento:** cada una de las unidades del corpus
- ▶ **tokens:** partes del documento pueden ser palabras, frases, oraciones, etc.
- ▶ **stems:** reducir una palabra a su raíz por ejemplo hablar → habl
- ▶ **lemmas:** toma en consideración el análisis morfológico de la palabra niñas/niños → niño
- ▶ **stop words:** palabras que por lo general se excluyen por no aportar nada al análisis, depende de la aplicación si las dejas o no. Ejemplos: un, una, es, en, por, pero
- ▶ **expresiones regulares:** secuencia de caracteres que conforman un patrón de búsqueda

# Normalización del texto

- ▶ **tokenizar** : va a depender del lenguaje
  - ▶ por lo general quieres considerar palabras como 'Nueva York' como un solo token
  - ▶ son reglas predeterminadas que dependen en expresiones regulares
- ▶ **normalización**: transformar las palabras o tokens en formas estandar
  - ▶ stems y lemmas
- ▶ **segmentación**: por lo general los signos de puntuación funcionan para segmentar el texto en oraciones o frases

## Edit distance / distancia de edición

- ▶ la mínima cantidad de operaciones de edición (insertar, borrar, sustituir) necesarias para transformar una palabra en otra
- ▶ ejemplos:
  - ▶ levenshtein: supresión, inserción y sustitución
$$dist(casa, calle) = 3$$
 $casa \rightarrow cala \rightarrow calla \rightarrow calle$
- ▶ intuitivamente: la distancia mínima de edición puede pensarse como una tarea de búsqueda de ruta mas corta entre dos palabras
- ▶ en la práctica: programación dinámica para encontrar la ruta mas corta entre X y Y

## Organizar el texto

## Preprocesamiento del texto

"Claro hay, en el sector privado, ciertos directivos que sí ganan mucho, pero la mayoría de la gente que trabaja en el sector privado no tiene muchos sueldos, es el caso de las universidades. Ustedes, qué bueno que tratamos el tema. A ver, que levante la mano quién gana 105 mil pesos mensuales. ¡Ah, te rayaste!"

# Preprocesamiento del texto

## 1. eliminar puntuación

“Claro hay en el sector privado ciertos directivos que sí ganan mucho pero la mayoría de la gente que trabaja en el sector privado no tiene muchos sueldos es el caso de las universidades Ustedes que bueno que tratamos el tema A ver que levante la mano quien gana 105 mil pesos mensuales Ah te rayaste”

# Preprocesamiento del texto

## 2. minúsculas (a veces)

“claro hay en el sector privado ciertos directivos que sí ganan mucho pero la mayoría de la gente que trabaja en el sector privado no tiene muchos sueldos es el caso de las universidades ustedes que bueno que tratamos el tema a ver que levante la mano quien gana 105 mil pesos mensuales ah te rayaste”

# Preprocesamiento del texto

## 3. quitar las stop words (a veces)

```
##  
## -- Column specification -----  
## cols(  
##   a = col_character()  
## )
```

# Preprocesamiento del texto

## 4. tokenizar

```
##      Token
## 1    claro
## 2    hay
## 3    en
## 4    el
## 5  sector
## 6 privado
```

# Preprocesamiento del texto

## 5. Lemmatize/Stemming

- ▶ proceso de reducir palabras a sus raíces
- ▶ reduces variaciones de la misma palabra
- ▶ stem corta el final de las palabras
- ▶ lemmatize considera el contexto y las corta a su base
- ▶ Ojo: casi todo lo que existe es para inglés y chino, en español (lenguas romances) puede no ser tan buena idea lematizar o hacer stemming

# Stemming

```
##      Token    stem
## 1  claro   clar
## 2    hay     hay
## 3    en      en
## 4    el      el
## 5 sector sector
## 6 privado  priv
```

## Lematización

```
##  
## Attaching package: 'rvest'  
  
## The following object is masked from 'package:readr':  
##  
##     guess_encoding  
  
##     Token    stem lemma_dict  
## 1   claro    clar      clarar  
## 2     hay     hay       haber  
## 3     en      en        en  
## 4     el      el        el  
## 5 sector  sector      sector  
## 6 privado priv      privado
```

## Preprocesamiento del texto

5. Otros pasos: remover cosas como

- ▶ urls
- ▶ html
- ▶ emojis
- ▶ números

## Pasos para limpiar/organizar el texto

1. eliminar puntuación
2. minúsculas
3. quitar stop words
4. tokenizar
5. lemmatize/stem
6. otros pasos

Definir características

## Cómo definimos que características necesitamos?

- ▶ caracteres
- ▶ palabras
  - ▶ stems
  - ▶ lemmas
  - ▶ quitamos stop words?
- ▶ grupos de palabras o n-grams
- ▶ oraciones

## Cómo definimos que características necesitamos?

- ▶ Todas?
  - ▶ ineficiente
  - ▶ palabras raras probablemente no informan mucho
  - ▶ palabras como 'a' tampoco
- ▶ Podar algunas palabras
  - ▶ frecuencia en el documento
  - ▶ frecuencia del termino
  - ▶ quitamos las stop words
  - ▶ usar un diccionario pre definido
  - ▶ sinónimos, antónimos, etc

## Transformar características a vectores (features to vectors)

- ▶ binario: 1 si la palabra está
- ▶ frecuencia: cuantas veces aparece cada palabra
- ▶ hashing: convierte cada palabra en una representación numérica
- ▶ ti-idf: contiene información de las palabras mas y menos importantes de todos los textos
- ▶ embeddings: toma en cuenta la semántica y el contexto

## Bolsa de palabras

- ▶ ignora el orden de las palabras
- ▶ cada documento es una bolsa que contiene varias palabras
- ▶ binario, frecuencias, hashing

# Bolsa de palabras

	about	bird	heard	is	the	word	you
About the bird, the bird, bird bird bird	1	5	0	0	2	0	0
You heard about the bird	1	1	1	0	1	0	1
The bird is the word	0	1	0	1	2	1	0

## Bolsa de palabras

1. vocabulario: todas las palabras del corpus
2. cada documento se representa como una vector del tamaño del documento
3. puede ser binario o contar cuantas veces aparece cada palabra

# Bolsa de palabras

- ▶ ventajas:
  - ▶ fácil de entender y programar
  - ▶ modelo base
  - ▶ funciona bien si tienes pocos documentos y son específicos
- ▶ desventajas:
  - ▶ el tamaño del vocabulario aumenta constantemente
  - ▶ sparse matrix → computacionalmente costoso
  - ▶ ignora la posición, grática, etc

# TD-IDF