

Introducción al Procesamiento Natural del Lenguaje

Fernanda Sobrino

6/22/2021

Introducción: PLM (NLP)

Por qué queremos usar texto como data?

Analísistico cuantitativo de texto

Proceso de análisis de texto

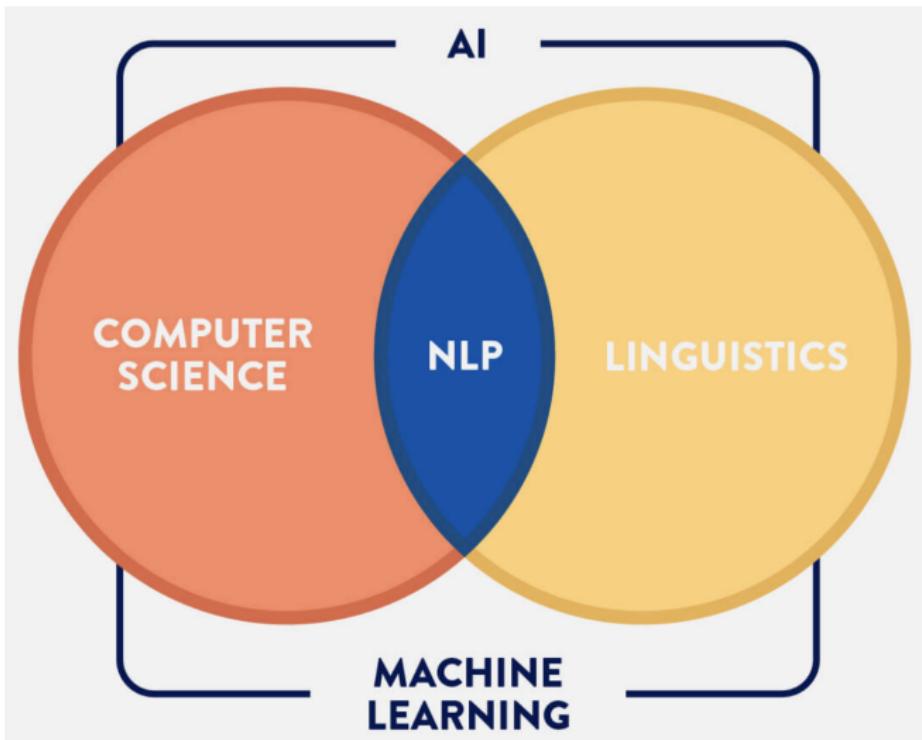
Definiciones

Definir las características

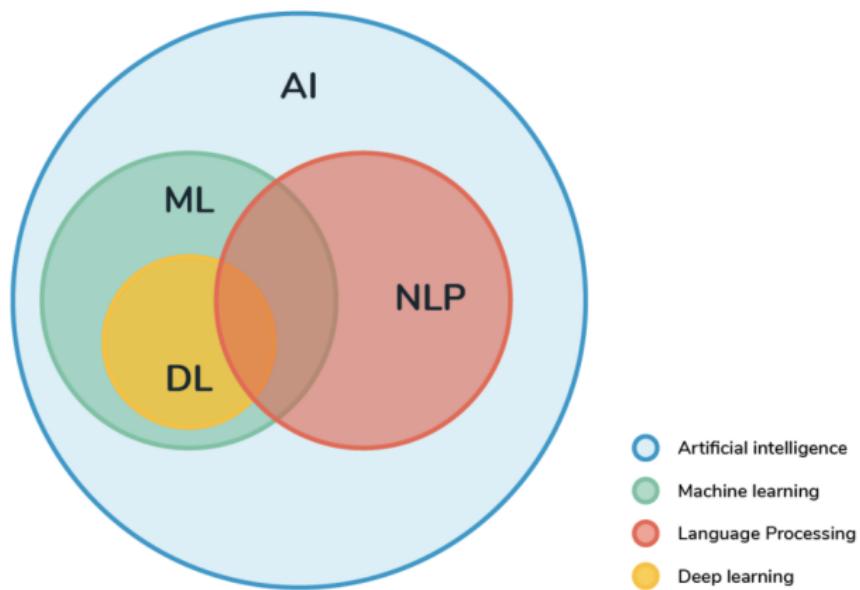
Procesamiento de texto básico

Introducción: PLM (NLP)

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



Aplicaciones de PLN

- ▶ Chatbots y asistentes de voz (conversaciones no estructuradas)
- ▶ Traductores automáticos (modelos del lenguaje)
- ▶ Monitoreo en redes sociales (análisis de sentimiento)
- ▶ Detección de noticias falsas (deep networks como BERT)
- ▶ Análisis de encuestas
- ▶ Revisores de gramática

Por qué queremos usar texto como data?

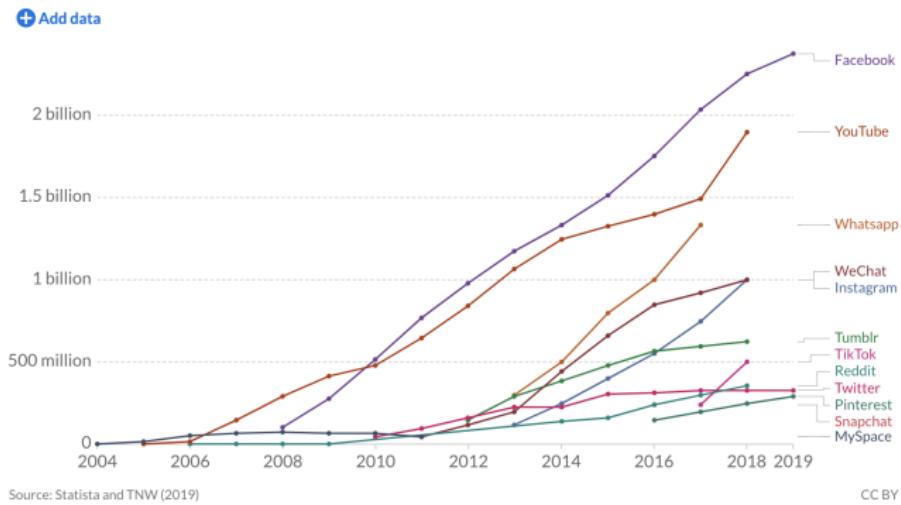
Historicamente (pre 2000)

- ▶ 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.

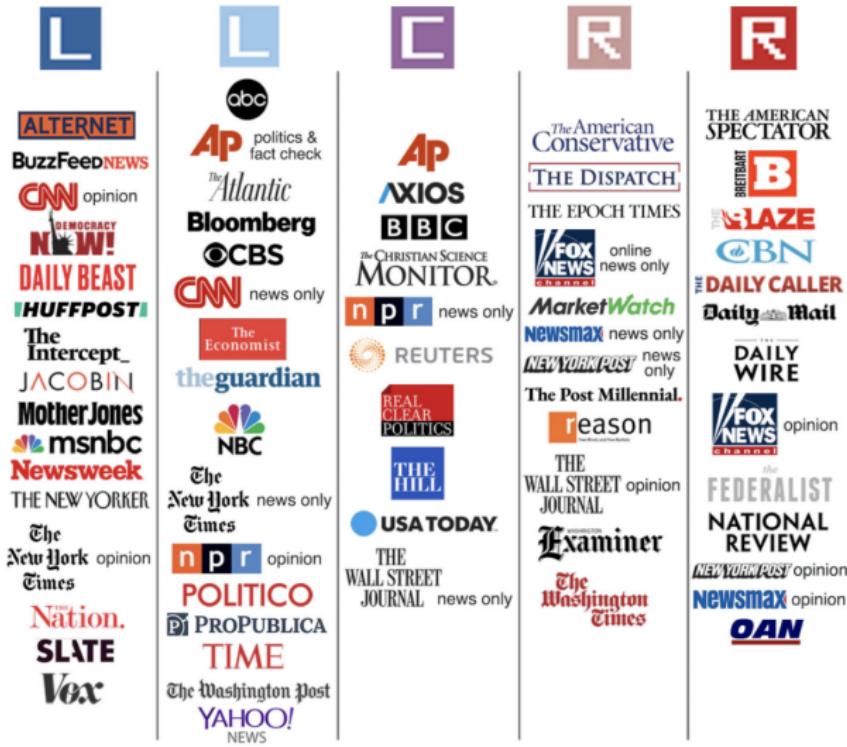
Our World
in Data



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, 18, 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 /.

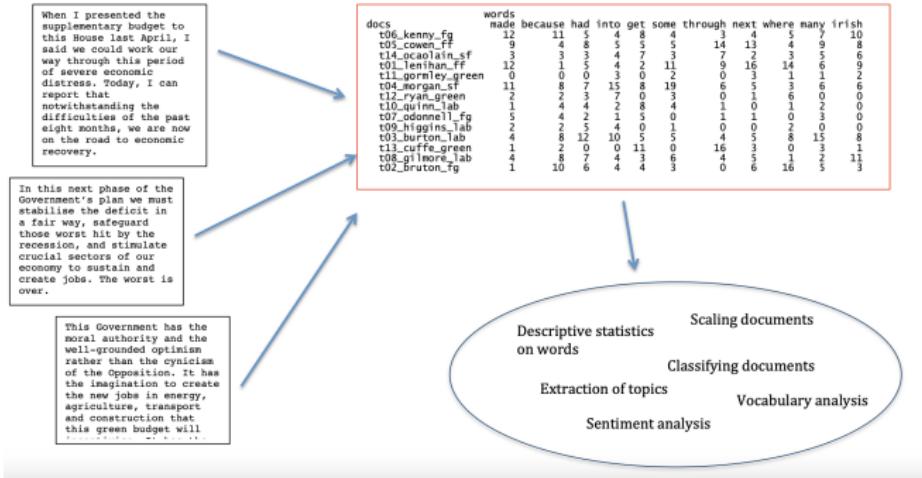
SCHEDULE 1.—Inhabitants in ~~the Eleventh District~~ 16th Ward, in the County of New York, State
of New York, enumerated by me on the 20 day of June, 1870.

Post Office: New York City

John Miller, Ass't Marshal.

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

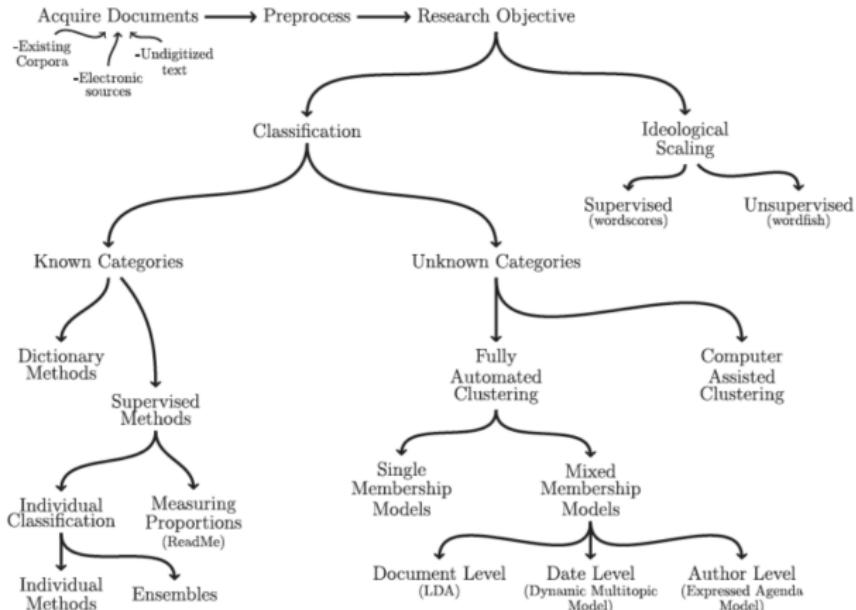
Principios del análisis cuantitativo de texto (Grimmer & Stuart 2013)

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

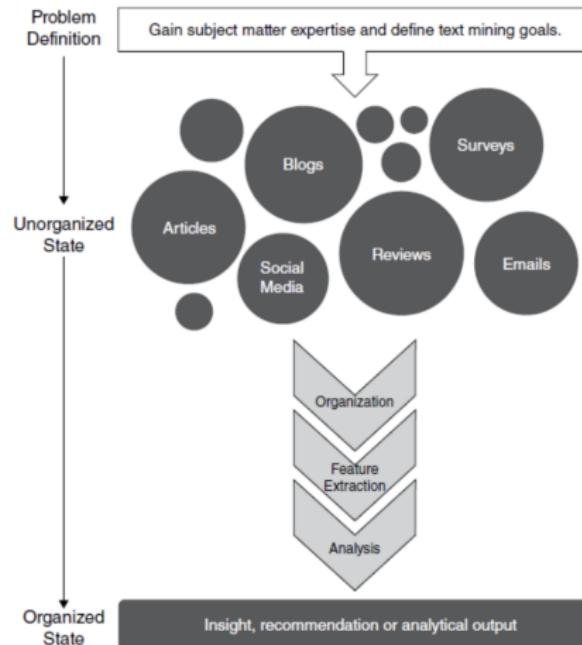
1. El texto representa alguna característica de interés
 - ▶ un atributo del autor
 - ▶ sentimiento/emoción
 - ▶ relevancia de alguna cuestión política, etc
2. El texto se puede representar extrayendo sus características
 - ▶ supuesto de la bolsa de palabras
 - ▶ word embeddings
3. Una matriz de características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

Métodos de análisis de texto



Proceso de análisis de texto

Proceso de análisis de texto



Proceso de análisis de texto

1. seleccionar textos: corpus
2. definir los documentos: la unidad de análisis (tweets, oraciones, párrafos, guiones)
3. definir las características: tokens, frases, segmentos, lenguaje, etc
4. convertir estas características a una matriz
5. proceso cuantitativo o estadístico para extraer información de la matriz
6. resumen, datos nuevos, etc

Ejemplo: Hoberg and Phillips 2016

- ▶ 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

Hoberg and Phillips 2016 (Análisis de texto)

1. corpus:
 - ▶ todas las empresas que cotizan en bolsa tienen reportes 10-K
 - ▶ estos reportes describen los productos que cada empresa provee
 - ▶ corpus: conjunto de todos estos reportes
2. documentos:
 - ▶ reporte por empresa y por año
3. características:
 - ▶ palabras en cada uno de los documentos

Hoberg and Phillips 2016 (Análisis de texto)

4. matriz:
 - ▶ matriz de frecuencias de palabras
 - ▶ $c_i(w)$ cuantas veces aparece la palabra w en el reporte 10-K de la empresa
 - ▶ tienen una matriz para cada año
5. proceso cuantitativo
 - ▶ calculan que tan diferentes son dos productos usando el angulo entre los vectores c_i y c_j para cada empresa
 - ▶ definen las industrias agrupando las empresas con las c 's mas cercanas
 - ▶ 300 grupos distintos para que coincida con la clasificación de industrias pre existentes
6. resultado:
 - ▶ la nueva clasificación de las industrias

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)

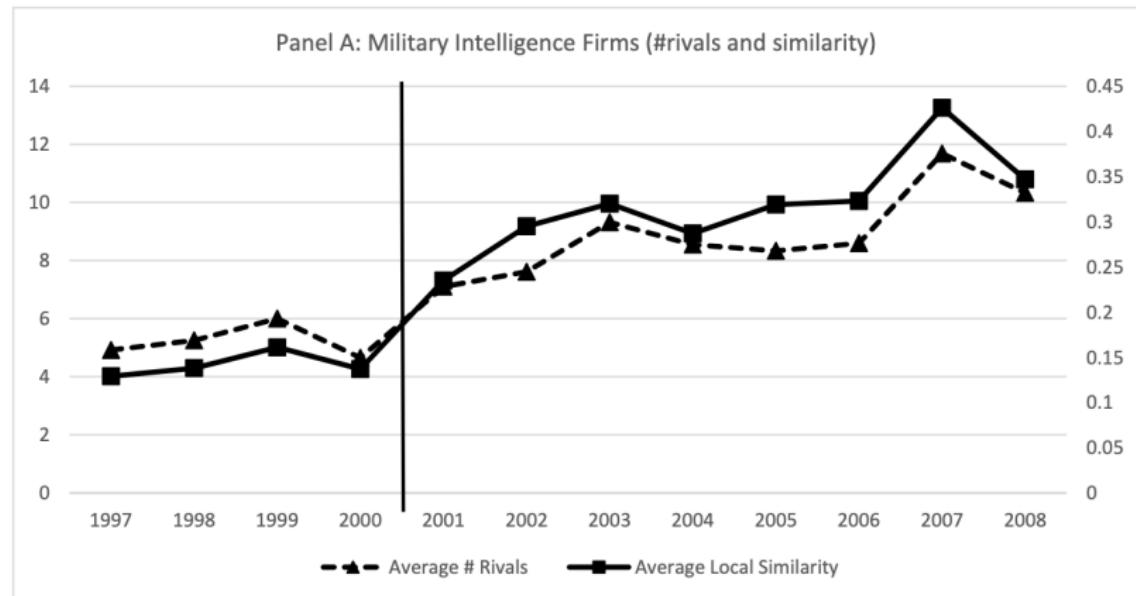
Core words: internet (236), telecommunications (211), interface (194), communication (188), solution (187), platform (184), architecture (182), call (177), infrastructure (173), voice (173), functionality (173), server (173), ...

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

Proceso de análisis de texto

1. seleccionar textos: corpus
2. definir los documentos: la unidad de análisis (tweets, oraciones, párrafos, guiones)
3. definir las características: tokens, frases, segmentos, lenguaje, etc
4. convertir estas características a una matriz
5. proceso cuantitativo o estadístico para extraer información de la matriz
6. resumen, datos nuevos, etc

Definir las características

Cómo definimos las características?

- ▶ objetivo del análisis
- ▶ tipo de documentos
- ▶ lenguaje del texto

Cómo definimos que características necesitamos?

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

Procesamiento de texto básico

Expresiones regulares (regex)

- ▶ cadena de caracteres que es utilizada para describir o encontrar patrones dentro de otros strings
- ▶ tedioso
 - ▶ tienes que pensar como una computadora
 - ▶ tienes que aprender la syntaxis
- ▶ es crucial para limpiar texto

Regex

regex	content
[abc]	matches a, b, or c
[^abc]	matches anything except a, b, or c
abc def	matches abc or def
ab(c d)ef	matches ab + (c or d) + ef

```
str_extract(c("grey", "gray"), "gr(e|a)y")
```

```
## [1] "grey" "gray"
```

Regex

regex	content
\[a-zA-Z]	matches any letter
\d	matches any digit
\D	matches any non-digit
\s	matches any whitespace (e.g. space, tab, newline)

```
str_extract(c("png", "123", "1n3"), "\\\d")
```

```
## [1] NA  "1" "1"
```

```
str_extract(c("A 1", "Z\n9"), "\\\D\\\\s\\\\d")
```

```
## [1] "A 1"  "Z\n9"
```

Regex

- ▶ cuidado con
- ▶ necesitamos \ para que la expresion regular genere una sola
- ▶ por ejemplo \. si queremos hacer match con un punto

```
str_extract(c("abc", "a.c", "bef"), "a\\.c")
```

```
## [1] NA      "a.c"  NA
```

Regex

- ▶ Match en la cantidad de ocurrencias

- ▶ ?, 0 o 1
- ▶ ▶ , 1 o mas
- ▶ ▶ , 0 o mas

```
x <- "1888 en numeros romanos es MDCCCLXXXVIII"  
str_extract(x, "CC?")
```

```
## [1] "CC"
```

```
str_extract(x, "CC+")
```

```
## [1] "CCC"
```

Regex

- ▶ Match en la cantidad de ocurrencias
 - ▶ $\{n\}$ exactamente n
 - ▶ $\{n, \}$ n o más
 - ▶ $\{ ,m\}$ cuando mucho m
 - ▶ $\{n, m\}$ entre n y m

```
x <- "1888 en numeros romanos es MDCCCLXXXVIII"  
str_extract(x, "C{2}")
```

```
## [1] "CC"  
  
str_extract(x, "C{2,}")
```

```
## [1] "CCC"  
  
str_extract(x, "C{2,3}")  
  
## [1] "CCC"
```

Regex

regex	content
[:lower:]	matches lower case
[:alpha:]	matches letters
[:alnum:]	matches alphanumeric characters
[:punct:]	matches punctuation

Por qué necesitamos Regex?

- ▶ sorpresivamente son muy importantes para NLP
 - ▶ secuencias sofisticadas de regex son el primer modelo que quieras correr
 - ▶ a veces basta una regex para resolver la pregunta/problema
- ▶ ayudan a limpiar el texto
- ▶ buenas para extraer ciertas características
- ▶ son buenas captando generalizaciones

Ejemplo: frase aleatoria de AMLO

“En este corredor habrá energía eléctrica y gas a precios bajos, así como subsidios fiscales para la instalación de fábricas y la creación de empleos, en tres años estará funcionando: me cансo, gанso.
<https://www.nacion321.com/gobierno/las-9-frases-mas-divertidas-que-nos-regalo-amlo-en-100-dias>”

Preprocesamiento del texto (Regex)

- ▶ Podemos usar regex para quitarle a este frase el html del final

```
frase <- gsub("http[:alnum:] [:punct:]"]*", "", frase)
```

“En este corredor habrá energía eléctrica y gas a precios bajos, así como subsidios fiscales para la instalación de fábricas y la creación de empleos, en tres años estará funcionando: me cансo, gансo.”

Preprocesamiento del texto (Regex)

Eliminar puntuación * Acá hay que decidir que vamos a hacer con las palabras acentuadas * las dejamos tal cual * las convertimos en palabras no acentuadas

```
frase <- gsub("[[:punct:]]", " ", frase)
```

“En este corredor habrá energía eléctrica y gas a precios bajos así como subsidios fiscales para la instalación de fábricas y la creación de empleos en tres años estará funcionando me canso ganso”

Normalización del texto

- ▶ poner las palabras en formato estandar
 - ▶ U.S.A , USA
 - ▶ Copa, copa
 - ▶ contracciones en ingles
 - ▶ Dr, Doctor
 - ▶ :), feliz

Ejemplo: normalización del texto

En este caso solo convertiremos el texto todo a minusculas

```
frase <- tolower(frase)
```

“en este corredor habrá energía eléctrica y gas a precios bajos así como subsidios fiscales para la instalación de fábricas y la creación de empleos en tres años estará funcionando me cango ganso”

Tokenización (segmentación de los documentos)

- ▶ la manera más sencilla de hacerla es usando los espacios
(depende del lenguaje)

```
## [1] "en"           "este"         "corredor"      "habrá"
## [6] "eléctrica"    "y"            "gas"          "a"
## [11] "bajos"        ""             "así"          "como"
## [16] "fiscales"     "para"         "la"           "instalac
## [21] "fábricas"     "y"            "la"           "creació
## [26] "empleos"      ""             "en"           "tres"
## [31] "estará"        "funcionando"  ""             "me"
## [36] ""              "ganso"        ""
```

Preprocesamiento del texto (stop words)

- ▶ quitar las stop words

```
## [1] "corredor"      "energía"       "eléctrica"     "gas"  
## [6] "bajos"          "subsídios"     "fiscales"      "instalac  
## [11] "creación"       "empleos"      "tres"         "años"  
## [16] "canso"          "ganso"
```

Preprocesamiento del texto

- ▶ 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  corredor corredor
## 2  energía     energ
## 3 eléctrica   electr
## 4      gas       gas
## 5  precios      preci
## 6    bajos      baj
```

Lematización

```
##  
## Attaching package: 'rvest'  
  
## The following object is masked from 'package:readr':  
##  
##     guess_encoding  
  
##      Token      stem lemma_dict  
## 1  corredor  corredor    corredor  
## 2   energía   energ    energía  
## 3 eléctrica  electr  eléctrico  
## 4      gas      gas      gas  
## 5  precios     preci     precio  
## 6     bajos     baj     bajo
```

Preprocesamiento del texto

- ▶ Objetivo: definir las características
- ▶ Posibles pasos:
 1. eliminar urls
 2. eliminar puntuación
 3. normalización de las palabras
 4. tokenizar
 5. quitar stop words
 6. lemmatize/stem
 7. otros pasos

Preprocesamiento del texto

- ▶ Ojo: esto no es una receta, que pasos y en qué orden aplicarlos va a depender de la aplicación en cuestión
- ▶ Resultado: texto listo para ser transformado en la matriz de documento-característica