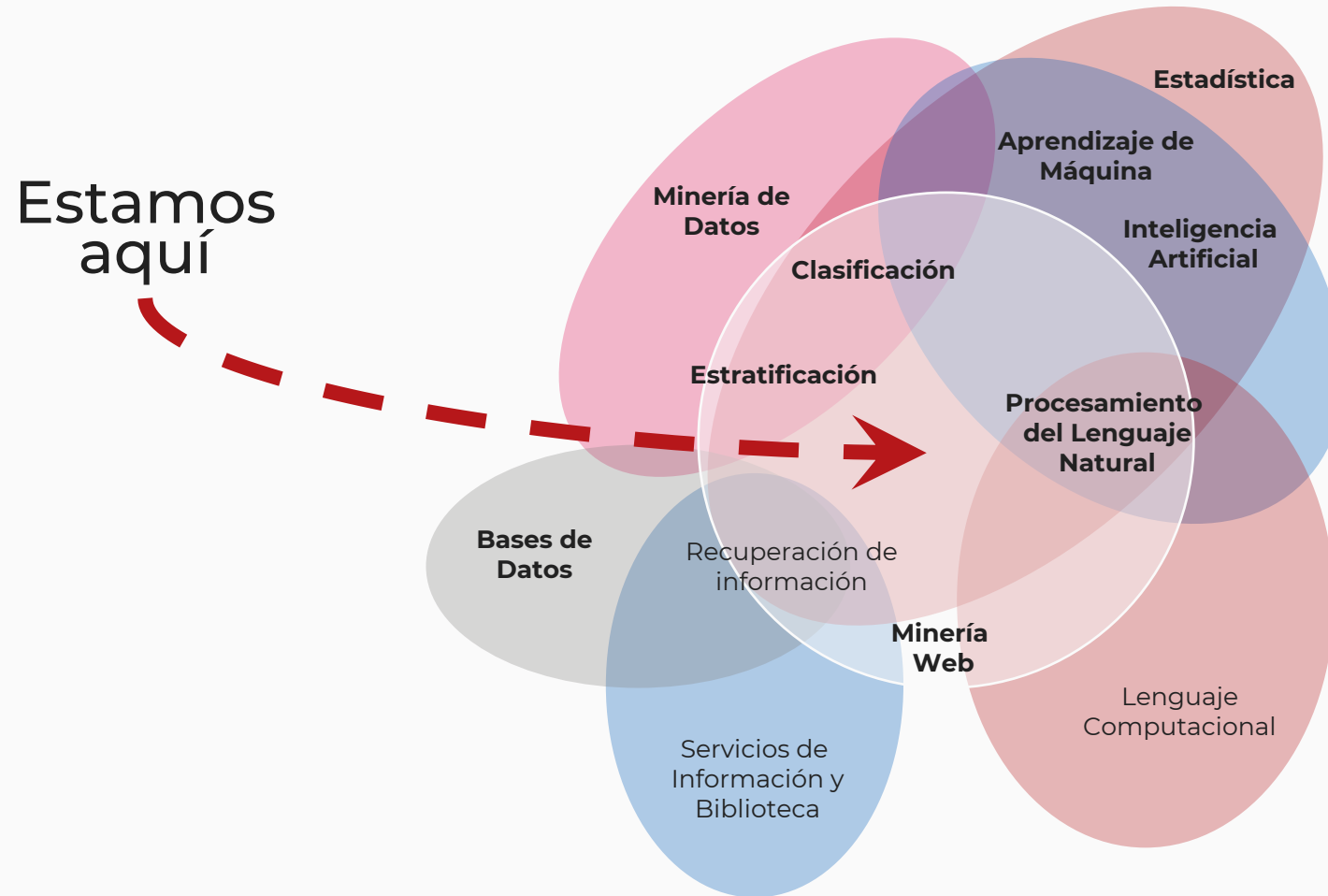


Minería y análisis de texto

Mtro. René Rosado González
Director de Programa LTP

Análisis de Texto

Text Analytics



Los 4 principios del análisis de texto automatizado

Four Principles of Automated Text Analysis



(1) Todos los modelos cuantitativos del lenguaje están equivocados, pero algunos son útiles.



(2) Los métodos cuantitativos para el texto amplían los recursos y exponencian la capacidad humana.



(3) No existe el mejor método global para el análisis de texto automatizado.



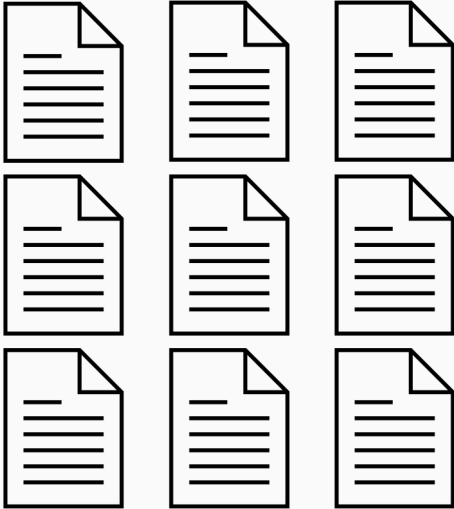
(4) Validar, Validar, Validar.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3), 267-297.

Conceptos clave

Key concepts

Corpus



Vocabulario

}]

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do
eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut
enim ad minim veniam, quis nostrud exercitation ullamco laboris
nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in
reprehenderit in voluptate velit esse cillum dolore eu fugiat
nulla pariatur. Excepteur sint occaecat cupidatat non proident,
sunt in culpa qui officia deserunt mollit anim id est laborum.
Sed ut perspiciatis unde omnis iste natus error sit voluptatem
accusantium doloremque laudantium, totam rem aperiam, eaque ipsa
quae ab illo inventore veritatis et quasi architecto beatae vitae

Token

}]

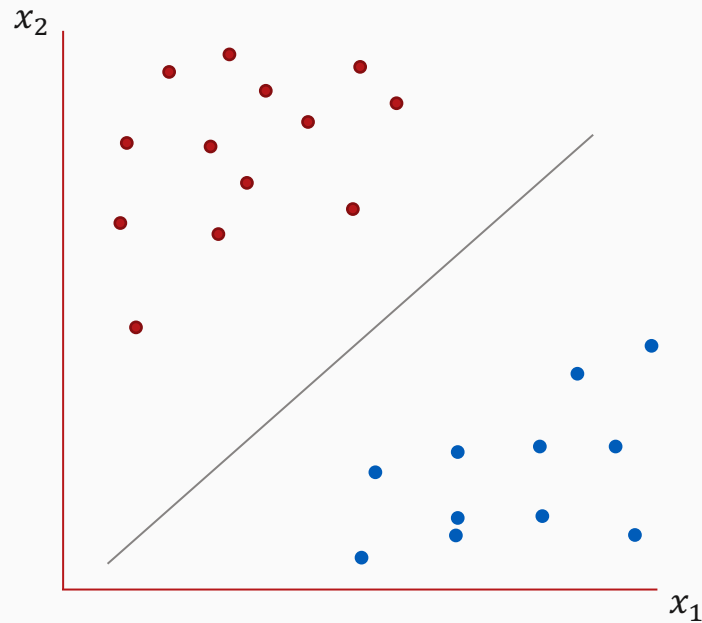
Lorem

Modelos Probabilísticos

Probabilistic Models

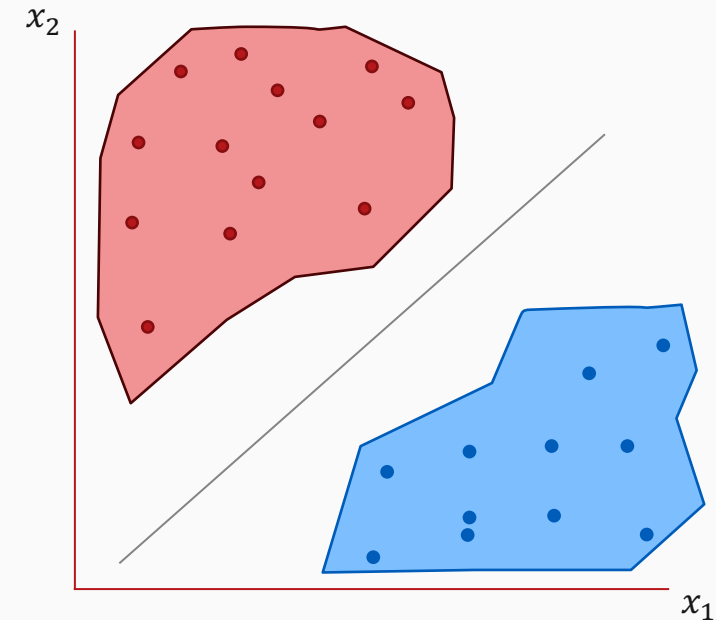
Discriminativos

Estimar $P(Y|X)$ directamente



Generativos

Estimar $P(X|Y)$ y deduce $P(Y|X)$



N-gramas

N-grams

Para que las probabilidades $P(W)$ reflejen la estructura del lenguaje pueden ser descritas como

$$W = w_1 w_2 \dots w_n$$

donde

w_i son las palabras contenidas en el texto W

Si consideramos la probabilidad de ocurrencia de una palabra en una frase respecto a las palabras inmediatas, no necesitamos considerar la frase completa

$$P(W) = P(w_1 w_2 \dots w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1 w_2) \dots P(w_n|w_1 w_2 w_3 \dots w_{n-1})$$

es decir, basta con calcular las probabilidades condicionales de la siguiente palabra:

$$P(w_{n+1}|w_1 w_2 w_3 \dots w_n)$$

N-gramas

N-grams

Un n-grama es la sucesión de longitud n de las palabras $w_1 w_2 \dots w_n$

$$P(w_1 w_2 \dots w_n) = \prod_{j=1}^n P(w_j | w_{j-1})$$

Ejemplo: $\langle s \rangle$ *Una vaca vestida de uniforme* $\langle /s \rangle$

Bigrama:

$$P(\text{Una} | \langle s \rangle) P(\text{vaca} | \text{una}) P(\text{vestida} | \text{vaca}) P(\text{de} | \text{vestida}) P(\text{uniforme} | \text{de}) P(\langle /s \rangle | \text{vestida})$$

Trigrama:

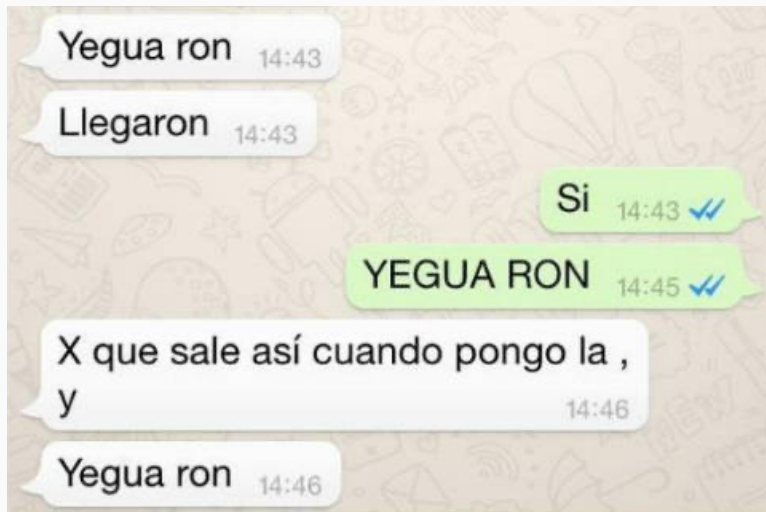
$$P(\text{vaca} | \langle s \rangle \text{Una}) P(\text{vestida} | \text{una vaca}) P(\text{de} | \text{vaca vestida}) P(\text{uniforme} | \text{vestida de}) P(\langle /s \rangle | \text{de uniforme})$$

Modelos de Lenguaje con N-gramas

N-grams Language Models

Desde el punto de vista estadístico, son una asignación de probabilidades $P(W)$ a cada posible frase del lenguaje $W = w_1 w_2 \dots w_n$

Ejemplo de un modelo de canal ruidoso:



Modelo de Canal Ruidoso

Noisy Channel Model

Verosimilitud (Modelo del Canal)
Probabilidad de observar el mensaje X
dado que es W

Probabilidad Posterior
Probabilidad de que y pertenezca
a una clase dado los datos

Probabilidad *Inicial*
(Modelo de Lenguaje)
Probabilidad observar el mensaje
 W dentro del contexto.

$$P(W|X) = \frac{P(X|W)P(W)}{P(X)}$$

Probabilidad de X
Probabilidad de observar X en
el mensaje

Modelo de Canal Ruidoso

Noisy Channel Model

X = Estoy a días minutos

	Probabilidad Inicial ↓		Resultado ↓
Modelo del Lenguaje (W)	$P(W)$	$P(X W)$	$P(X W)P(W)$
Estoy a veinte minutos	0.0002	0.01	0.00002
Estoy a diez minutos	0.0001	0.12	0.00012
Voy un poco tarde	0.0008	0.00	0.00000
Estoy a tías minutos	0.0000006	0.05	0.00000
		↑ Contexto	

Expresiones Regulares

Regular Expressions [RegEx]

Reg[ular]
Ex[pression]

Some people, when confronted with a problem, think: “I know, I’ll use regular expressions.” Now they have two problems.

— Jamie Zawinski

Inicio (^)

Conteo con Regex

“^y”

Conteo Sin Regex

“y”

rosa icela rodríguez velázquez: el aviso. se llega a la comunidad y se hace por tres cuestiones:

una, es avisando primero y acordando con las autoridades del pueblo, de la comunidad.

dos, con un perifoneo, porque pues allá no hay redes sociales como aquí.

y la otra es pegando cartulinas en las comunidades en donde tenemos más posibilidades de que pase la gente.

y lo más importante es el boca en boca que van teniendo las comunidades y que se ponen de acuerdo para acudir.

entonces, estamos muy agradecidos y muy sorprendidos por el orden y la participación, y muy agradecidos con todos los mexicanos y mexicanos que están acudiendo y que están participando, así como de los servidores públicos, que lo están haciendo muy bien, por eso reconozco a mis compañeros y a las dependencias.

y a cumplir con esta instrucción que nos dio el señor presidente, él nos dijo en marzo: ‘quiero hacer esto’, pero estaba el proceso electoral, entonces tuvimos que esperar hasta agosto para poderlo hacer, pero ya estamos listos desde hace algunas semanas.

entonces, decirles que es muy satisfactorio este programa, porque es devolver al pueblo lo incautado, lo decomisado y lo robado.

Inicio (^)

Conteo con Regex

"^y" = 3

Conteo Sin Regex

"y" = 15

rosa icela rodríguez velázquez: el aviso. se llega a la comunidad y se hace por tres cuestiones:
una, es avisando primero y acordando con las autoridades del pueblo, de la comunidad.
dos, con un perifoneo, porque pues allá no hay redes sociales como aquí.
y la otra es pegando cartulinas en las comunidades en donde tenemos más posibilidades de que pase la gente.
y lo más importante es el boca en boca que van teniendo las comunidades y que se ponen de acuerdo para acudir.
entonces, estamos muy agradecidos y muy sorprendidos por el orden y la participación, y muy agradecidos con todos los mexicanos y mexicanos que están acudiendo y que están participando, así como de los servidores públicos, que lo están haciendo muy bien, por eso reconozco a mis compañeros y a las dependencias.
y a cumplir con esta instrucción que nos dio el señor presidente, él nos dijo en marzo: 'quiero hacer esto', pero estaba el proceso electoral, entonces tuvimos que esperar hasta agosto para poderlo hacer, pero ya estamos listos desde hace algunas semanas.
entonces, decirles que es muy satisfactorio este programa, porque es devolver al pueblo lo incautado, lo decomisado y lo robado.

Fin (\$)

Conteo con Regex

`"\?$"` = 3

Conteo Sin Regex

`"\?"` = 4

interlocutora: ¿el programa aprende en casa continuará si los padres deciden no llevarlos?

presidente andrés manuel lópez obrador: sí, sí va a continuar, estamos contemplando eso.

interlocutora: y si hay una propuesta para los trayectos, para el cuidado.

presidente andrés manuel lópez obrador: sí, vamos a que haya seguridad en todo lo que se requiera.

interlocutora: ¿en qué consistirá?

presidente andrés manuel lópez obrador: vamos a dar a conocer el plan para que haya seguridad. tiene que protegerse a los niños, a las maestras, a los maestros, las escuelas. eso corresponde a los gobiernos estatales, siempre lo hacen y ahora vamos a pedir que se apliquen más.

interlocutora: y, bueno, una última pregunta, presidente, por favor. ¿qué pasó con la coordinadora nacional de trabajadores de la educación?, ¿ya tuvieron alguna negociación sobre el regreso a clases?

Fin (\$)

Conteo con Regex

"\?\$" = 3

Conteo Sin Regex

"\?" = 4

interlocutora: ¿el programa aprende en casa
continuará si los padres deciden no llevarlos?  





presidente andrés manuel lópez obrador: sí, sí va a
continuar, estamos contemplando eso.

interlocutora: y si hay una propuesta para los
trayectos, para el cuidado.

presidente andrés manuel lópez obrador: sí,
vamos a que haya seguridad en todo lo que se
requiera.

interlocutora: ¿en qué consistirá?  

presidente andrés manuel lópez obrador: vamos a
dar a conocer el plan para que haya seguridad. tiene
que protegerse a los niños, a las maestras, a los
maestros, las escuelas. eso corresponde a los
gobiernos estatales, siempre lo hacen y ahora
vamos a pedir que se apliquen más.

interlocutora: y, bueno, una última pregunta,
presidente, por favor. ¿qué pasó con la coordinadora
nacional de trabajadores de la educación?  
¿tuvieron alguna negociación sobre el regreso a
clases?  

Caracteres

\d	Dígitos
\w	Alfanuméricos
\s	Espacios en blanco
\D	Todo menos dígitos
\W	Todo menos alfanuméricos
\S	Todo menos espacios en blanco

Cuantificadores

?	Cero o una vez
+	Una o más veces
*	Cero o más veces
{2}	Dos veces exactas
{2,3}	Dos o tres veces
{2,}	Dos o más veces

Lógicas

 	O (<i>or</i>)
[az]	la a o la z
[a-z]	Uno entre la a y la z
[^a-z]	Uno que no esté entre la a y la z
[a-z]+	Uno o más entre la a y la z
[a-z]{2}	Dos entre la a y la z

Portable Operating System Interface

POSIX

POSIX class	Equivalente	Significa
<code>[:upper:]</code>	<code>[A-Z]</code>	Mayúsculas
<code>[:lower:]</code>	<code>[a-z]</code>	Minúsculas
<code>[:alpha:]</code>	<code>[A-Za-z]</code>	Cualquier letra
<code>[:digit:]</code>	<code>[0-9]</code>	Dígitos
<code>[:xdigit:]</code>	<code>[0-9A-Fa-f]</code>	hexadecimales
<code>[:alnum:]</code>	<code>[A-Za-z0-9]</code>	Alfanumericos
<code>[:punct:]</code>		Puntuación
<code>[:blank:]</code>	<code>[\t]</code>	Espacios y tabuladores
<code>[:space:]</code>	<code>[\t\n\r\f\v]</code>	Espacios en blanco

Te recomiendo

<https://regexlearn.com/>

<https://regexone.com/>