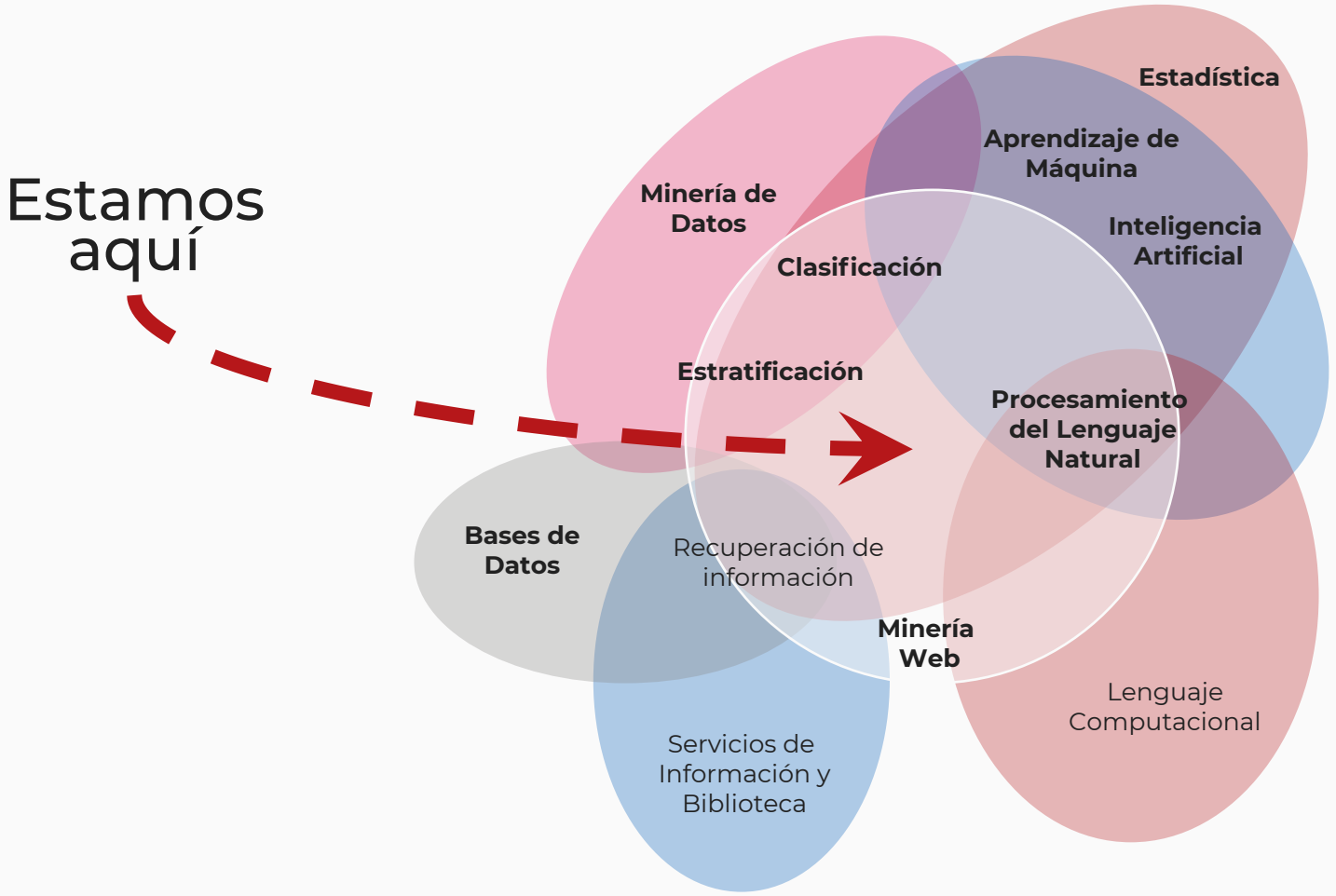
Tecnológico de Monterrey
Escuela de Ciencias Sociales
y Gobierno

Análisis de Frecuencias

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 adipisicing 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

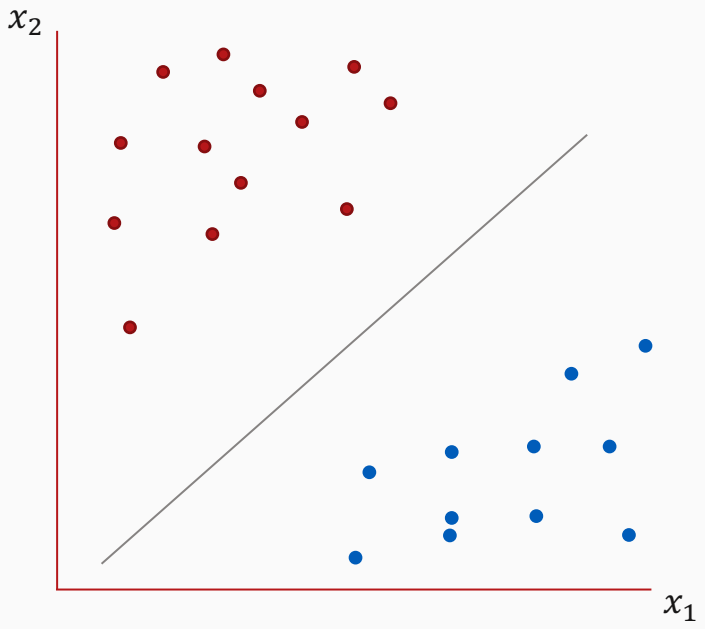
Lorem

Modelos Probabilísticos

Probabilistic Models

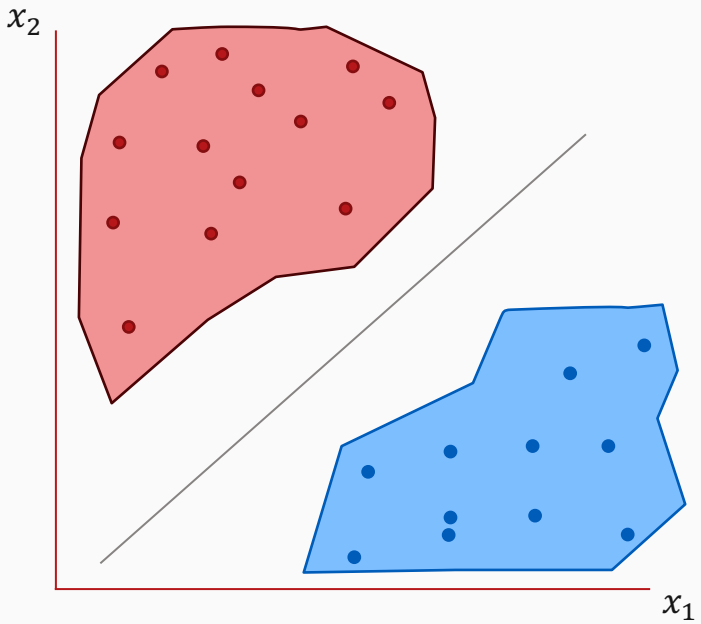
Discriminativos

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 consideramo la probabilidad de ocurrencia de una palabra en una frase respecto a las palabras inmediateas, 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: <s>*Una vaca vestida de uniforme*</s>

Bigrama:

$$P(Una| < s >)P(vaca|una)P(vesida|vaca)P(de|vestida)P(uniforme|de)P(</s > |vestida)$$

Trigrama:

$$P(vaca| < s > Una)P(vestida|una\ vaca)P(de|vaca\ vestida)P(uniforme|vestida\ de)P(</s > |de\ uniforme)$$

Bolsa de Palabras

Bag of Words

- Es la representación más simple de un conjunto de documentos.
- Cada documento es una bolsa con un conjunto de palabras *(tokens)*.
- El orden de las palabras no importa.
- Conserva la frecuencia de las palabras de forma cruda.
- El conjunto de todas las palabras del modelo forma un diccionario.

Bolsa de Palabras

Bag of Words

Oración1 {Era el mejor de los tiempos, era el peor de los tiempos, la edad de la sabiduría, y también de la locura; la época de las creencias y de la incredulidad; la era de la luz y de las tinieblas; la primavera de la esperanza y el invierno de la desesperación.}

BoW1 {Era : 3, el : 3, mejor : 1, de : 10, los : 2, tiempos 2: , peor : 1, la : 10, edad : 1, sabiduría : 1, y : 4, también : 1 , locura : 1, época : 1, las : 2, creencias : 1, incredulidad : 1, luz : 1, tinieblas : 1, primavera : 1, esperanza : 1, invierno : 1, desesperación : 1 }

Oración2 { Todo lo poseíamos, pero no teníamos nada; caminábamos en derechura al cielo y nos extraviábamos por el camino opuesto. En una palabra, aquella época era tan parecida a la actual, que nuestras más notables autoridades insisten en que, tanto en lo que se refiere al bien como al mal, sólo es aceptable la comparación en grado superlativo.}

BoW2 { Todo : 1, lo : 2, poseíamos : 1, pero : 1, no : 1, teníamos : 1, nada : 1, caminábamos : 1, en : 5 , derechura : 1, al 3: , cielo 1: , y : 1, nos : 1, extraviábamos : 1, por : 1, el : 1, camino:1, opuesto : 1, una : 1, palabra : 1, aquella : 1, época : 1, era : 1, tan : 1, parecida : 1, a : 1, la : 2, actual : 1, que : 3, nuestras : 1, más. : 1, notables : 1, autoridades : 1, insisten : 1, tanto: 1, se : 1, refiere : 1, bien : 1, como : 1, mal : 1, sólo : 1, es : 1, acceptable : 1, comparación : 1, grado : 1, superlativo : 1}

Matriz de Términos y Documentos

Term-Document Matrix

Es una matriz matemática que describe la frecuencia de términos que ocurren en una colección de documentos

	TOKEN-1	**TOKEN-2**	**...**	**TOKEN-N**
DOC 1	1	1	...	0
DOC 2	0	0	...	0
DOC 3	0	1	...	1
DOC 4	1	0	...	1
...	...	...	...	...
DOC N	1	1	...	1

Frecuencia de Términos

Term Frequency (TF)

El peso de un término que aparece en un documento es simplemente proporcional a la frecuencia del término.

$$TF = \frac{w_i}{\sum_n^N w_n}$$

Donde w_i es la palabra y $\sum_n^N w_n$ es el total de palabras dentro de un documento.

Frecuencia de Términos

Term Frequency (TF)

Doc1 : Este ejemplo es un buen ejemplo.
Doc2 : Este es un ejemplo.
Doc3 : Un ejemplo no es bueno si no es claro.

	este	ejemplo	es	un	buen	no	si	claro	bueno
DOC 1	$\frac{1}{2}$	$\frac{2}{4}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{1}$	0	0	0	0
DOC 2	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{3}$	0	0	0	0	0
DOC 3	0	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{3}$	0	$\frac{1}{1}$	$\frac{1}{1}$	$\frac{1}{1}$	$\frac{1}{1}$

Frecuencia Inversa del Documento

Inverse Documen Frequency (IDF)

Nos dice la información que aporta un término en relación inversa al número de documentos en los que aparece

$$IDF = \log\left(\frac{N}{n_{w_i}}\right)$$

Donde N es el total de documentos y n_{w_i}es el número de documentos donde aparece el término

Frecuencia Inversa del Documento

Inverse Documen Frequency (IDF)

Doc1 : Este ejemplo es un buen ejemplo.
Doc2 : Este es un ejemplo.
Doc3 : Un ejemplo no es bueno si no es claro.

	este	ejemplo	es	un	buen	no	si	claro	bueno
DOC 1	$log\left(\frac{3}{2}\right)$	0	0	0	$log\left(\frac{3}{1}\right)$	$log\left(\frac{3}{1}\right)$	$log\left(\frac{3}{1}\right)$	$log\left(\frac{3}{1}\right)$	$log\left(\frac{3}{1}\right)$
DOC 2	$log\left(\frac{3}{2}\right)$	0	0	0	$log\left(\frac{3}{1}\right)$	$log\left(\frac{3}{1}\right)$	$log\left(\frac{3}{1}\right)$	$log\left(\frac{3}{1}\right)$	$log\left(\frac{3}{1}\right)$
DOC 3	$log\left(\frac{3}{2}\right)$	0	0	0	$log\left(\frac{3}{1}\right)$	$log\left(\frac{3}{1}\right)$	$log\left(\frac{3}{1}\right)$	$log\left(\frac{3}{1}\right)$	$log\left(\frac{3}{1}\right)$

TF-IDF

TF-IDF

Refleja la importancia de una palabra para un registro en una colección o corpus.

$$TF * IDF = \frac{w_i}{\sum_n^N w_n} * \log\left(\frac{N}{n_{w_i}}\right)$$

TF-IDF

TF-IDF

Doc1 : Este ejemplo es un buen ejemplo.
Doc2 : Este es un ejemplo.
Doc3 : Un ejemplo no es bueno si no es claro.

	este	ejemplo	es	un	buen	no	si	claro	bueno
DOC 1	$\frac{log\left(\frac{3}{2}\right)}{2}$	0	0	0	$log\left(\frac{3}{1}\right)$	0	0	0	0
DOC 2	$\frac{log\left(\frac{3}{2}\right)}{2}$	0	0	0	0	0	0	0	0
DOC 3	0	0	0	0	0	$log(3)$	$log(3)$	$log(3)$	$log(3)$

Lematización

Lemmatization

- Es un proceso lingüístico que consiste en, dada una forma flexionada (plural, en femenino, conjugada, etc), hallar la raíz (lema) correspondiente.
- El lema es la forma que por convenio se acepta como representante de todas las formas flexionadas de una misma palabra.
- Lematizar implica estandarizar, desambiguar, segmentar y, en caso de usar programas de lematización automática, también etiquetar.

Lematización

Lemmatization

Doc1 : Nos vamos a ir yendo
Doc2 : Te vas a ir yendo
Doc3 : Me voy a ir yendo

TDM con tokens

	nos	vamos	a	ir	yendo	te	vas	me	voy
DOC 1	1	1	1	1	1	0	0	0	0
DOC 2	0	0	1	1	1	1	1	0	0
DOC 3	0	0	1	1	1	0	0	1	1

Lematización

Lemmatization

Doc1 : Nos vamos a ir yendo
Doc2 : Te vas a ir yendo
Doc3 : Me voy a ir yendo

Doc1 : Nosotros ir a ir ir
Doc2 : Tú ir a ir ir
Doc3 : Yo a ir ir

TDM con lemas

	nosotros	**ir**	**a**	**tú**	**yo**
DOC 1	1	3	1	0	0
DOC 2	0	3	1	1	0
DOC 3	0	3	1	0	1

Lematización

Lemmatization

Doc1 : Nos vamos a ir yendo
Doc2 : Te vas a ir yendo
Doc3 : Me voy a ir yendo

Doc1 : Nosotros ir a ir ir
Doc2 : Tú ir a ir ir
Doc3 : Yo ir a ir ir

TDM con TF-IDF de los lemas

	nosotros	**ir**	**a**	**tú**	**yo**
DOC 1	$\frac{\log(1/3)}{5}$	0	0	0	0
DOC 2	0	0	0	$\frac{\log(1/3)}{5}$	0
DOC 3	0	0	0	0	$\frac{\log(1/3)}{5}$

Derivación

Stemming

- Es un método para reducir todas las formas flexionadas de palabras a una raíz o tallo (stem) cuando estas comparten una misma raíz.

- A diferencia de la lematización, en donde cada lema es una palabra que existe en el vocabulario del lenguaje correspondiente, los tallos que se obtienen no necesariamente existen como palabra.

El estudiante estudia sus estudios en el estudio.

El estudi estudi sus estudi en el estudi.