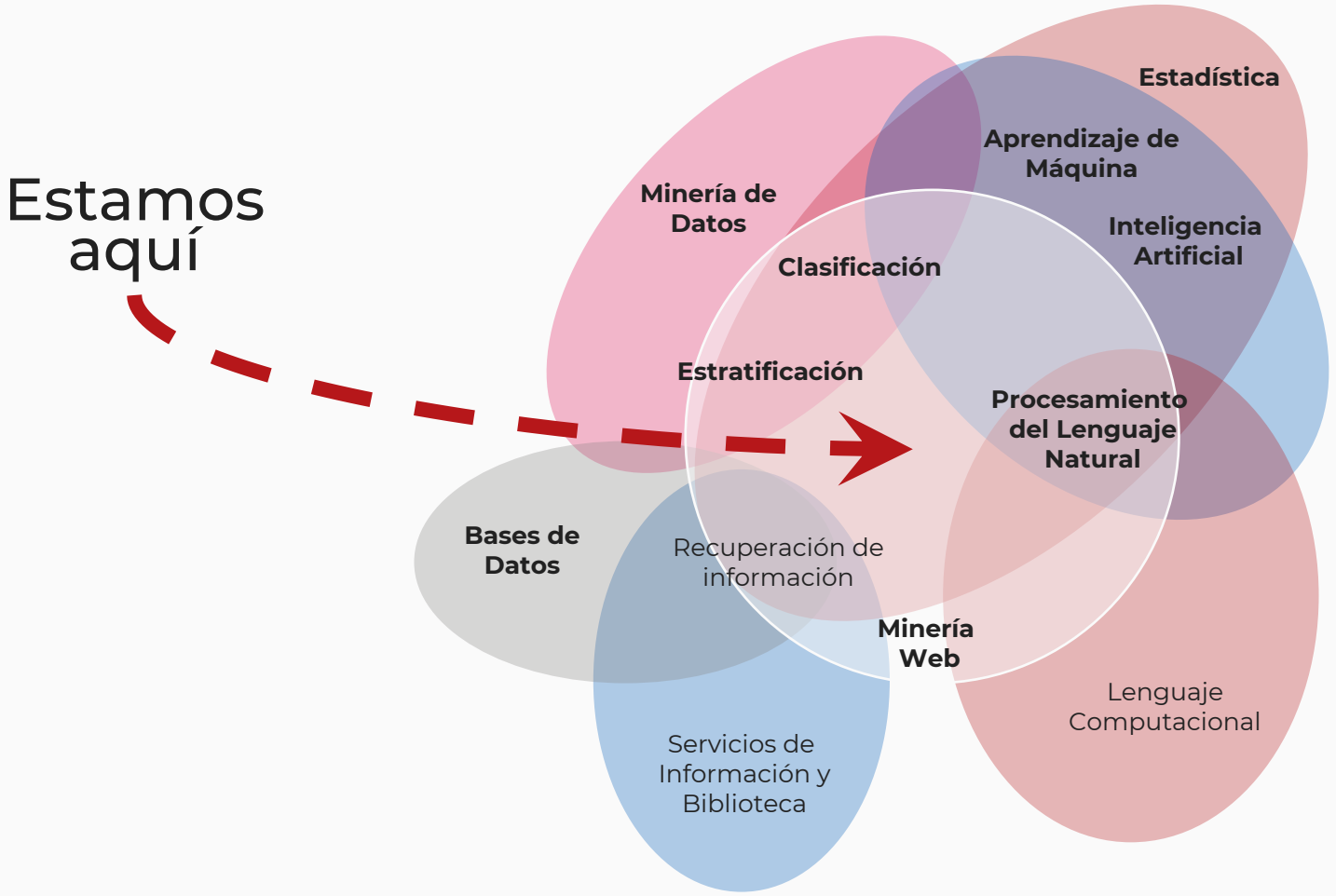
Tecnológico de Monterrey
Escuela de Ciencias Sociales
y Gobierno

Introducción a 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 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

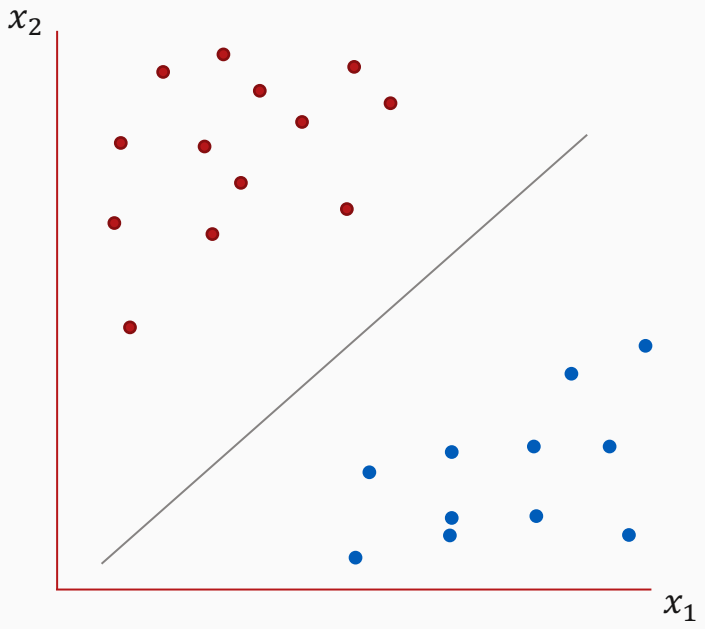
Lorem

Modelos Probabilísticos

Probabilistic Models

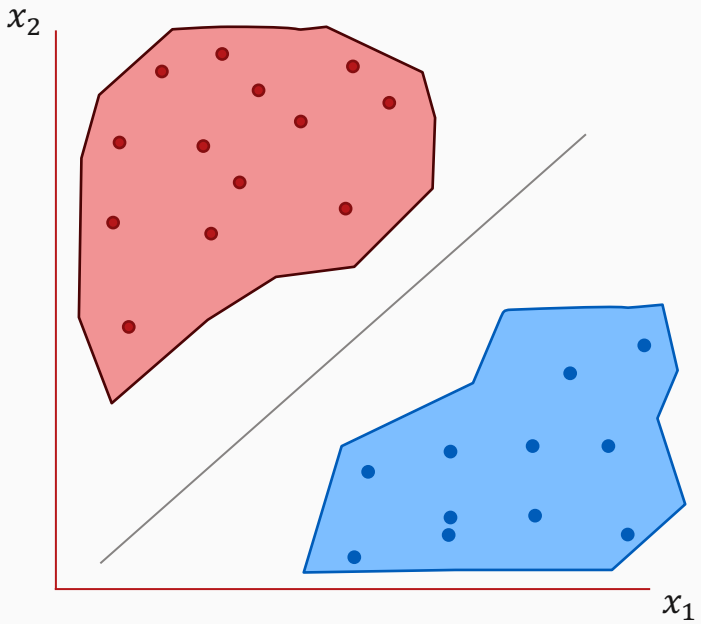
Discriminativos

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:

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