Modelado de tópicos

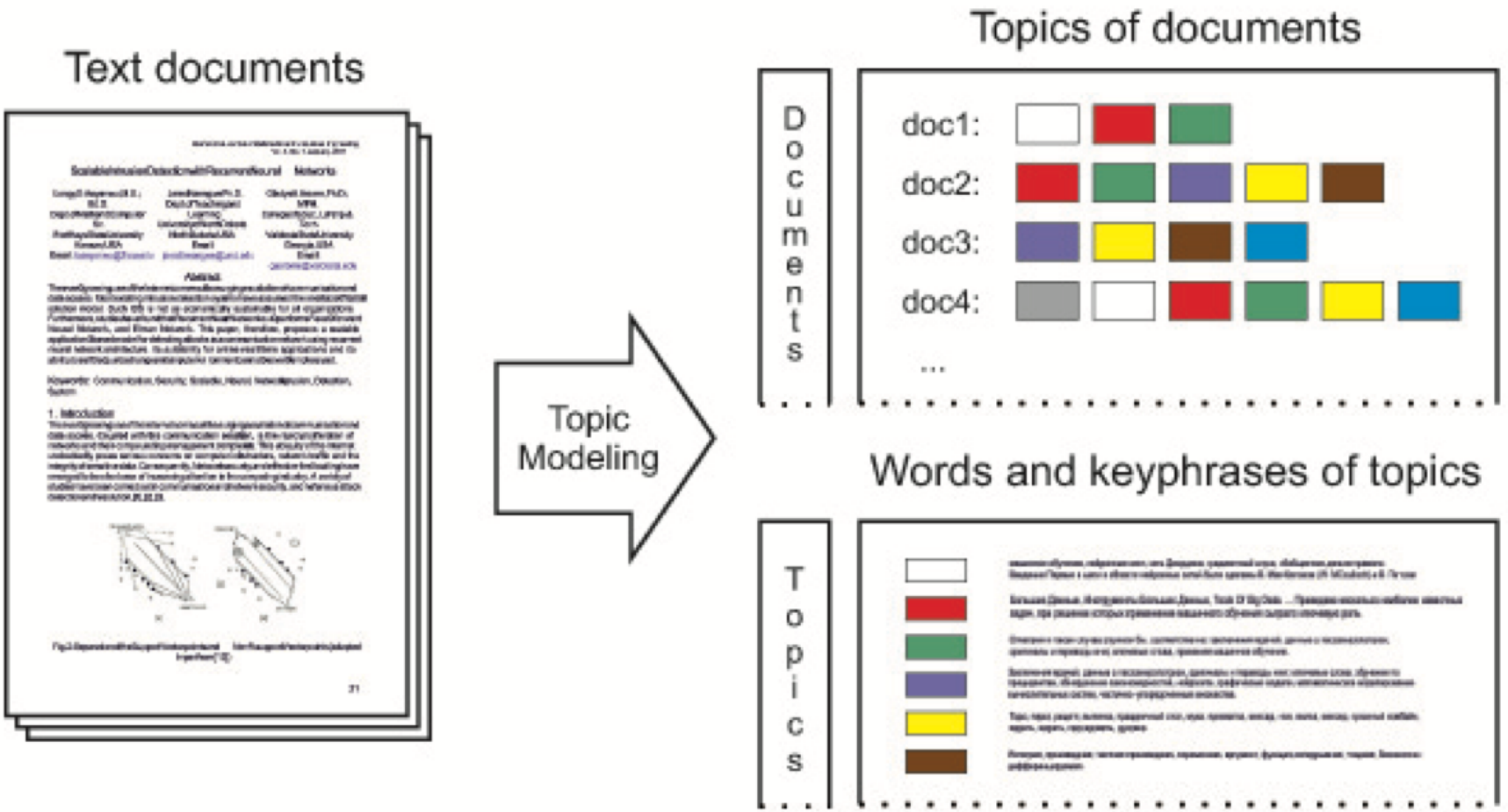
Curso de procesamiento de lenguaje natural

Olivia Gutú y Julio Waissman

Maestría en Ciencia de Datos
Universidad de Sonora

5 de mayo de 2021

¿Que es modelado de tópicos?



Definición formal

- Dado:
  - $\{d_1, \ldots, d_D\}$ un conjunto de documentos (corpus),
  - $\{w_1, \ldots, w_W\}$ un conjunto de palabras (vocabulario),
  - n_{dw} el número de veces que la palabra w aparece en el documento d
- Encontrar:
  - Para un conjunto de T tópicos $\{t_1, \ldots, t_T\}$
  - $\phi_{wt} = \Pr(w|t)$ una distribución de palabras en cada tópico,
  - $\theta_{td} = \Pr(t|d)$ una distribución de tópicos por cada documento
- Basado en las hipótesis:
  - Un *tópico* es un conjunto coherente de palabras que co-ocurren en un subconjunto de documentos
  - Un documento está representado con una BOW con cuentas (o proporcional)
  - Toda palabra observada en un documento tiene un *tópico latente*

¿Para qué sirve el modelado de tópicos?

El modelado de tópicos provée una representación semántica inherente a un conjunto de documentos

Se utiliza en:

1. Categorización de textos
2. Agregación y resumen de noticias
3. Sistemas de recomendación
4. Recuperación de la información
5. Segmentación de corpus

Modelo probabilístico

1. Ley de probabilodad total (marginalización)

$$\Pr(w) = \sum_{t \in T} \Pr(w|t) \Pr(t)$$

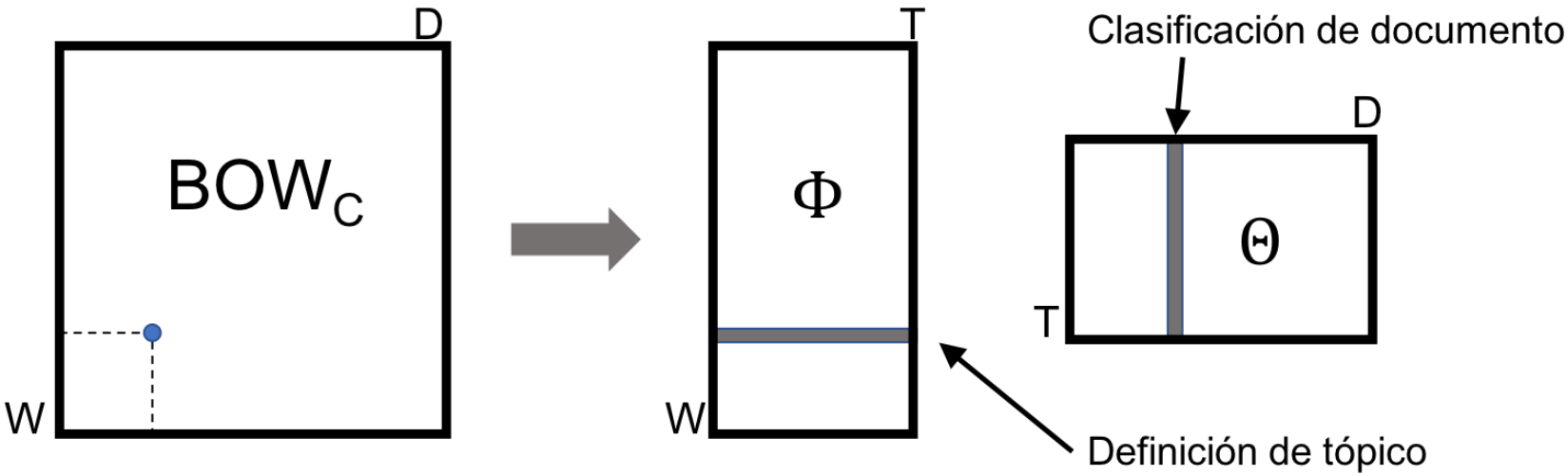
2. Hipótesis de independencia condicional

$$\Pr(w|t, d) = \Pr(w|t)$$

Planteamiento

$$\Pr(w|d) = \sum_{t \in T} \Pr(w|t, d) \Pr(t|d) = \sum_{t \in T} \Pr(w|t) \Pr(t|d) = \phi_{wt}\theta_{td}$$

Visto en forma matricial



El problema de descomposición matricial en este caso está pobremente definido, y hay que utilizar algún criterio de optimización para encontrar las matrices Φ y Θ

PLSA

Se optimiza la verosimilitud logarítmica

$$\Phi^*, \Theta^* = \arg\max_{\Phi,\Theta} \sum_d \sum_{w \in d} n_{dw} \log \sum_t \phi_{wt}\theta_{td}$$

Si conociéramos los tópicos sería muy parecido a los vectores de palabra, pero resulta que solo sabemos el número de tópicos que imponemos

Algoritmo EM

Expectation

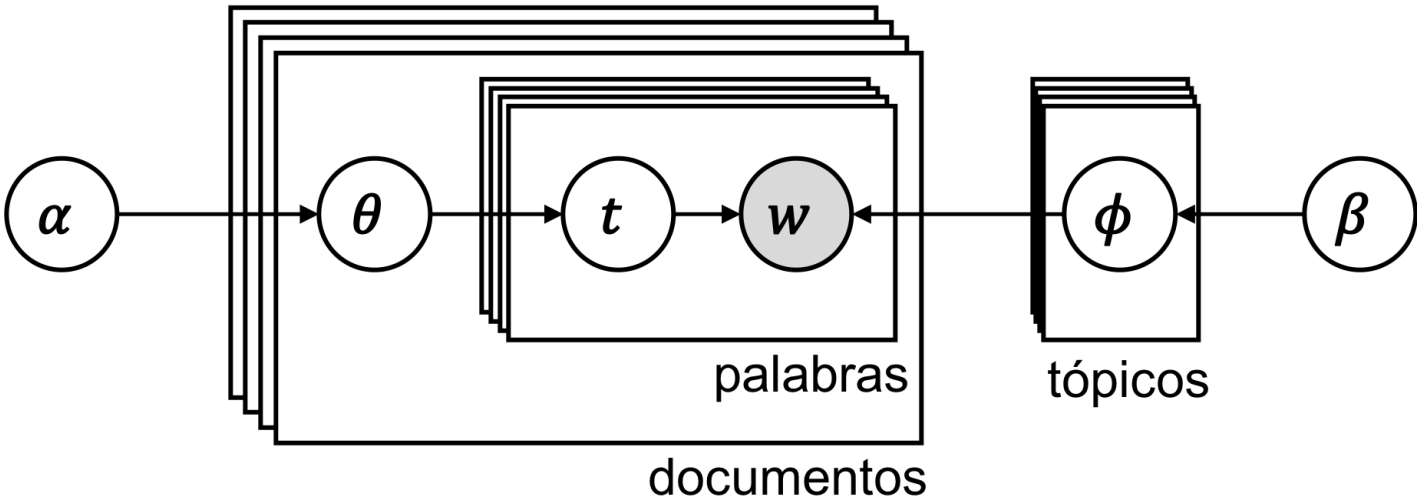
$$\Pr(t|d,w) = \frac{\Pr(w|t)\Pr(t|d)}{\Pr(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}$$

Maximization

$$\phi_{wt} = \frac{n_{wt}}{\sum_v n_{vt}} \quad \text{donde} \quad n_{wt} = \sum_d n_{dw}\Pr(t|d,w)$$

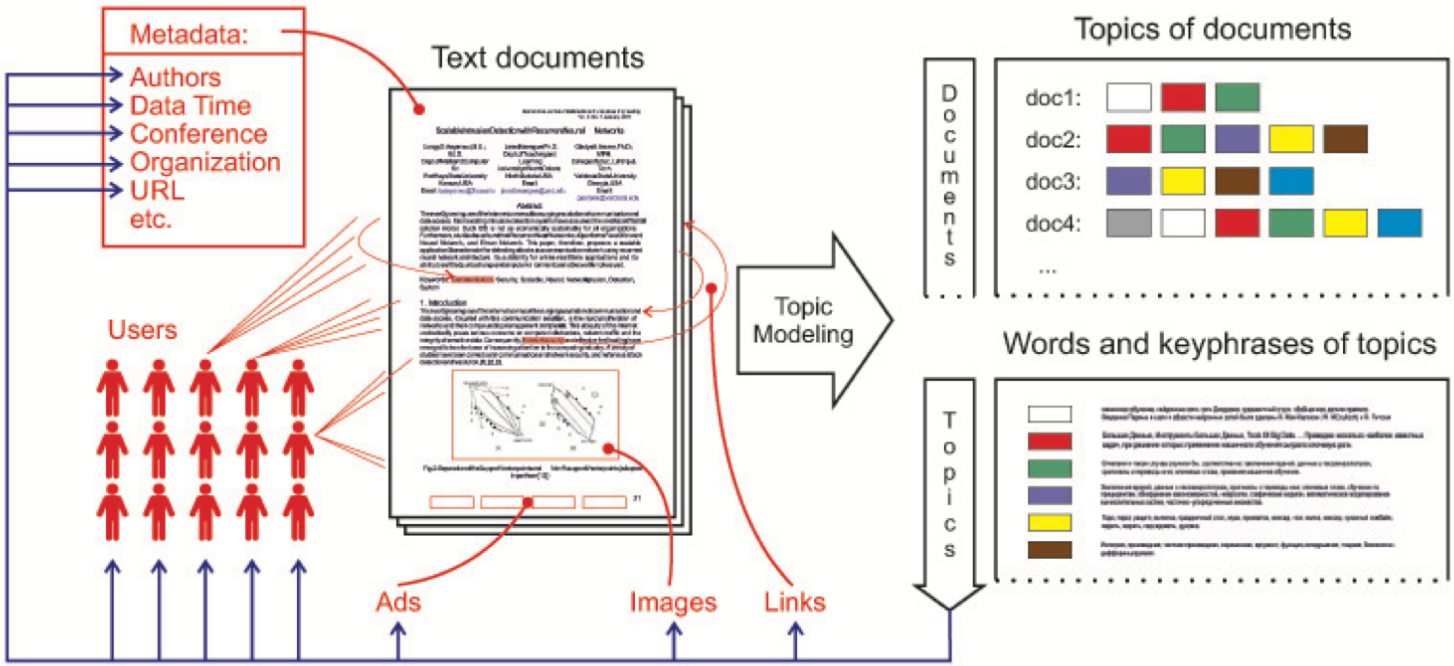
$$\theta_{td} = \frac{n_{td}}{\sum_{t'} n_{t'd}} \quad \text{donde} \quad n_{td} = \sum_w n_{dw}\Pr(t|d,w)$$

Latent Dirichlet Allocation (LDA)



1. La distribución de palabras para el tópico t (ϕ_t, el t-ésimo un renglón de Φ) es generada por una distribución de *Dirichlet* con parámetros $\beta \in \mathbb{R}^W$
2. La distribución de tópicos para el documento d (θ_d una columna de Θ) también se genera a partir de una distribución de *Dirichlet* con parámetros $\alpha \in \mathbb{R}^T$

Modelado de tópicos multimodal



Biblioteca especializada *BigARTM* en http://bigartm.org

Revisemos el modelado de tópicos

Ejemplo con los sonetos del siglo de oro español
https://github.com/juliowaissman/lda-ejemplo