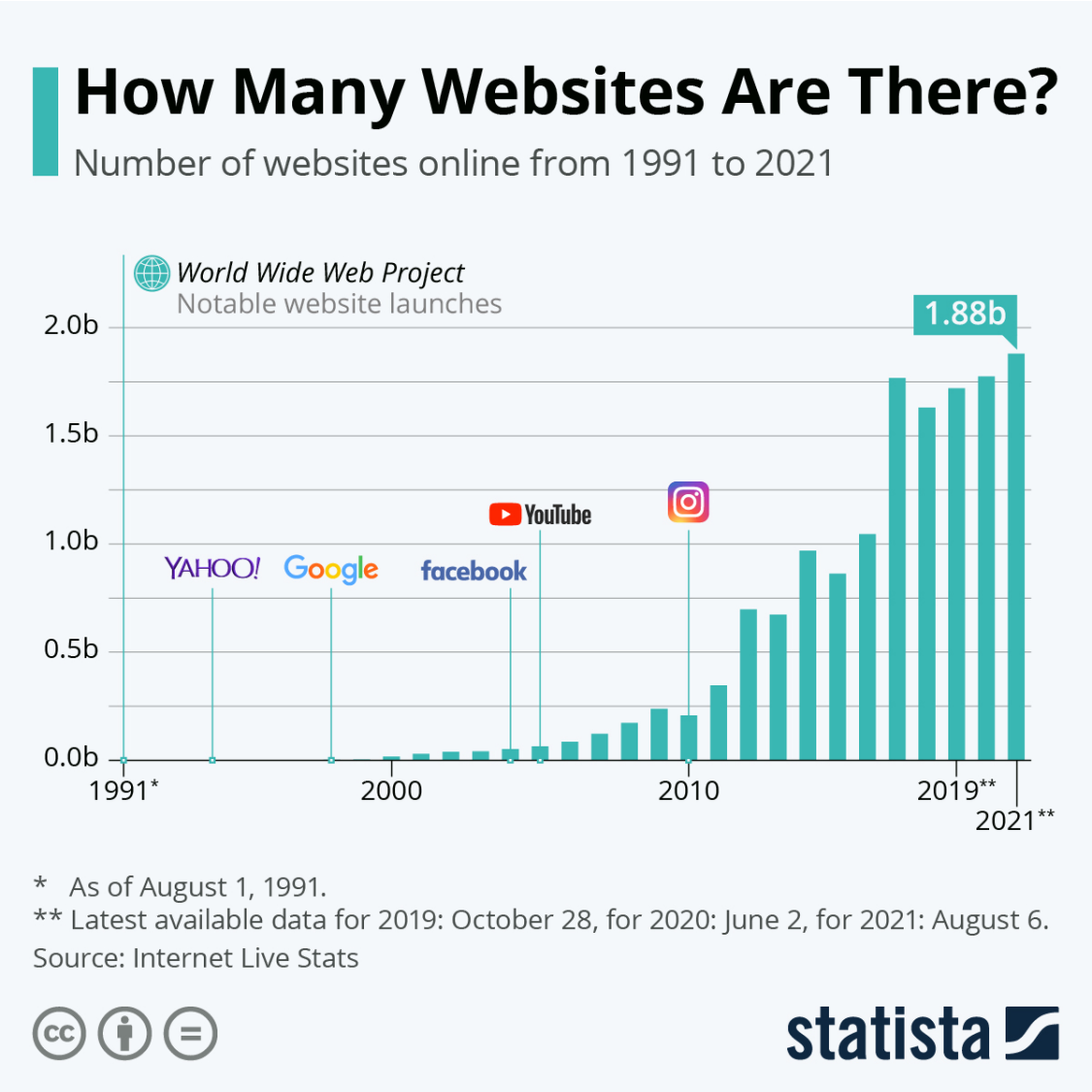
Modelado de Temas

Luis Norberto Zúñiga Morales

7 de abril de 2022

Introducción



Introducción

Consideren la cantidad de noticias en un sitio web como el New York Times o similares.

- ¿Cómo pueden obtener el contenido de interés?
- ¿Cómo explorar un tema en particular considerando toda la colección de documentos?

Introducción

¡Alerta!

Mientras más y más textos están disponibles en línea, nosotros no tenemos el poder humano para leerlos y estudiarlos para proporcionar el tipo de experiencia de navegación que se desea.

Introducción

Con este fin, se ha desarrollado el modelado de temas (probabilístico).

Definición: Modelado de Temas

Conjunto de algoritmos de aprendizaje automático cuyo fin es descubrir y anotar grandes conjuntos de datos con información temática [1].

Introducción

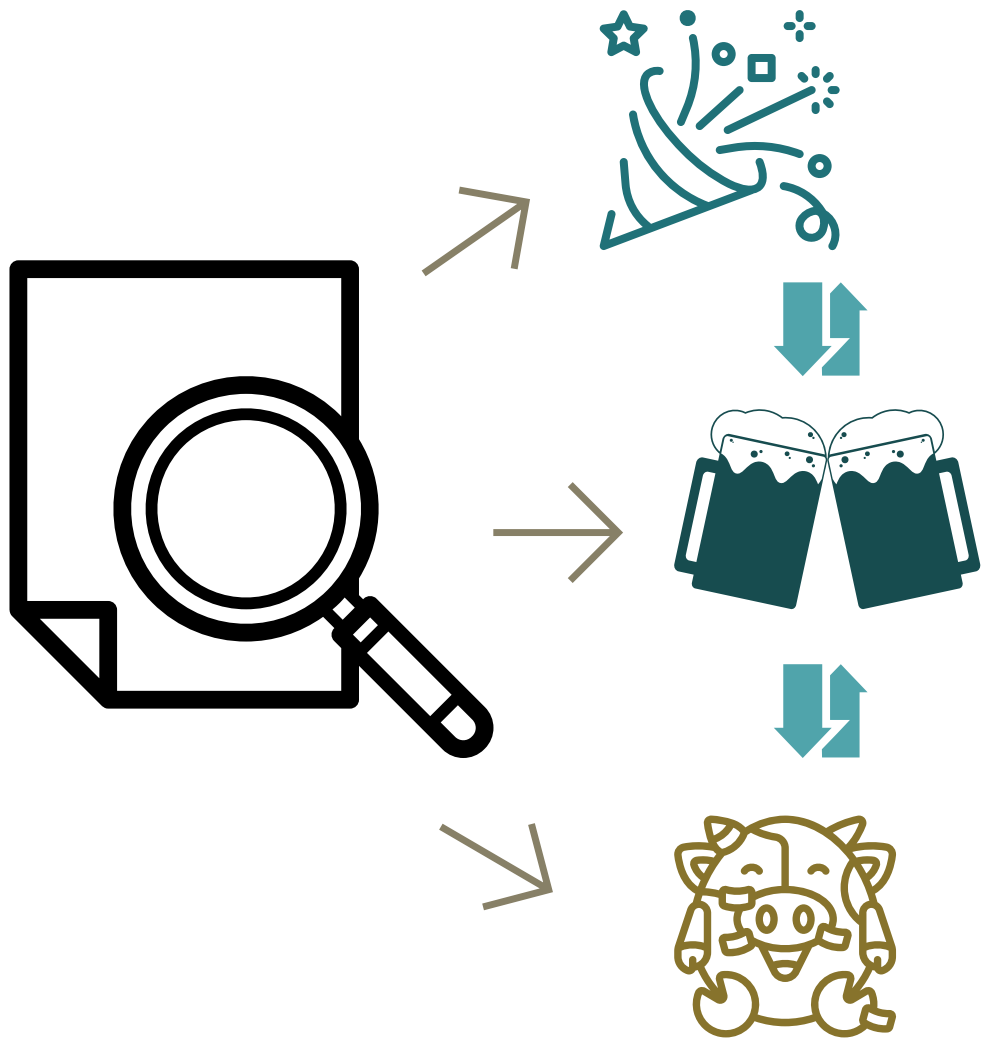


Figura: Idea general del modelado de temas: descubrir temas, cómo interactúan entre ellos y cómo evolucionan con el tiempo.

Introducción

- El modelado de tema son algoritmos que permiten descubrir temas en una colección sin estructura de documentos.
- Se pueden aplicar a una colección masiva de documentos.
- No son exclusivos de PLN: se pueden emplear para encontrar patrones en información genética, imágenes y redes sociales.

Latent Dirichlet Allocation

Por otro lado, la nutrióloga comentó que la anemia, a nivel mundial, sigue siendo un problema de salud pública. En México, **aproximadamente 30% de las mujeres cursan su embarazo con anemia**, un problema que puede provocar la pérdida del o la bebé. Otra consecuencia es que las y los niños pueden nacer con bajo peso y puede afectar su neurodesarrollo. Además de que las y los bebés corren el riesgo de también presentar anemia en sus primeros años de vida.

⋮

Como parte de esta colaboración entre la IBERO y la Universidad de Copenhague, la cual cuenta con el financiamiento de la Agencia Danesa de Desarrollo Internacional (Danida), este año iniciarán con la **recolección de información de 600 mujeres embarazadas** atendidas en el IMSS de Morelos, esto con apoyo del Instituto Nacional de Salud Pública.

Tema 1
Salud

Tema 2
Colaboración
Académica

Tema 3
Estadística

Figura: Idea general detrás de Latent Dirichlet Allocation. Artículo: Ibero.

Latent Dirichlet Allocation

Definición: Tema

Un tema es una distribución de probabilidad sobre cierta colección de palabras que conforman un vocabulario.

Latent Dirichlet Allocation

LDA es un método de modelado probabilístico:

- En el modelado probabilístico generativo, la información se considera que surge de un proceso generativo que incluye variables ocultas.
- El proceso generativo define una función de probabilidad conjunta sobre las variables aleatorias ocultas y observadas.
- Se realiza análisis de datos utilizando la función de probabilidad conjunta para determinar la distribución condicional de las variables ocultas dadas las variables observadas.
- Lo anterior se llama distribución posterior.

Latent Dirichlet Allocation

Cómo se ajusta en el marco de LDA:

- Las variables observadas son las palabras en los documentos.
- Las variables ocultas son las estructuras que dan forma a los distintos temas de los documentos.
- El problema de determinar la estructura oculta de los temas es el problema de calcular la distribución posterior.

Latent Dirichlet Allocation

Notación para una definición formal:

- $\beta_{1:K}$: los temas, donde β_k es una distribución sobre el vocabulario.
- θ_d: es la proporción de temas para el d-ésimo documento.
- $\theta_{d,k}$: es la proporción del tema k en el documento d.
- z_d: la asignación de temas para el documento d.
- $z_{d,n}$: es la asignación del tema para la n-ésima palabra en el documento d.
- w_d: las palabras observadas en el documento d.
- $w_{d,n}$: la n-ésima palabra en el documento d, un elemento de un vocabulario fijo.

Latent Dirichlet Allocation

Con las variabales anteriores, el proceso generativo para LDA corresponde a la siguiente distribución conjunto de variables ocultas y observadas:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{n} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n}) \right)$$

Latent Dirichlet Allocation

Document Word Matrix

	W1	W2	W3	W4	W5	W6	W7	W8
D1	0	1	1	0	1	1	0	1
D2	1	1	1	1	0	1	1	0
D3	1	0	0	0	1	0	0	1
D4	1	1	0	1	0	0	1	0
D5	0	1	0	1	0	0	1	0

Latent Dirichlet Allocation

Document Term Matrix (DTM)

	W1	W2	W3	W4	W5	W6	W7	W8
D1	0	1	1	0	1	1	0	1
D2	1	1	1	1	0	1	1	0
D3	1	0	0	0	1	0	0	1
D4	1	1	0	1	0	0	1	0
D5	0	1	0	1	0	0	1	0

Shape: 5 * 8

Document Topic Matrix

	K1	K2	K3	K4	K5	K6
D1	1	0	0	0	0	0
D2	0	1	0	0	1	1
D3	1	1	0	0	0	0
D4	1	0	0	1	0	1
D5	0	0	1	1	0	0

Shape: 5 * 6

Topic Word Matrix

	W1	W2	W3	W4	W5	W6	W7	W8
K1	0	1	1	0	1	0	1	0
K2	1	1	1	1	0	1	1	1
K3	1	0	0	0	0	1	0	0
K4	1	1	0	1	1	0	0	1
K5	0	0	1	1	0	1	1	1
K6	1	0	1	1	1	0	0	1

Shape: 6 * 8

Latent Dirichlet Allocation

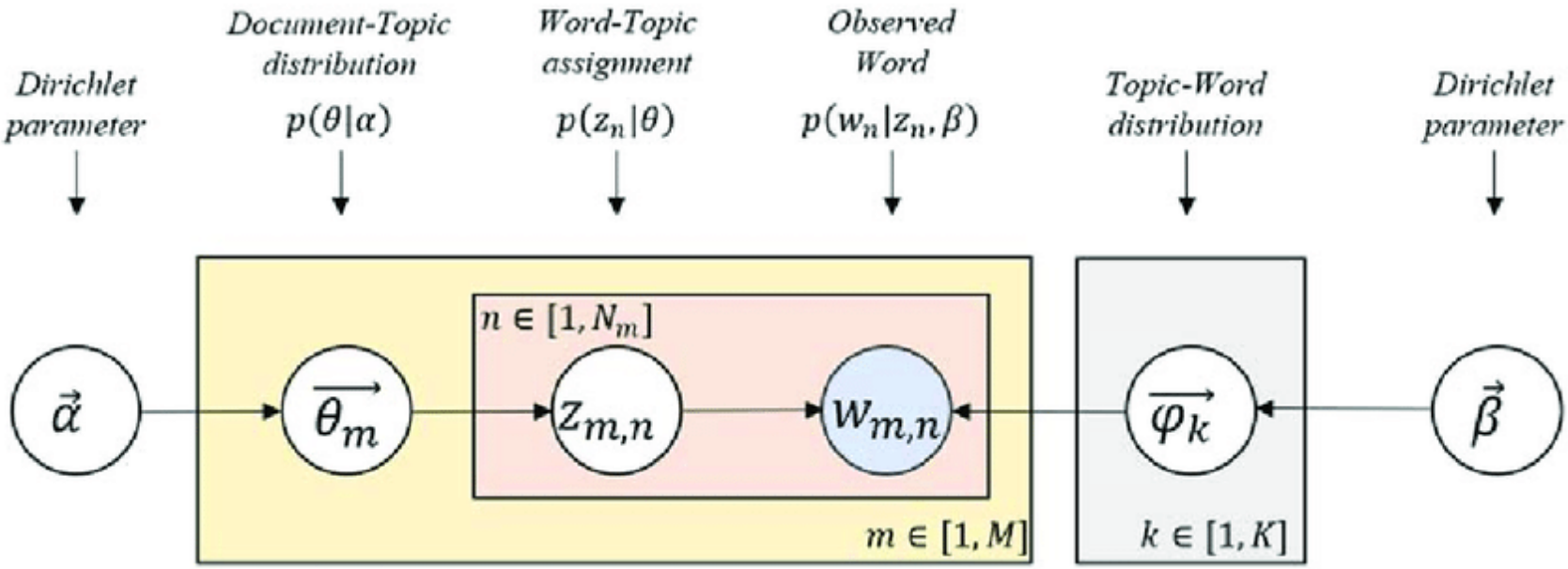
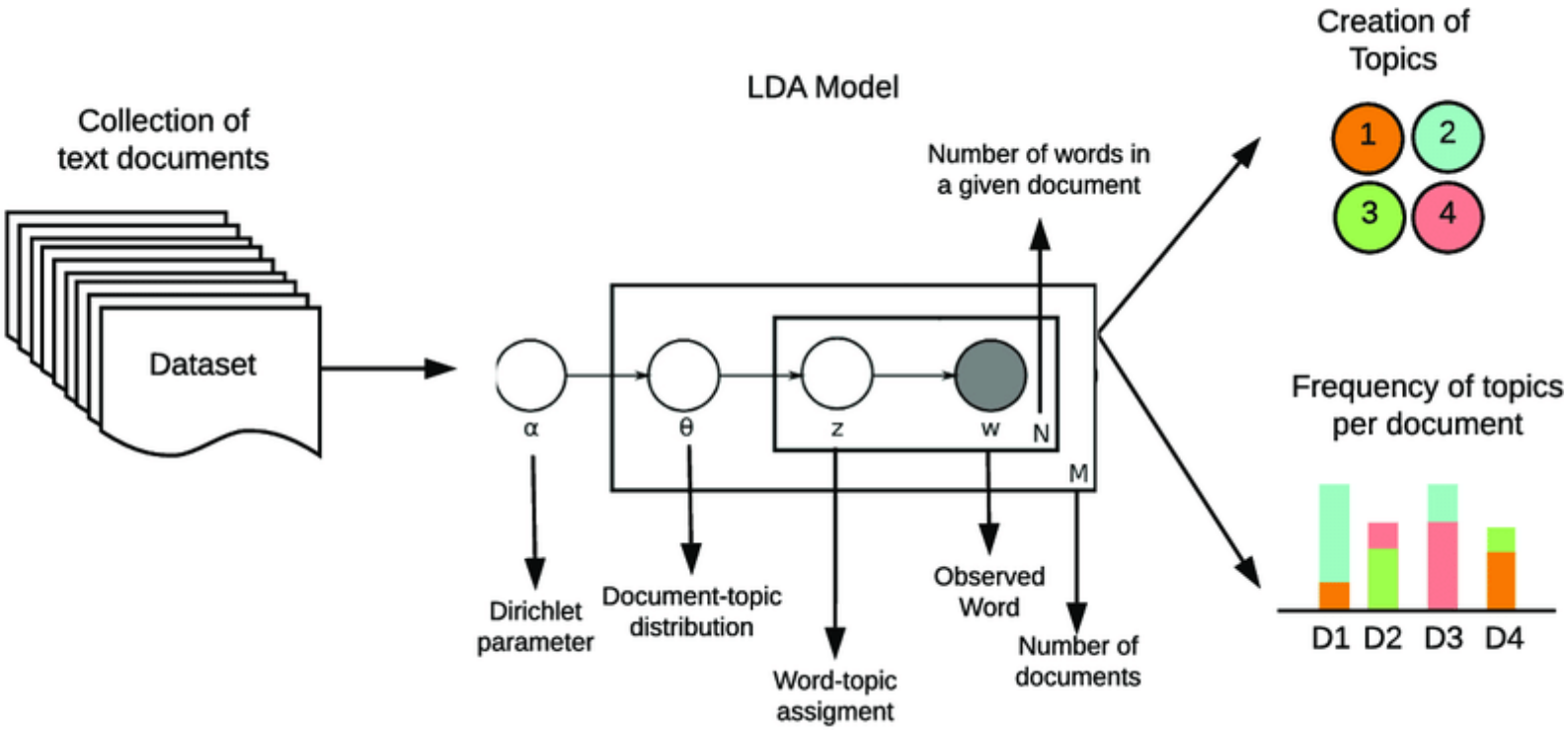


Figura: La caja amarilla representa los documentos en el corpus. La caja rosa es el número de palabras en un documento.

Latent Dirichlet Allocation



Bibliografía

[1] David M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, apr 2012.