

M5. Minería de Texto + webscraping

Clase 5. Modelado de tópicos II STM

Latent Dirichlet Allocation (LDA)

Ventajas de los modelos generativos:

- Supuestos explícitos: si el modelo falla (por ejemplo, no encuentra los tópicos correctos en un corpus bien definido) se puede chequear si es porque los datos no cumplen alguna.
- Variando esas hipótesis/supuestos se producen las extensiones de LDA
- Hoy: Structural Topic Modeling (STM)

Structural Topic Modeling (STM)

- Permitir vincular los tópicos detectados con covariables (o metadata)
- Variación de LDA => modelo generativo del conteo de palabras
- Las covariables pueden afectar la **prevalencia** (*prevalence*) de los tópicos sobre cada documento, es decir, permite que la composición de tópicos sobre cada documento varíe en función de alguna covariable
- También pueden afectar el **contenido** (*content*), es decir, permite que la probabilidad de cada palabra de pertenecer a cada tópico se vea afectada por esas covariables.

Latent Dirichlet Allocation (LDA)

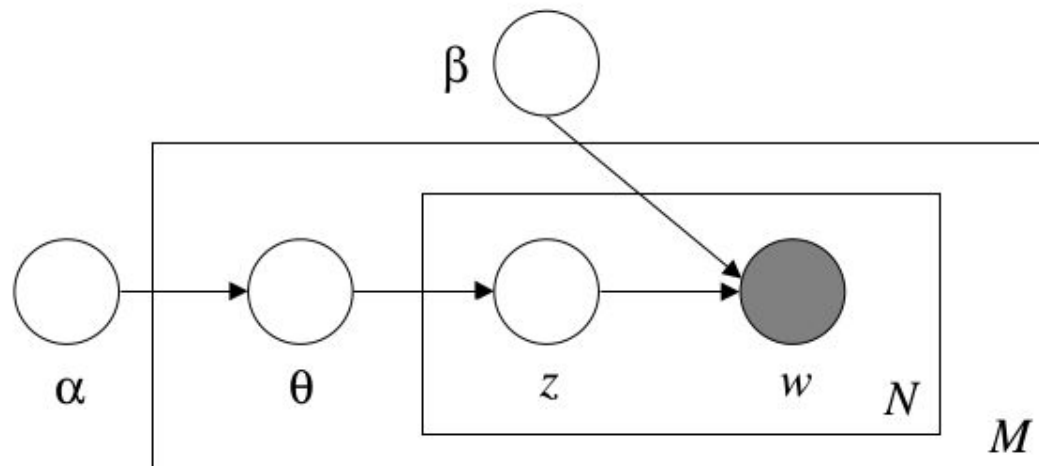
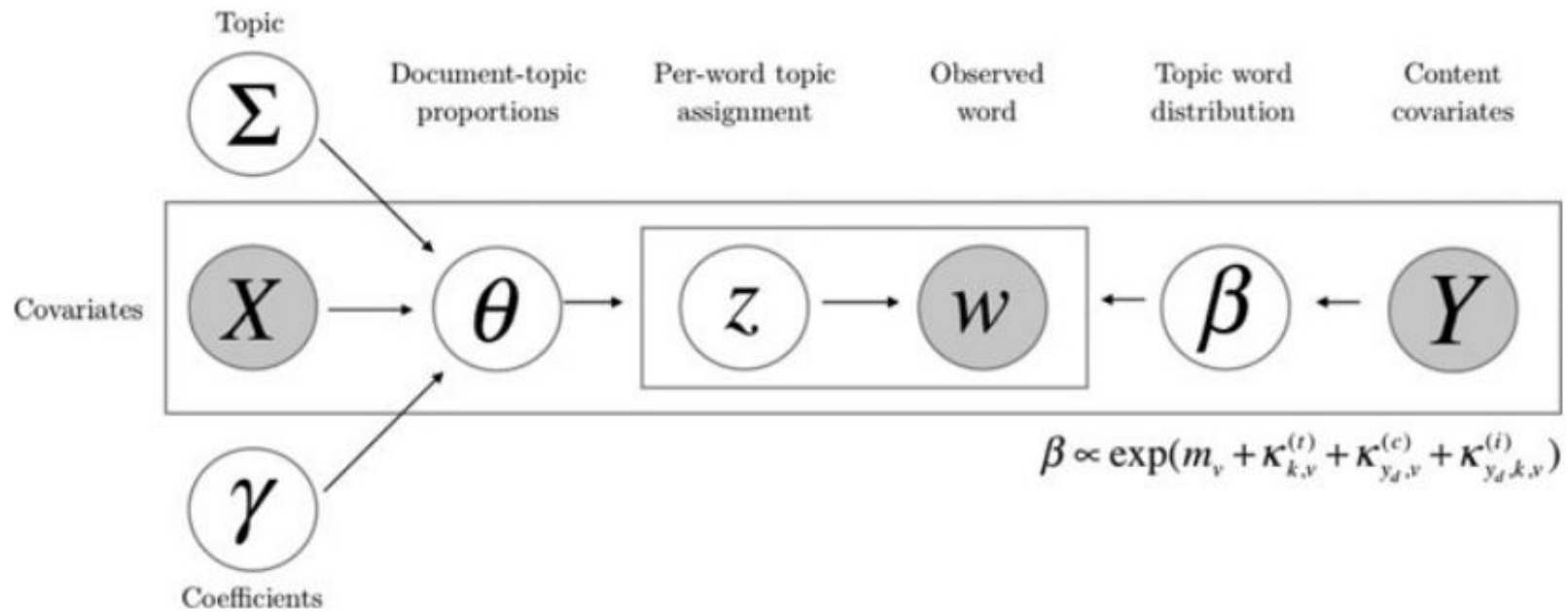
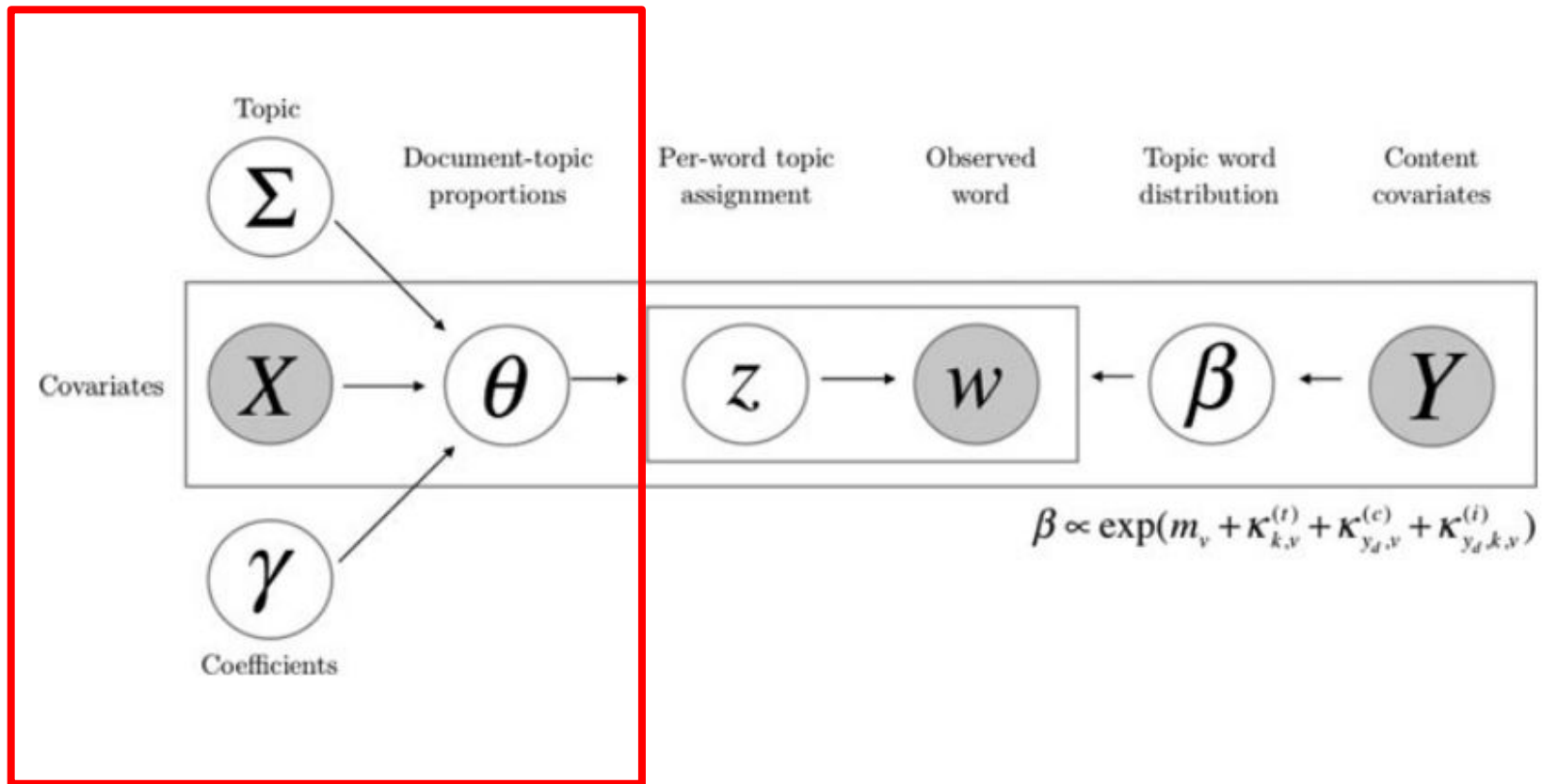
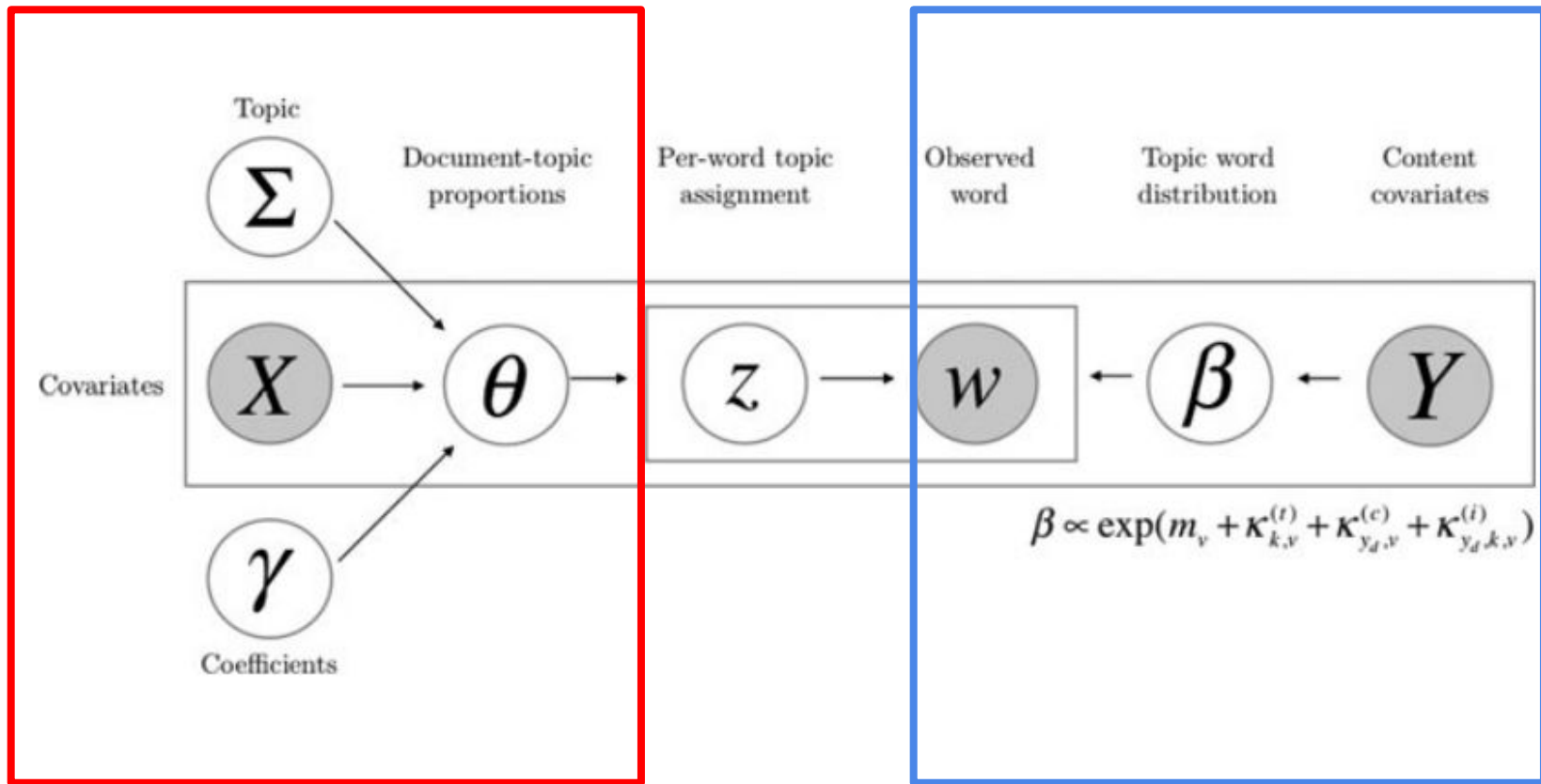


Figure 1: Graphical model representation of LDA. The boxes are “plates” representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.





Prevalencia: X afecta a la composición de tópicos por documento



Prevalencia: X afecta a la composición de tópicos por documento

Contenido: Y afecta a la composición de palabras por tópico

Vamos al notebook...

Un flujo de trabajo posible para topic modeling

1. Definir objetivos

Adecuación entre problemas, preguntas y objetivos de investigación y supuestos del modelo (en este caso, STM/LDA u otros)

Definición conceptual de “tópico”

Ejemplo: “agenda setting”

Un flujo de trabajo posible para topic modeling

1. Definir objetivos

2. Construir un corpus

- Corpus con una distribución “balanceada”
- Evitar sesgos (o al menos conocerlos y cuantificarlos)
- Curar los corpus
- Unidad de análisis (documentos, párrafos, oraciones, etc.)

Un flujo de trabajo posible para topic modeling

1. Definir objetivos
2. Construir un corpus
3. Pre-procesar el texto
 - Filtrado de palabras
 - Stopwords por lista
 - Palabras de alta frecuencia
 - Eliminar caracteres
 - Reducción de complejidad (stemming/lemmatización, etc.)
 - Etc.

Un flujo de trabajo posible para topic modeling

1. Definir objetivos
 2. Construir un corpus
 3. Pre-procesar el texto
 4. Seleccionar el modelo adecuado
- K (nro. de tópicos) es el parámetro fundamental
 - Dispersión de tópicos y palabras son otros opcionales
 - Prueba de diferentes valores de K
 - Algunas métricas (perplexity)

Un flujo de trabajo posible para topic modeling

1. Definir objetivos
 2. Construir un corpus
 3. Pre-procesar el texto
 4. Seleccionar el modelo adecuado
 5. Interpretar y validar el modelo
- Analizar palabras por tópicos
 - Diferentes métricas (palabras más probables, FREX, etc.)
 - Muestrear documentos con altos valores en los diferentes tópicos y realizar una lectura cercana para validar
 - Colapsar tópicos similares (lógica similar a Grounded Theory)

Un flujo de trabajo posible para topic modeling

1. Definir objetivos
2. Construir un corpus
3. Pre-procesar el texto
4. Seleccionar el modelo adecuado
 - Pocos estándares
 - Tabla de palabras por tópico
 - LDAviz
 - Diferentes métricas: relevancia, FREX, exclusividad, etc.
 - Tópicos genéricos vs tópicos exclusivos
5. Interpretar y validar el modelo
6. Comunicar el modelo.