LDA

Latent Dirichlet Allocation

- Modelo creado por Andrew Ng, David Blei y **Michael I. Jordan** durante sus doctorados en Columbia.
- La idea es descubrir los tópicos ocultos que explican las diferencias en una colección de textos.
- Modelos previos (LSI) asignaban un tópico por documento, pero esto no es realista y no se podía usar en documentos nuevos.
- La idea aquí es descubrir tópicos (sin ninguna información previa) usando una red bayesiana que simula un generador de documentos.

Modelo Generativo

- La idea es hacer una suposición que simplifica el proceso de crear documentos.
- Un documento es creado como una **combinación** de tópicos.
- Un tópico es una distribución sobre un vocabulario finito.
- Cada documento exhibe estos tópicos en proporciones diferentes.

Variables ocultas

- Más aún, LDA busca imponer una estructura latente con variables ocultas.
- Las **variables observadas** son los documentos y las palabras.
- Las **variables ocultas** son los tópicos como tal y como los exhibe cada documento.
- La idea es inferior la **probabilidad posterior** de las variables ocultas, dado que observamos las variables observadas.

Distribución Dirichlet

- Es una distribución sobre el conjunto de símpleces que suman 1.
- Es una distribución de probabilidad para generar distribuciones discretas.
- Tiene como parámetro un vector $\overline{\alpha}$ que pondera la probabilidad de cada dimensión en el vector resultante.

$$P(\theta|\overline{\alpha}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma\left(\alpha_i\right)} \prod_i \theta_i^{\alpha_i - 1}.$$

Lógica del modelo LDA

1. Para cada Tópico,
   - Escoja una distribución sobre el vocabulario $\beta_k \sim \text{Dir}_V(\eta)$
2. Para cada Documento,
   - Escoja un vector de proporciones de tópicos usando $\overline{\theta_d} \sim \text{Dir}_K(\overline{\alpha})$
   - Genere sucesivamente cada palabra así:
     - Escoja un tópico $Z_{d,n} \sim \text{Mult}(\overline{\theta_d}), \quad Z_{d,n} \in \{1, \ldots, K\}$
     - Escoja la palabra de modo que $W_{d,n} \sim \text{Mult}(\overline{\beta}_{Z_{d,n}})$

Representación Gráfica

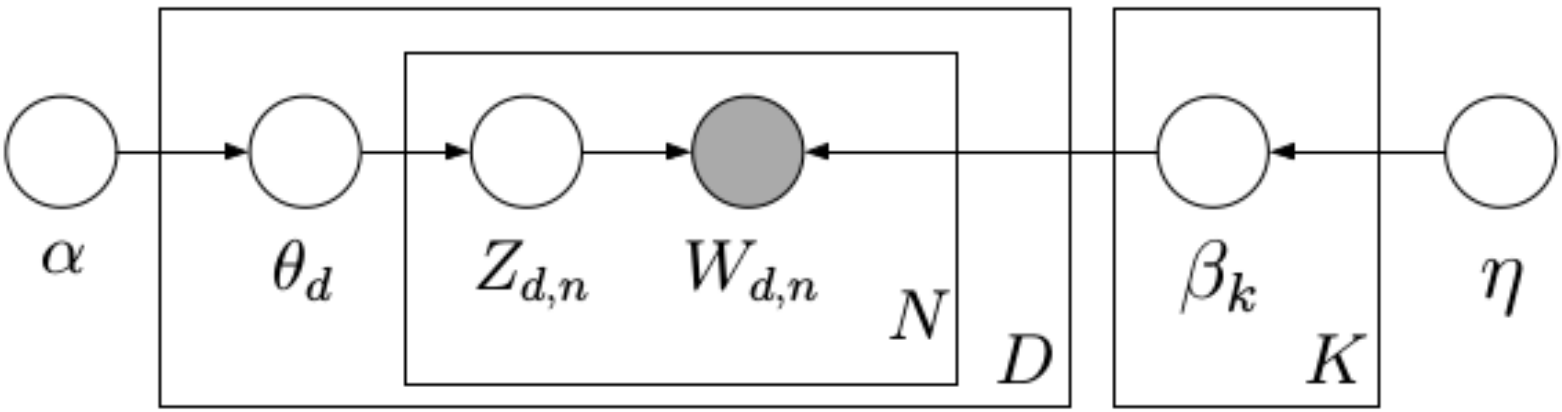

Figura: Los nodos coloreados son variables observadas, los vértices marcan dependencias entre variables aleatorias y las cajas denotan repetición. N,D y K representan respectivamente el número de palabras, de documentos y de tópicos.

Explicación sobre parámetros α y β

- La Dirichlet que genera β_k es simétrica.
- Es decir que a priori no queremos que los tópicos tengan un sesgo hacia algunas palabras.
- La Dirichlet que genera las mixturas de tópicos típicamente no lo es.
- Ahí se quiere capturar que en general los documentos pueden tener un tópico preponderante (y también querríamos modelar dispersión).

Incomputabilidad de la probabilidad posterior

$$p(\mathbf{w}|\alpha,\beta) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)} \int \left(\prod_{i=1}^{k}\theta_i^{\alpha_i-1}\right)\left(\prod_{n=1}^{N}\sum_{i=1}^{k}\prod_{j=1}^{V}(\theta_i\beta_{ij})^{w_n^j}\right) d\theta$$

Figura: Usualmente estas distribuciones son computables sacando logaritmos. Pero la no independencia de θ y β resulta en que haya una sumatoria donde están acoplados. Entonces la integral no se puede resolver (los métodos numéricos no funcionan bien en una dimensionalidad tan alta)

Visualizar un Tópico

- Para ver un tópico, podemos mostrar las probabilidades de sus términos β_i, o en cambio:

$$\text{term score}_{k,v} = \beta_{k,v} \log \left(\frac{\beta_{k,v}}{\left(\prod_{j=1}^{K} \beta_{j,v} \right)^{\frac{1}{K}}} \right)$$

Visualizar un Documento

- Usando la distribución posterior de la mixtura de tópicos podemos representar un documento sintéticamente como una proporción de tópicos.
- También es útil buscar documentos que tengan un contenido alto de cierto tópico.
- Si uno asigna cada palabra a su tópico más probable, también puede ver como se distribuyen los tópicos **dentro de un documento**.

Distribución dentro de un documento

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figura: Documento con modelo de tópicos, TREC AP Corpus

Encontrar Documentos Similares

- Podemos usar esa representación de baja dimensión para encontrar documentos similares.
- En principio queremos documentos con distribuciones similares, así que podemos usar la distancia de Hellinger para comparar distribuciones:

$$\text{document similarity}_{d,f} = \sum_{k=1}^{K} \left(\sqrt{\theta_{d,k}} - \sqrt{\theta_{f.k}} \right)^2 \qquad (4)$$