Text as Data - Modelado de Tópicos

Ciencia de Datos y Econometría Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Modelos de Tópicos

- El texto es de dimensionalidad muy alta
- Algunas veces un modelo factorial no supervisado es una estrategia popular y útil con datos de texto
- Puedes primero ajustar un modelo factorial a un corpus gigante y usar estos factores para aprendizaje supervisado en un subconjunto de documentos etiquetados.
- La reducción de dimensionalidad no supervisada facilita el aprendizaje supervisado

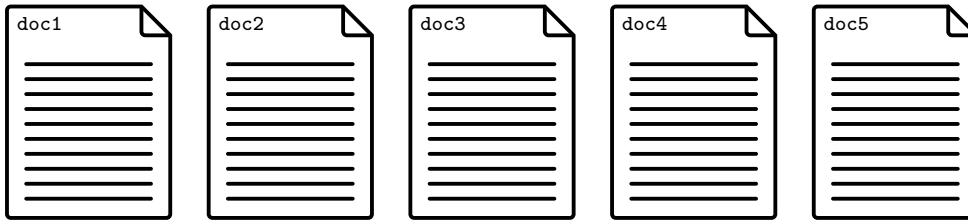
Modelos de Tópicos

LDA: Resumen

- El enfoque de usar PCA para factorizar texto era común antes de los años 2000.
- Las versiones de este algoritmo se denominaban bajo la etiqueta de análisis semántico latente.
- Sin embargo, esto cambió con la introducción del modelado de tópicos, también conocido como Asignación Latente de Dirichlet (LDA), por Blei et al. en 2003.
- Estos autores propusieron tomar en serio la representación de bolsa de palabras y modelar los conteos de tokens como realizaciones de un marco probabilístico.

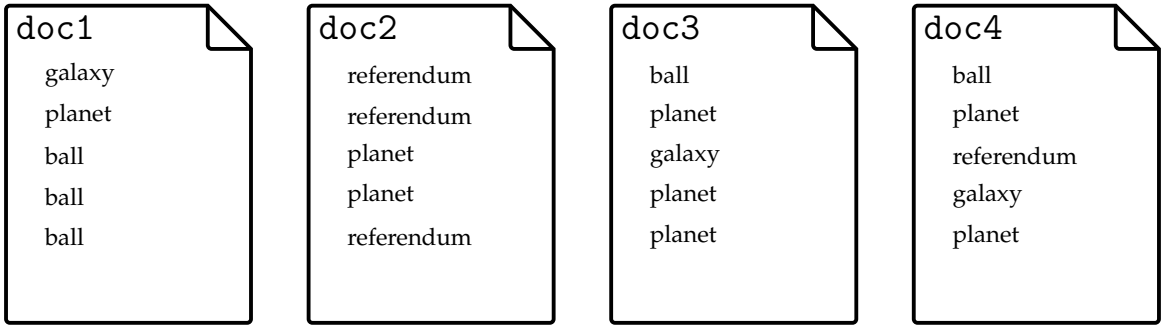
Topic Models

El Problema



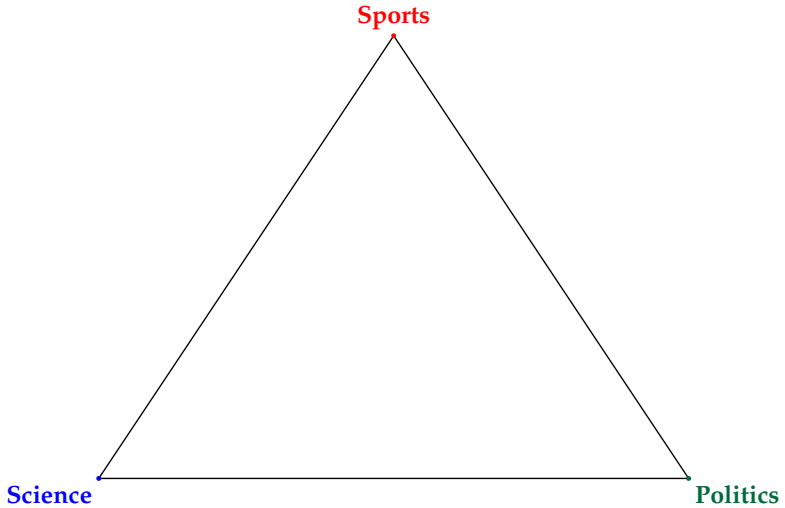
Topic Models

El Problema



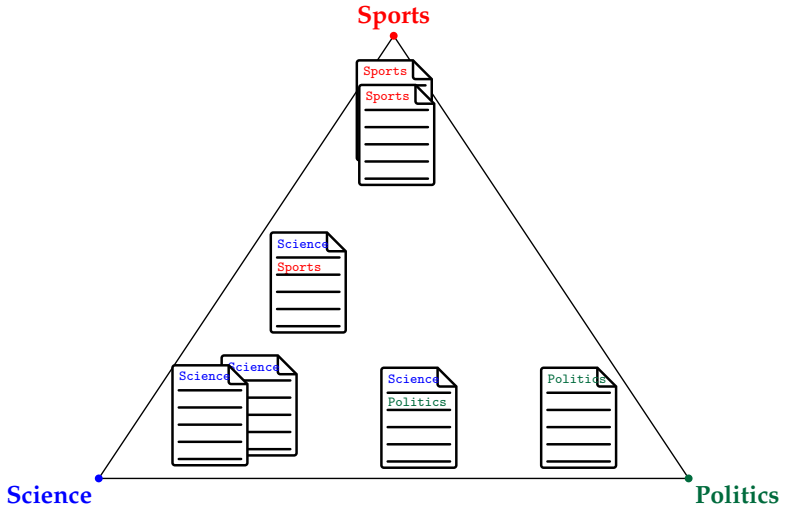
Topic Models

Latent Dirichlet Allocation



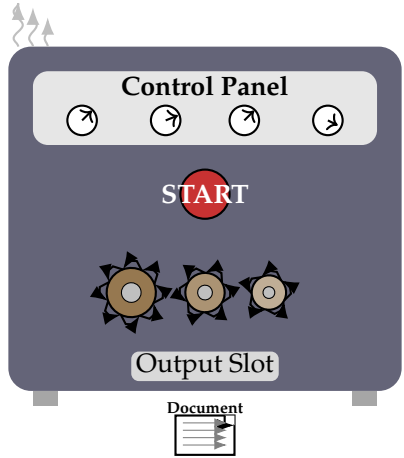
Topic Models

Latent Dirichlet Allocation



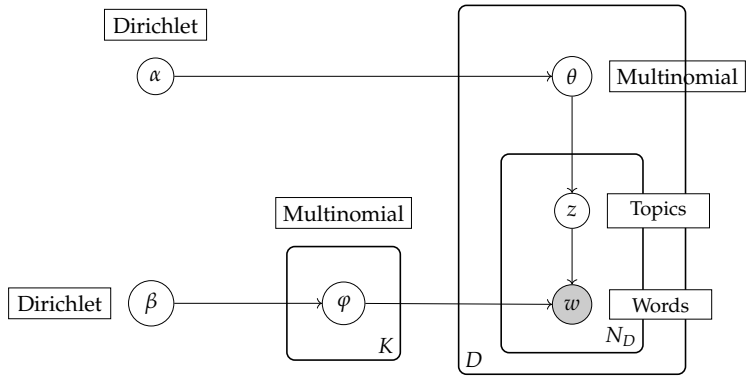
Topic Models

Latent Dirichlet Allocation



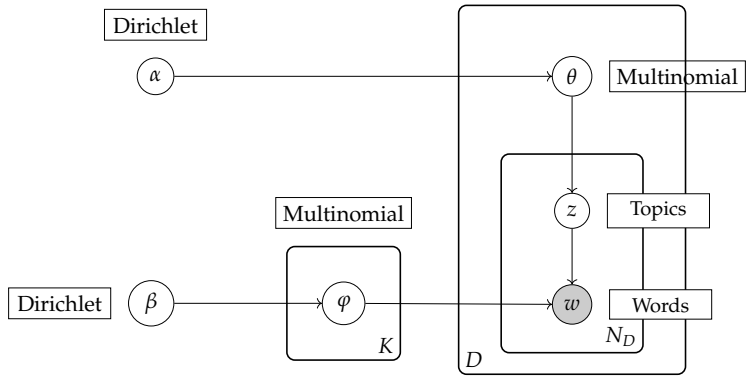
Topic Models

Latent Dirichlet Allocation: Blueprint



Topic Models

Latent Dirichlet Allocation: Blueprint



$$P(W, Z, \theta, \varphi; \alpha, \beta) = \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j) P(W_{j,t} \mid \varphi Z_{j,t})$$

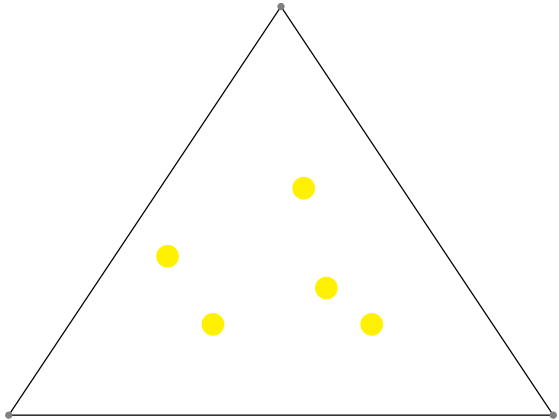
Topic Models

Latent Dirichlet Allocation: Blueprint

$$P(W, Z, \theta, \varphi; \alpha, \beta) = \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j) P(W_{j,t} \mid \varphi Z_{j,t})$$

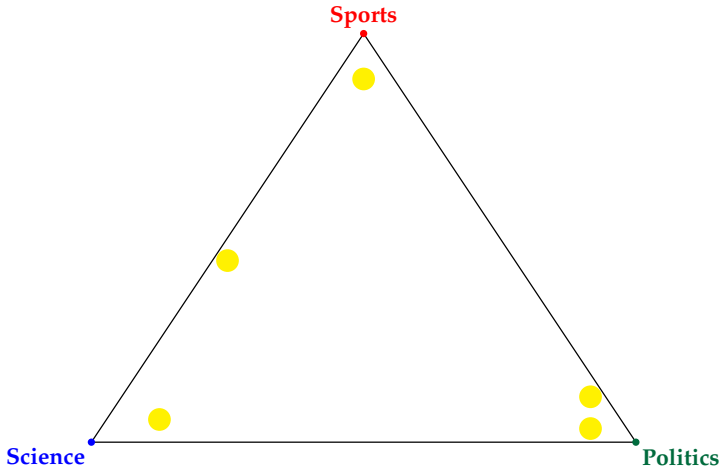
Topic Models

Dirichlet Distributions



Topic Models

Dirichlet Distributions



Topic Models

Dirichlet Distributions

- The Dirichlet distribution lives on something called the $(K-1)$**-simplex**. For our triangular house (when $K = 3$):

$$\Delta^{K-1} = \left\{ \boldsymbol{\theta} \in \mathbb{R}^K : \theta_k \geq 0 \; \forall k, \sum_{k=1}^{K} \theta_k = 1 \right\}$$

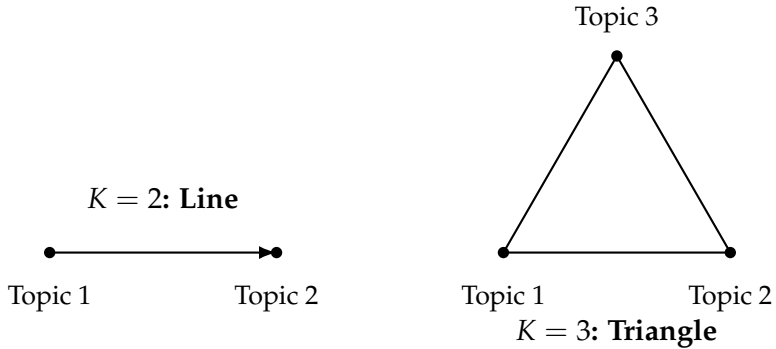
- This just says: $\boldsymbol{\theta}$ is a vector of K non-negative numbers that sum to 1
- We write $\boldsymbol{\theta} \sim \text{Dir}(\boldsymbol{\alpha})$ where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ is the concentration vector.
- **The probability density function is:**

$$f(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

- where $B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\alpha_0)}$ and $\alpha_0 = \sum_{k=1}^{K} \alpha_k$.
- $B(\boldsymbol{\alpha})$ is just a normalizing constant that ensures probabilities sum to 1.

Topic Models

Dirichlet Distributions



Topic Models

Dirichlet Distributions

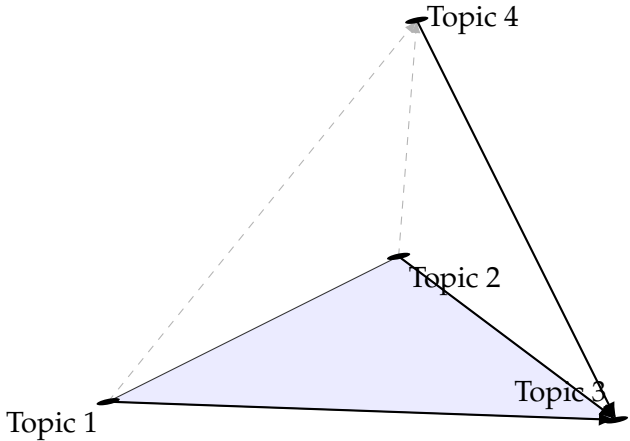


Figure 1: A tetrahedron for $K = 4$ topics. For $K > 4$, we enter higher dimensions!

Topic Models

Two Dirichlet Distributions in LDA

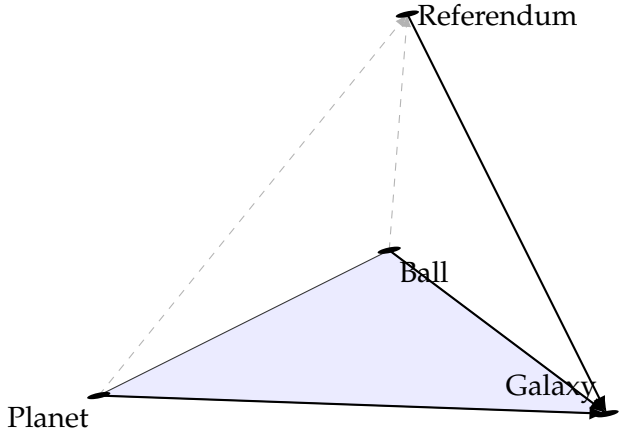
$$P(W, Z, \theta, \varphi; \alpha, \beta) = \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j) P(W_{j,t} \mid \varphi Z_{j,t})$$

LDA actually uses the Dirichlet distribution **twice**:

1. **Document-Topic distribution** ($\boldsymbol{\theta} \sim \text{Dir}(\boldsymbol{\alpha})$):
   - Lives on a triangle with corners = topics (Sports, Science, Politics)
   - Each point = a document's topic mixture
   - Usually sparse ($\alpha < 1$) because documents focus on one or two topics
2. **Topic-Word distribution** ($\boldsymbol{\varphi} \sim \text{Dir}(\boldsymbol{\beta})$):
   - Lives on a tetrahedron with corners = words (ball, planet, galaxy, referendum)
   - Each point = a topic's word distribution

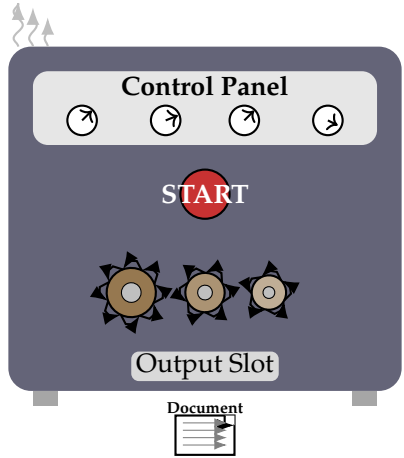
Topic Models

Two Dirichlet Distributions in LDA



Topic Models

LDA



Topic Models

LDA

$$P(W, Z, \theta, \varphi; \alpha, \beta) = \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j) P(W_{j,t} \mid \varphi Z_{j,t})$$

Topic Models: Examples LDA