MSC-011

Juan Zamora O.

Noviembre, 2023.

Estructura de la Presentación

Clustering y Modelos estadísticos de Texto

- Abundante en diversos dominios (redes sociales, medios digitales, registros en salud …)
- Resulta útil poder explorar estas coleciones de alguna manera asistida
- Clustering permite caracterizar de manera *automática* una colección de documentos
- A finales de los 90, aparecieron varios modelos estadístico de texto usando un modelo de mezcla sobre variables aleatorias multinomiales
  - LSI
  - pLSI

- LDA aparece a principio del 2000
- Incluye un modelo generativo para los documentos, además de los tópicos
- La idea es que documentos son representados como mezclas aleatorias de tópicos latentes
- Cada tópico se caracteriza como una una distribución sobre las palabras
- Distribución apriori de tópicos es una Dirichlet

Referencias: Blei et al. 2003

Idea fundamental

- Consideremos un conjunto de D documentos $\{W_1, W_2 \dots W_D\}$
- Cada documento W_d contiene N_d palabras, es decir $W_d = \{w_{d1}, w_{d2} \dots w_{dN_d}\}$
- El objetivo es agrupar los documentos y sus palabras en G grupos, denominados *tópicos*

El Modelo *generativo* LDA

- Para cada documento W_d de los D documentos, θ_d representa su proporción de tópicos y $\theta_d \sim Dir(\alpha)$
  - $\alpha = (\alpha_1, \alpha_2 \dots \alpha_G)$ es un apriori común sobre los tópicos para todos los documentos
- La distribución ψ_g, para el tópico $g \in \{1 \dots G\}$ también cumple con que $\psi_g \sim Dir(\beta)$
  - $\beta = (\beta_1, \beta_2 \dots \beta_V)$ y V es el tamaño del vocabulario
- Luego, el tópico z_{id} de la i-ésima palabra en el documento d sigue una distribución multinomial, es decir $z_{id} \sim M(1, \theta_d)$

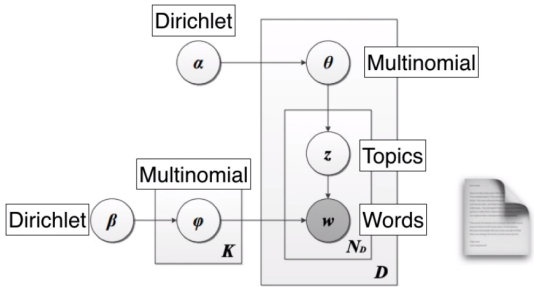
El Modelo *generativo* LDA

- Dado este tópico z_{id}, la i-ésima palabra en el documento d es también tomada a partir de una distribución multinomial, es decir $w_{id} \sim M(1, \psi_{z_{id}})$
- Notar que la distribución de palabras sobre los tópicos es una mezcla de multinomiales

$$w_{id}|\theta \sim \sum_{g=1}^{G} \theta_{dg} M(1, \psi_g)$$

Resumen gráfico de LDA

- La inferencia puede ser realizada usando el algoritmo *EM Bayesiano Variacional*, MCMC o Expectation-Propagation



Ejemplo de las palabras más representativas en 11 tópicos

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11
<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>
internacionalización	región	desarrollo	importante	prestigio	problemas	mejoras	tiempo	inclusiva	formación	académicos
nacional	valparaíso	medio	áreas	futuro	sociales	universidades	profesores	innovadora	profesional	comunidad
investigación	institución	vinculación	vanguardia	educación	desarrollo	aporte	profesional	estudiantes	calidad	estudiantes
región	comunidad	investigación	personas	estudiantes	temas	país	cambio	aprendizaje	nuevas_tecnologías	personas
liderazgo	entorno	estudiantes	estudios	programas	excelente	calidad	mejoras	abierta	procesos	funcionarios
reconocida	compromiso	sostenible	compleja	calidad	institución	desarrollo	mundo	personas	valores	conocimientos
referente	local	innovadora	espacios	carreras	resolver	conocimientos	tiempos	desarrollar	continua	oportunidad
proyectos	tradición	ambiente	nuevas_tecnologías	chile	público	infraestructura	puedan	investigación	institución	preocupada
áreas	ciudad	permitan	carreras	mejoras	país	enseñanza	gestión	más_inclusiva	trabajo	espacios
manteniendo	nacional	institución	territorio	mundo	comprometida	tres	investigación	calidad	estudiantes	servicio

Asociación texto y tópico

¿De qué sirve esta perspectiva generadora de documentos?

- Existen técnicas estadísticas y computacionales para invertir este procedimiento a partir de documentos existentes (…nuestros documentos), pudiendo así inferir la composición *más probable* de los tópicos que permitieron generar esta colección de documentos.
- Los tópicos estimados tienen un significado identificado por el/la analista
- Para encontrar la cantidad de tópicos se utiliza una medida denominada *Perplexity*
  - Se calcula tomando la log-verosimilitud de los documentos con los tópicos resultantes
  - Que tanto es posible reproducir la composición de los documentos dados los tópicos
  - El bjetivo es escoger el número de tópicos que minimiza la Perplexity