

EST-297

Fundamentos de Ciencia de Datos

Juan Zamora O.

Junio, 2024.



PONTIFICIA
UNIVERSIDAD
CATÓLICA DE
VALPARAÍSO



Estructura de la Presentación

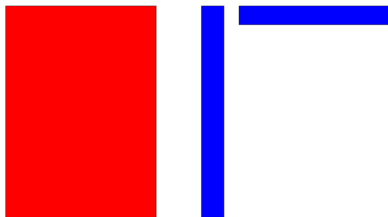
- 1 Aproximaciones Low-Rank para Clustering
- 2 (NMF) Factorización de matrices no-negativas
- 3 NMF
- 4 Aplicación de NMF para extracción de tópicos en texto

Aproximaciones Low-Rank para Clustering

Una matriz X de rango r admite una factorización de la forma

$$X = BC^T, B \in \mathbf{R}^{m \times r}, C \in \mathbf{R}^{n \times r}$$

X es aproximada con bajo rango (low-rank) cuando $\text{rango}(X) \ll \min(m, n)$



(NMF) Factorización de matrices no-negativas

- Grupo de algoritmos de de análisis multivariado y algebra lineal donde una matriz X es factorizada en dos matrices W y H
- Cada columna de X es aproximada por una combinación lineal no-negativa de las columnas de W , donde los coeficientes de mezcla corresponden a las columnas de H
- Las tres matrices tienen elementos no-negativos
- Usado en sistemas recomendadores, procesamiento de audio, agrupamiento de texto.

NMF

- Dada una matriz no-negativa $X \in \mathbf{R}^{m \times n}$ y un $k \in \mathbf{Z} \ll \min(m, n)$
- Encuentra matrices no-negativas $W \in \mathbf{R}^{m \times k}$ y $H \in \mathbf{R}^{k \times n}$ tales que minimizan

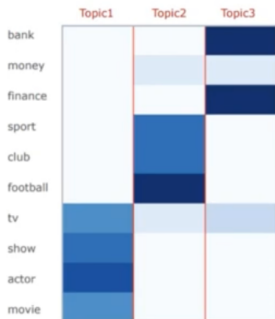
$$\|X - WH\|_F^2 = \sum_i \sum_j (X_{ij} - [WH]_{ij})^2$$

- W : base para un espacio k -dimensional, la i -ésima columna de H : corresponde a representación k -dim de i -ésima columna de X

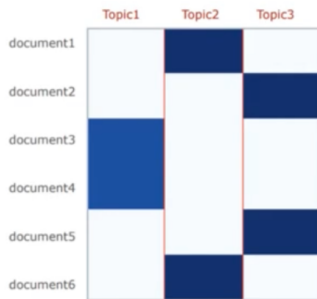
Aplicación de NMF para extracción de tópicos en texto

- Se construye matriz de terminos vs documentos
- Se aplica NMF para obtener W y H

Basis vectors W : topics (clusters)



Coefficients H : memberships for documents



Clustering y Modelos estadísticos de Texto

- Abundante en diversos dominios (redes sociales, medios digitales, registros en salud ...)
- Resulta útil poder explorar estas colecciones 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 $z_{id} \sim M(1, \theta_d)$

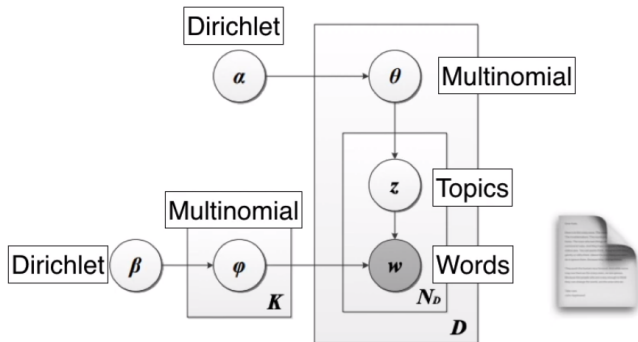
El Modelo *generativo* LDA

- Dado este t3pico z_{id} , la i -3sima palabra en el documento d es tambi3n tomada a partir de una distribuci3n multinomial, es decir $w_{id} \sim M(1, \psi_{z_{id}})$
- Notar que la distribuci3n de palabras sobre los t3picos es una mezcla de multinomiales

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

Resumen gráfico de LDA

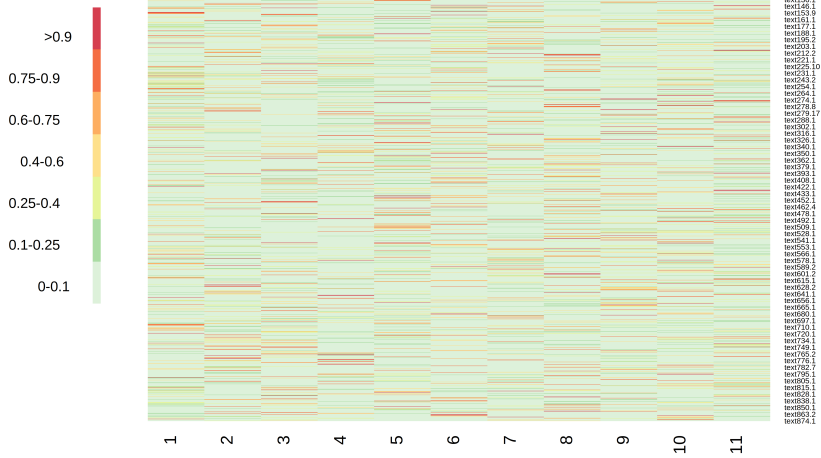
- La inferencia puede ser realizada usando el algoritmo *EM Bayesiano Variacional*, MCMC o Expectation-Propagation



Ejemplo de las palabras más representativas en 11 tópicos

| Topic 1 <chr> | Topic 2 <chr> | Topic 3 <chr> | Topic 4 <chr> | Topic 5 <chr> | Topic 6 <chr> | Topic 7 <chr> | Topic 8 <chr> | Topic 9 <chr> | Topic 10 <chr> | Topic 11 <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 objetivo es escoger el número de tópicos que minimiza la Perplexity