Filtrado Basado en Contenido

IIC 3633 - Sistemas Recomendadores

Denis Parra
Profesor Asistente, DCC, PUC CHile

TOC

En esta clase

1. Contenido en lugar de ratings
2. Representación de Espacio Vectorial
3. TF-IDF
4. Buscando Items Similares
5. Representación en Espacio Latente

Por Qué un Recomendador Basado en Contenido

- El filtrado colaborativo tiene algunas desventajas: cold-start, sparcity, transparency.

PROS

- A diferencia del Filtrado Colaborativo, si los items tienes descripciones suficientes, nos evitamos el "new-item problem""
- Las representaciones del contenido son variadas y permiten utilizar diversas técnicas de procesamiento del texto, uso de información semántica, inferencias, etc.
- Es sencillo hacer un sistema más transparente: usamos el mismo contenido para explicar las recomendaciones.

CONS

- Tienden a la sobre-especialización: va a recomendar items similares a los ya consumidos, creando una tendencia al "filter bubble".
- Los métodos basados en filtrado colaborativo han mostrado ser, empíricamente, más precisos al momento de generar recomendaciones.

Arquitectura de un Sistema de Recomendación CB

- Los componentes principales son: (1) Analizador del Contenido, (2) Aprendizaje del Perfil de Usuario, (3) Filtrado de Contenido

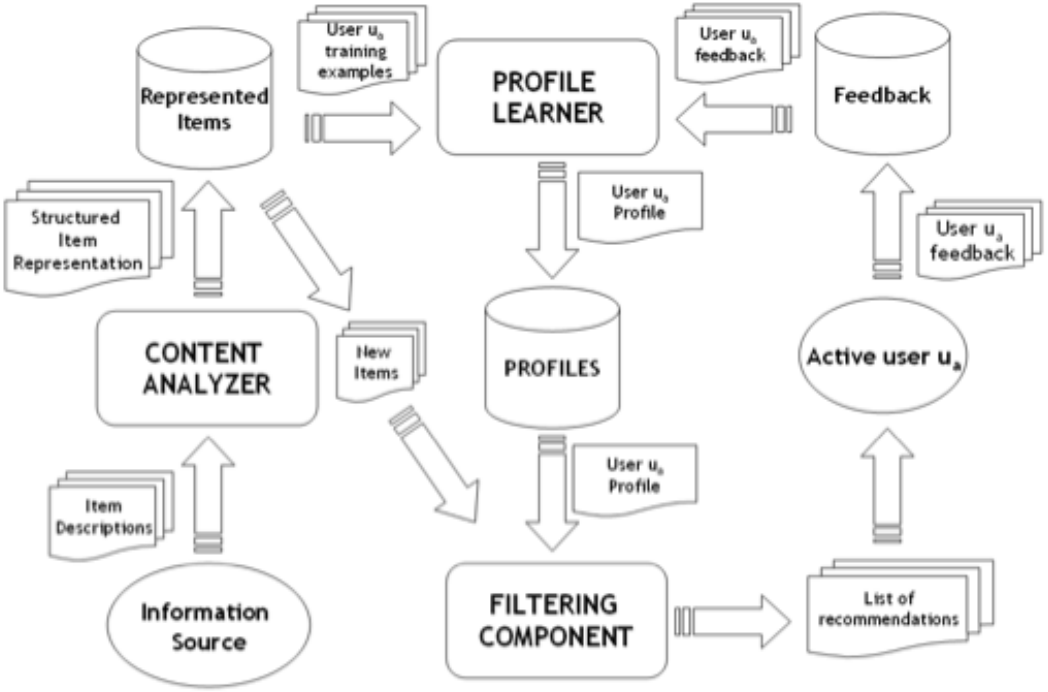
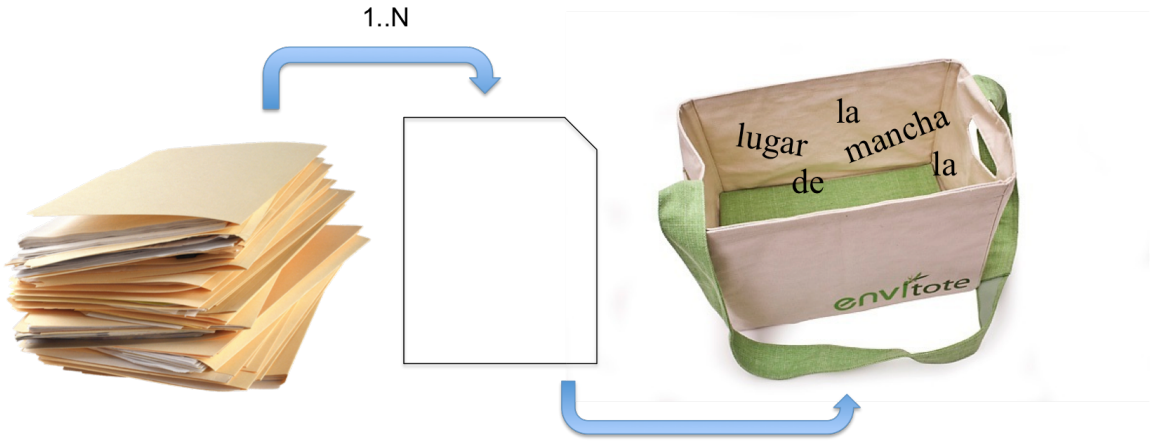


Fig. 3.1: High level architecture of a Content-based Recommender

Representación del Contenido: Bolsa de Palabras

- Se suele representar a los documentos como "bolsas de palabras"; de esta forma es fácil pasar a representar cada documento como un vector (Vector Space Model)



Representación del Contenido: VSM

- El corpus completo puede entonces representarse como una matriz donde las filas son términos y las columnas son documentos.

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth . . .
ANTHONY	1	1	0	0	0	1
BRUTUS	1	1	0	1	0	0
CAESAR	1	1	0	1	1	1
CALPURNIA	0	1	0	0	0	0
CLEOPATRA	1	0	0	0	0	0
MERCY	1	0	1	1	1	1
WORSER	1	0	1	1	1	0
. . .						

- Luego, ¿Cúal es la mejor forma de representar los pesos de los términos?

Representación del Contenido: VSM II

Frecuencia de los términos

Cada documento se representa como un vector, el "peso" de cada palabra para ese documento

$$\mathrm{TF}(t_k, d_j) = \frac{f_{k,j}}{max_z f_{z,j}}$$

puede darse en base a la frecuencia del término en el documento.

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth . . .
ANTHONY	157	73	0	0	0	1
BRUTUS	4	157	0	2	0	0
CAESAR	232	227	0	2	1	0
CALPURNIA	0	10	0	0	0	0
CLEOPATRA	57	0	0	0	0	0
MERCY	2	0	3	8	5	8
WORSER	2	0	1	1	1	5
. . .						

Podemos normalizar el valor en función de la frecuencia máxima de cualquier término en el documento.

Representación del Contenido: VSM III

Log de Frecuencia de los términos

Pero el hecho que un término x aparece 100 veces y otro término y sólo 10 veces, no hace a x 10 veces más relevantes; por lo tanto podemos usar un logaritmo.

- La log-frecuencia del término t en d se define como

$$w_{t,d} = \begin{cases} 1 + \log_{10} \mathrm{tf}_{t,d} & \text{if } \mathrm{tf}_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$$

- $\mathrm{tf}_{t,d} \rightarrow w_{t,d}$:

  $0 \rightarrow 0,\ 1 \rightarrow 1,\ 2 \rightarrow 1.3,\ 10 \rightarrow 2,\ 1000 \rightarrow 4,$ etc.

Representación del Contenido: VSM IV

TF-IDF

Bajo la intuición de que un término que aparece en sólo unos poco documentos podría ser descriptivo, podemos considerar la "Inverse Document Frequency" y combinarla con la "Term Frequency":

$$\text{TF-IDF}(t_k, d_j) = \underbrace{\text{TF}(t_k, d_j)}_{\text{TF}} \cdot \underbrace{log\frac{N}{n_k}}_{\text{IDF}}$$

Done t_k es el término k, d_j es el documento j.

Resumen de Componentes del TF-IDF

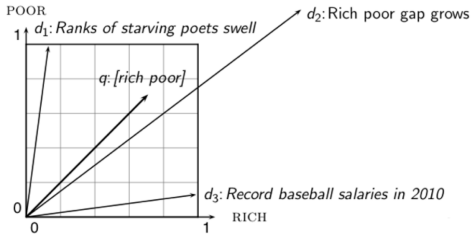
Term frequency		Document frequency		Normalization	
n (natural)	$\mathrm{tf}_{t,d}$	n (no)	1	n (none)	1
l (logarithm)	$1 + \log(\mathrm{tf}_{t,d})$	t (idf)	$\log \frac{N}{\mathrm{df}_t}$	c (cosine)	$\frac{1}{\sqrt{w_1^2+w_2^2+\ldots+w_M^2}}$
a (augmented)	$0.5 + \frac{0.5 \times \mathrm{tf}_{t,d}}{\max_t(\mathrm{tf}_{t,d})}$	p (prob idf)	$\max\{0, \log \frac{N-\mathrm{df}_t}{\mathrm{df}_t}\}$	u (pivoted unique)	$1/u$
b (boolean)	$\begin{cases} 1 & \text{if } \mathrm{tf}_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$			b (byte size)	$1/CharLength^{\alpha}$, $\alpha < 1$
L (log ave)	$\frac{1+\log(\mathrm{tf}_{t,d})}{1+\log(\mathrm{ave}_{t \in d}(\mathrm{tf}_{t,d}))}$				

Representación Semántica del Contenido

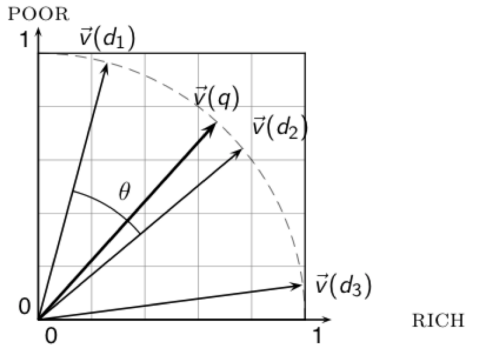
- No todo el contenido del documento corresponde a la misma categoría.
- Autor, palabras clave, fechas, tópicos pueden dar una noción adicional de filtrado.
- Opción 1: Representación semántica explícita (No lo veremos en detalle en esta clase)
  - Ontologías
  - WordNet
  - ConceptNet
- Opción 2: Inferir representación semántica (LSI, LDA)
- Opción 3: Word Vectors (Word2Vec, Glove)

Buscando Items Similares

Distancia Euclidiana

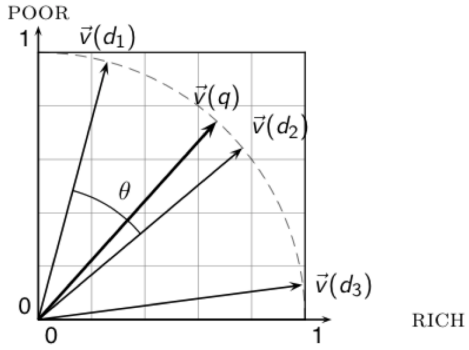


Distancia Coseno



Buscando Items Similares

Distancia Coseno



Fórmula

$$sim(d_i, d_j) = \frac{\sum_k w_{ki} \cdot w_{kj}}{\sqrt{\sum_k {w_{ki}}^2} \cdot \sqrt{\sum_k {w_{kj}}^2}}$$

Buscando Items Similares II

Okapi BM25

$$RSV_d = \sum_{t \in q} IDF \cdot \frac{(k_1 + 1) tf_{td}}{k_1((1-b) + b \times (L_d / L_{ave})) + tf_{td}} \cdot \frac{(k_3 + 1) tf_{tq}}{k_3 + tf_{tq}}$$

Ref: Denis Parra and Peter Brusilovsky. 2009. Collaborative filtering for social tagging systems: an experiment with CiteULike. In Proceedings of the third ACM conference on Recommender systems (RecSys '09) http://doi.acm.org/10.1145/1639714.1639757

Buscando Items Similares III

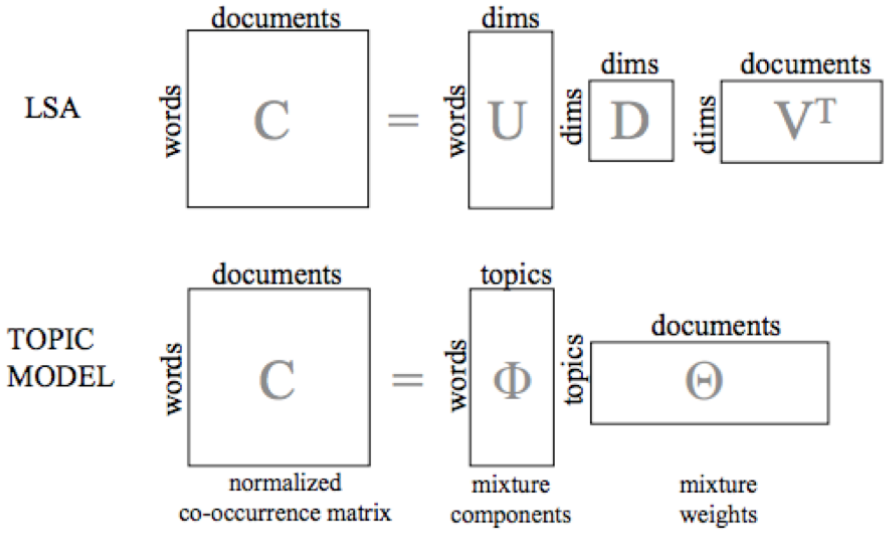
Técnicas de Procesamiento adicionales

- Pasar a mayúsculas/minúsculas
- Tokenization
- Stemming (Porter, Krovetz)
- Lemmatization

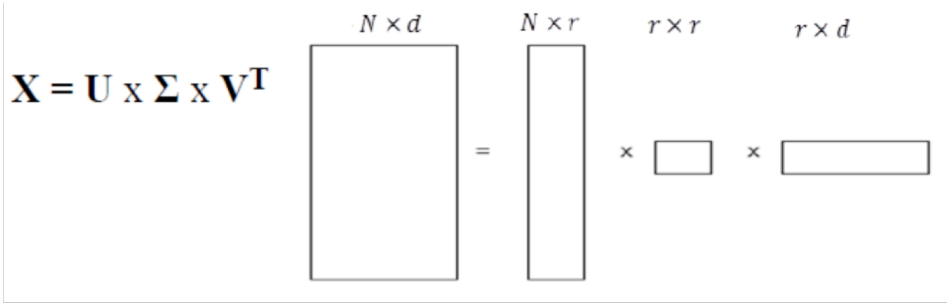
Buscando Items Similares

Representación en espacio latente

- Latent Semantic Indexing
- Latent Dirichlet Allocation



LSI I



$$(\mathbf{t}_i^T) \rightarrow \begin{bmatrix} x_{1,1} & \dots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \dots & x_{m,n} \end{bmatrix} = (\hat{\mathbf{t}}_i^T) \rightarrow \begin{bmatrix} \begin{bmatrix} \\ \mathbf{u}_1 \\ \\ \end{bmatrix} \dots \begin{bmatrix} \\ \mathbf{u}_l \\ \\ \end{bmatrix} \end{bmatrix} \cdot \begin{bmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_l \end{bmatrix} \cdot \begin{bmatrix} [& \mathbf{v}_1 &] \\ & \vdots & \\ [& \mathbf{v}_l &] \end{bmatrix}$$

X ($\mathbf{d}_j \downarrow$) U Σ V^T ($\hat{\mathbf{d}}_j \downarrow$)

LSI II

		d_1	d_2	d_3	d_4	d_5	d_6	
	ship	1	0	1	0	0	0	
	boat	0	1	0	0	0	0	
	ocean	1	1	0	0	0	0	
	voyage	1	0	0	1	1	0	
	trip	0	0	0	1	0	1	

LSI III

	1	2	3	4	5
ship	−0.44	−0.30	0.57	0.58	0.25
boat	−0.13	−0.33	−0.59	0.00	0.73
ocean	−0.48	−0.51	−0.37	0.00	−0.61
voyage	−0.70	0.35	0.15	−0.58	0.16
trip	−0.26	0.65	−0.41	0.58	−0.09

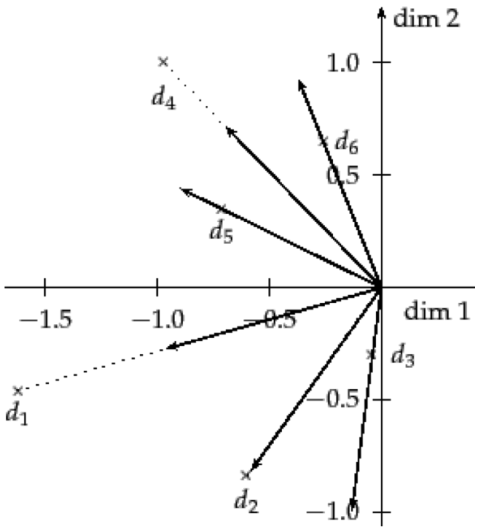
$$\begin{bmatrix} 2.16 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 1.59 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 1.28 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.39 \end{bmatrix}$$

	d_1	d_2	d_3	d_4	d_5	d_6
1	−0.75	−0.28	−0.20	−0.45	−0.33	−0.12
2	−0.29	−0.53	−0.19	0.63	0.22	0.41
3	0.28	−0.75	0.45	−0.20	0.12	−0.33
4	0.00	0.00	0.58	0.00	−0.58	0.58
5	−0.53	0.29	0.63	0.19	0.41	−0.22

19/28

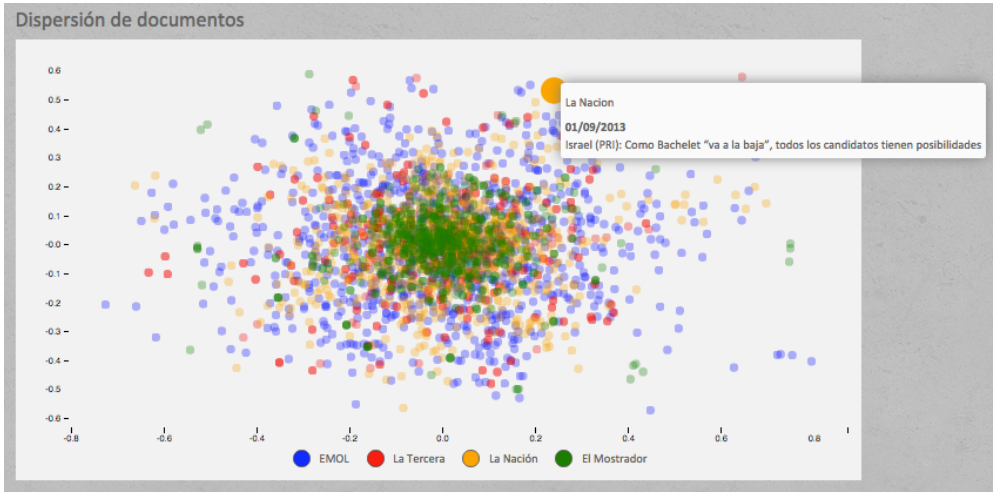
LSI IV

		d_1	d_2	d_3	d_4	d_5	d_6
	1	−0.75	−0.28	−0.20	−0.45	−0.33	−0.12
	2	−0.29	−0.53	−0.19	0.63	0.22	0.41
	3	0.28	−0.75	0.45	−0.20	0.12	−0.33
	4	0.00	0.00	0.58	0.00	−0.58	0.58
	5	−0.53	0.29	0.63	0.19	0.41	−0.22



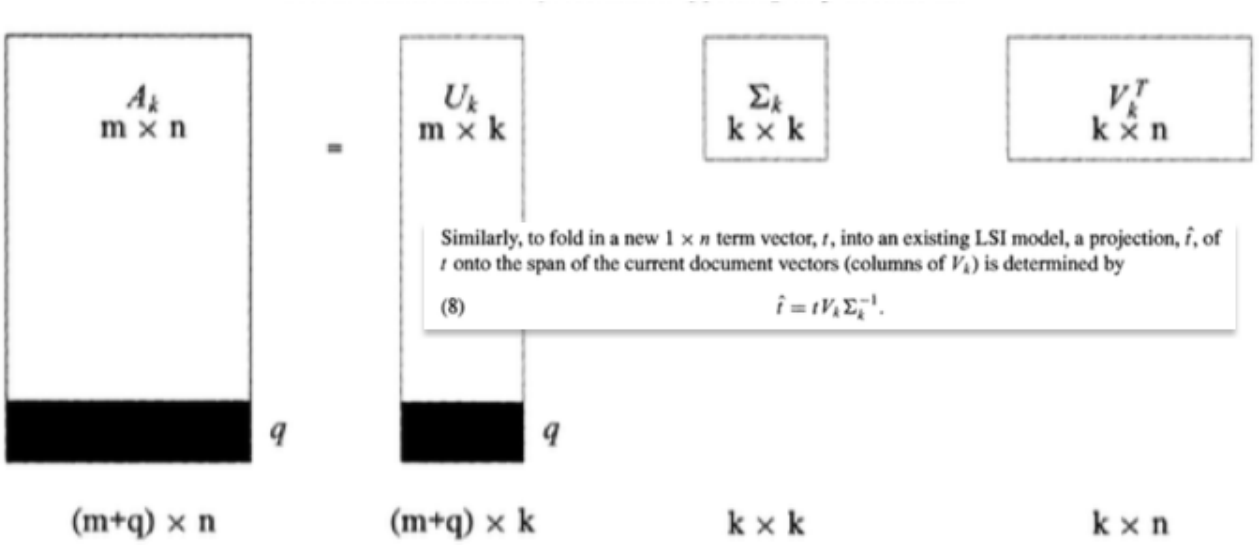
LSI IV

Demo: http://dfao-uc.github.io/



Proyección de documentos o términos nuevos

- Folding in: Using Linear Algebra for Intelligent Information Retrieval



Similarly, to fold in a new $1 \times n$ term vector, t, into an existing LSI model, a projection, $\hat{t}$, of t onto the span of the current document vectors (columns of V_k) is determined by

(8) $$\hat{t} = tV_k\Sigma_k^{-1}.$$

FIG. 3. *Mathematical representation of folding-in q terms.*

NMF

- Non-Negative Matrix Factorization

Document Clustering Based On Non-negative Matrix Factorization

Wei Xu, Xin Liu, Yihong Gong

NEC Laboratories America, Inc.
10080 North Wolfe Road, SW3-350
Cupertino, CA 95014, U.S.A.

{xw,xliu,ygong}@ccrl.sj.nec.com

LDA I

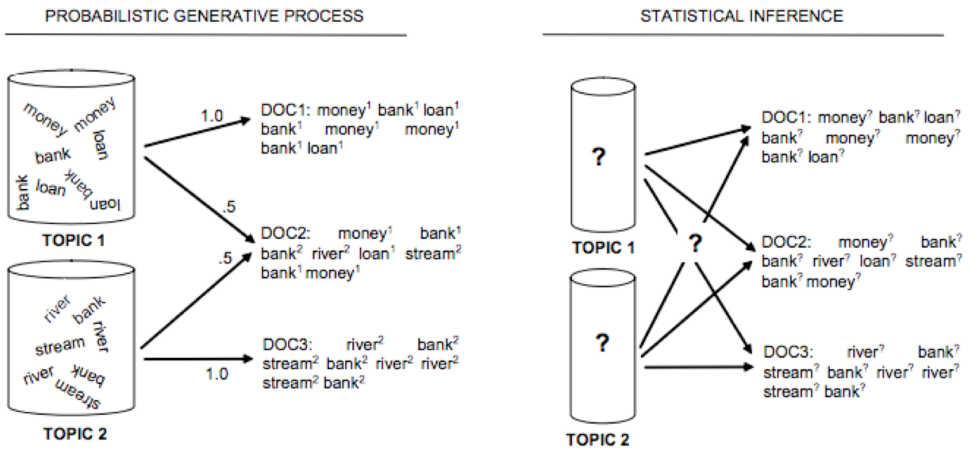
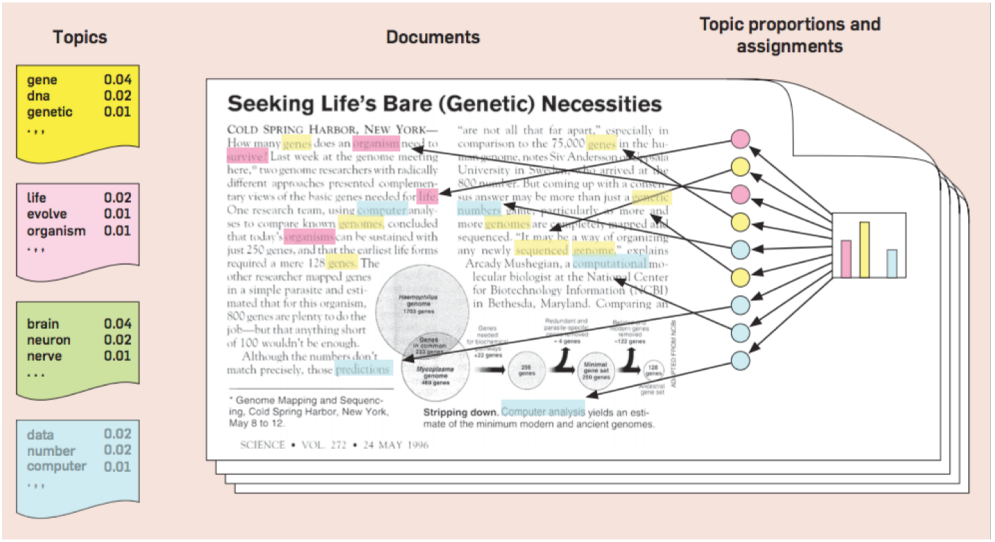


Figure 2. Illustration of the generative process and the problem of statistical inference underlying topic models

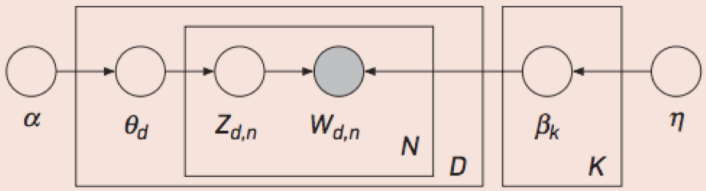
LDA II


Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
life 0.02
evolve 0.01
organism 0.01
brain 0.04
neuron 0.02
nerve 0.01
data 0.02
number 0.02
computer 0.01
Seeking Life's Bare (Genetic) Necessities
COLD SPRING HARBOR, NEW YORK—
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996

25/28

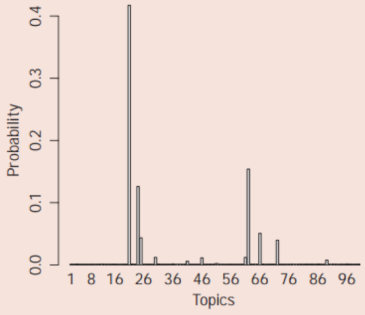
LDA III

Figure 4. The graphical model for latent Dirichlet allocation. Each node is a random variable and is labeled according to its role in the generative process (see Figure 1). The hidden nodes—the topic proportions, assignments, and topics—are unshaded. The observed nodes—the words of the documents—are shaded. The rectangles are "plate" notation, which denotes replication. The *N* plate denotes the collection words within documents; the *D* plate denotes the collection of documents within the collection.



26/28

LDA IV



"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

27/28

Referencias

- Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval (Vol. 1, p. 6). Cambridge: Cambridge university press.
- Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. Handbook of latent semantic analysis, 427(7), 424-440.
- Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.