Topic Models

Tecnologías de búsqueda en la web

Marcelo Mendoza

Departamento de Informática
Universidad Técnica Federico Santa María

Topic Models: PLSI/ LDA

Topic Models

Topic models (idea):

Topic 247

word	prob.
DRUGS	.069
DRUG	.060
MEDICINE	.027
EFFECTS	.026
BODY	.023
MEDICINES	.019
PAIN	.016
PERSON	.016
MARIJUANA	.014
LABEL	.012
ALCOHOL	.012
DANGEROUS	.011
ABUSE	.009
EFFECT	.009
KNOWN	.008
PILLS	.008

Topic 5

word	prob.
RED	.202
BLUE	.099
GREEN	.096
YELLOW	.073
WHITE	.048
COLOR	.048
BRIGHT	.030
COLORS	.029
ORANGE	.027
BROWN	.027
PINK	.017
LOOK	.017
BLACK	.016
PURPLE	.015
CROSS	.011
COLORED	.009

Topic 43

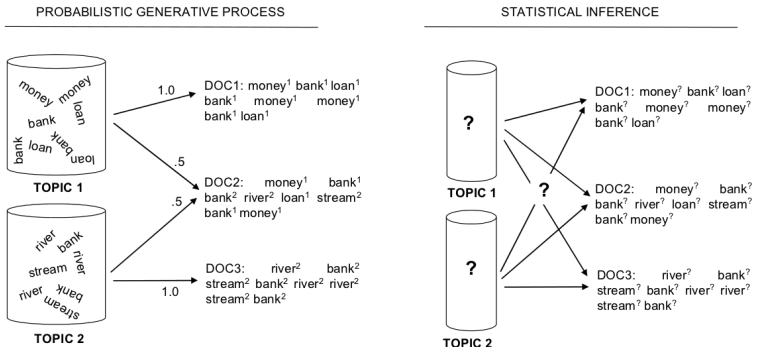
word	prob.
MIND	.081
THOUGHT	.066
REMEMBER	.064
MEMORY	.037
THINKING	.030
PROFESSOR	.028
FELT	.025
REMEMBERED	.022
THOUGHTS	.020
FORGOTTEN	.020
MOMENT	.020
THINK	.019
THING	.016
WONDER	.014
FORGET	.012
RECALL	.012

Topic 56

word	prob.
DOCTOR	.074
DR.	.063
PATIENT	.061
HOSPITAL	.049
CARE	.046
MEDICAL	.042
NURSE	.031
PATIENTS	.029
DOCTORS	.028
HEALTH	.025
MEDICINE	.017
NURSING	.017
DENTAL	.015
NURSES	.013
PHYSICIAN	.012
HOSPITALS	.011

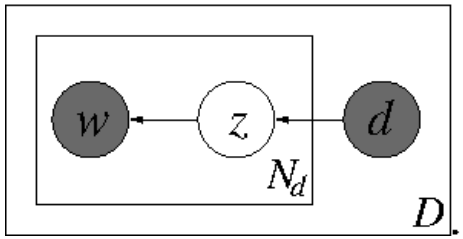
Topic Models

Proceso probabilístico generativo (inferencia):



PLSA (Hofmann 1999[1])

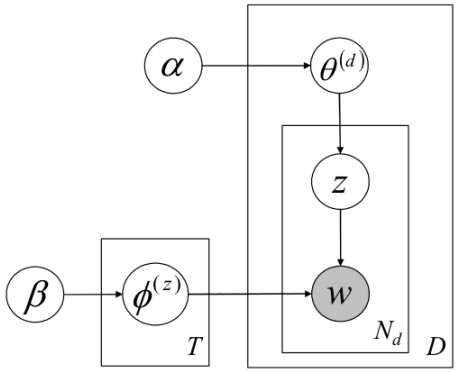
Modelo gráfico probabilístico:



[1]Hofmann, T. (1999). Probabilistic Latent Semantic Indexing. In Proceedings of the 22nd International ACM SIGIR Conference, 50-57.

LDA (Blei et al 2003[2])

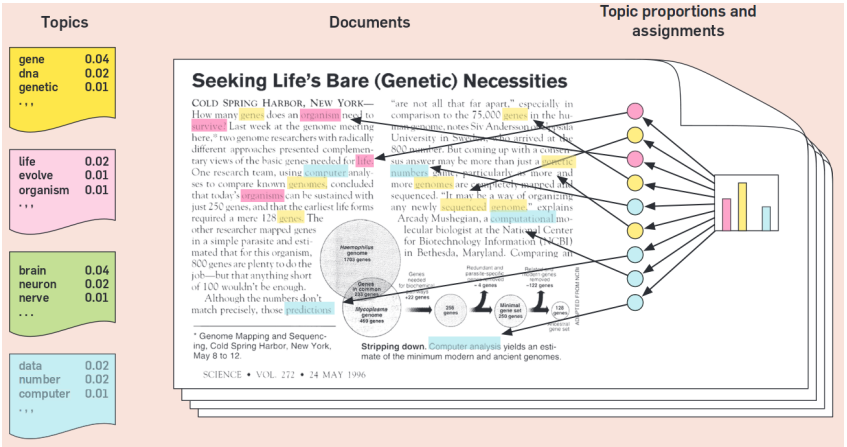
Modelo gráfico probabilístico:



[2]D. Blei, A. Ng, and M. Jordan (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993–1022.

Modelo generativo LDA

El modelo generativo:[3]



[3]Más en: Blei. D. (2012). Probabilistic Topic Models, Communications of the ACM, 55(4):77-84.

El efecto de α

Bias/smoothing: