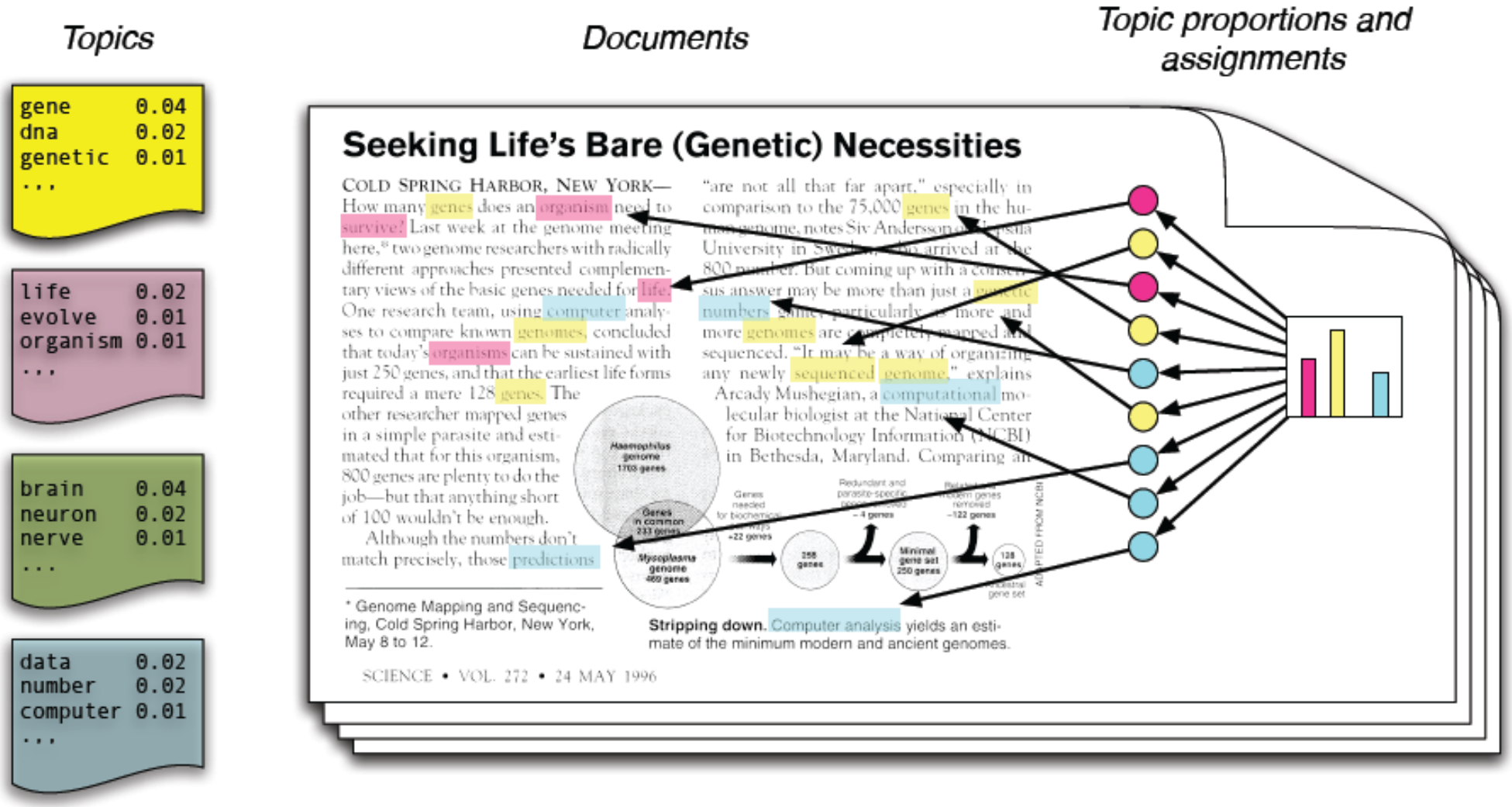
Latent Dirichlet Allocation

Patrice Béchard Théophile Gervet

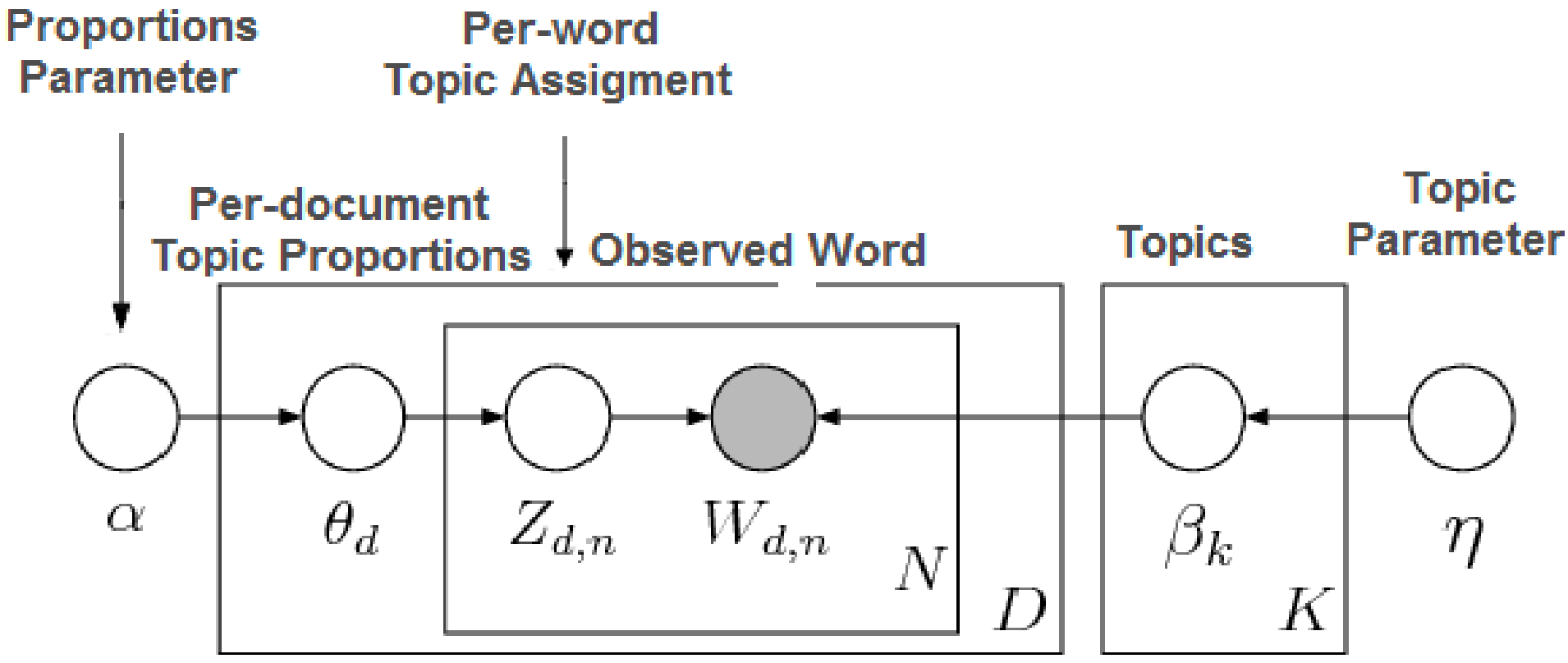
Université de Montréal

December 16, 2017

Context


Topics
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
...
Documents
Topic proportions and assignments
Seeking Life's Bare (Genetic) Necessities
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.
Although the numbers don't match precisely, those predictions
"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996

Graphical Model



$$\prod_{k=1}^{K} p(\beta_k|\eta) \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N_d} p(z_{d,n}|\theta_d) p(w_{d,n}|z_{d,n}, \beta_{1:K})$$

Generative Process

Formally, LDA assumes the following generative process for each document $\mathbf{w}_d$:

1. Sample a distribution over topics $\boldsymbol{\theta}_d \sim \mathrm{Dir}(\boldsymbol{\alpha})$
2. For each of the N_d words $w_{d,n}$ independently:
   - 2.1 Draw a topic from the distribution over topics : $z_{d,n} \sim \mathrm{Mult}(\boldsymbol{\theta}_d)$
   - 2.2 Draw a word from this topic : $w_{d,n} \sim \mathrm{Mult}(\boldsymbol{\beta}_{z,n})$

Intractable posterior mean :

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\eta}) = \frac{p(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\eta})}{p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\eta})}$$

Variational EM

Intractable posterior : $$p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})}$$

Log-likelihood : $$\log p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \log \int \sum_{\mathbf{z}} p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) d\boldsymbol{\theta}$$

Usual trick : $$\begin{aligned}\log p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \log \int \sum_{\mathbf{z}} \frac{p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) q(\boldsymbol{\theta}, \mathbf{z})}{q(\boldsymbol{\theta}, \mathbf{z})} d\boldsymbol{\theta} \\ &= \mathbb{E}_q \left[\log \int \sum_{\mathbf{z}} \frac{p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})}{q(\boldsymbol{\theta}, \mathbf{z})} d\boldsymbol{\theta} \right] \\ &\geq \mathbb{E}_q \left[\log p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) \right] - \mathbb{E}_q \left[q(\boldsymbol{\theta}, \mathbf{z}) \right] \\ &= \mathcal{L}(q, \boldsymbol{\alpha}, \boldsymbol{\beta})\end{aligned}$$

Variational EM

E-step : $\max_q \mathcal{L} \Leftrightarrow \min_q \mathrm{KL}\left(q(\boldsymbol{\theta}, \mathbf{z}) || p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})\right)$

Variational distribution : $q(\boldsymbol{\theta}, \mathbf{z}|\gamma, \boldsymbol{\phi}) = q(\boldsymbol{\theta}|\gamma) \prod_{n=1}^{N} q(z_n|\boldsymbol{\phi}_n)$

New E-step : $(\gamma^*, \boldsymbol{\phi}^*) = \arg\min_{\gamma, \boldsymbol{\phi}} \mathrm{KL}(q(\boldsymbol{\theta}, \mathbf{z}|\gamma, \boldsymbol{\phi}) \;||\; p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}))$

- Overall algorithm :

E-step : $\boldsymbol{\phi}_{n,k} \propto \boldsymbol{\beta}_{k,w_n} \exp\{\Psi(\gamma_k)\}$

$$\gamma_k = \boldsymbol{\alpha}_k + \sum_{n=1}^{N} \boldsymbol{\phi}_{n,k}$$

M-step : $\boldsymbol{\beta}_{k,j} \propto \sum_{d=1}^{D} \sum_{n=1}^{N_d} \boldsymbol{\phi}_{d,n,k} w_{d,n}^{(j)}$

Gibbs Sampling

Collapsed (or *Rao-Blackwellized*) Gibbs sampling :

- $\boldsymbol{\theta}_d$ and $\boldsymbol{\beta}_k$ can be computed using only $z_{d,n}$.
- We sample $z_{d,n}$ from posterior $p(z_n|\mathbf{z}_{\neg n}, \boldsymbol{\alpha}, \boldsymbol{\eta}, \mathbf{w})$.

Developed posterior in terms of counts :

$$p(z_n = k|\mathbf{z}_{\neg n}, \mathbf{w}) \propto \frac{n_{\neg n,k}^{(w_n)} + \boldsymbol{\eta}_n}{\sum_{n'=1}^{V} \left(n_{\neg n,k}^{(w_{n'})} + \boldsymbol{\eta}_{n'} \right)} \frac{n_{\neg n,k}^{(\mathbf{w}_n)} + \boldsymbol{\alpha}_k}{\sum_{k'=1}^{(K)} \left(n_{\neg n,k'}^{(\mathbf{w}_n)} + \boldsymbol{\alpha}_{k'} \right)}$$

Computation of parameters :

$$\boldsymbol{\beta}_{k,n} = \frac{n_k^{(w_n)} + \boldsymbol{\eta}_t}{\sum_{t=1}^{V} n_k^{(t)} + \boldsymbol{\eta}_t} \qquad \boldsymbol{\theta}_{d,k} = \frac{n_m^{(k)} + \boldsymbol{\alpha}_k}{\sum_{k=1}^{K} n_m^{(k)} + \boldsymbol{\alpha}_k}$$

References

Blei, David M, Ng, Andrew Y, and Jordan, Michael I. *Latent dirichlet allocation*. The Journal of Machine Learning Research, 3:993-1022, 2003.

Heinrich, Gregor. *Parameter estimation for text analysis*. Technical report, vsonix GmbH + University of Leipzig, Leipzig, Germany, 2005.

Darling, William M. *A theoretical and practical implementation tutorial on topic modeling and gibbs sampling*. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies, 2011.