Applied Text Mining in Python

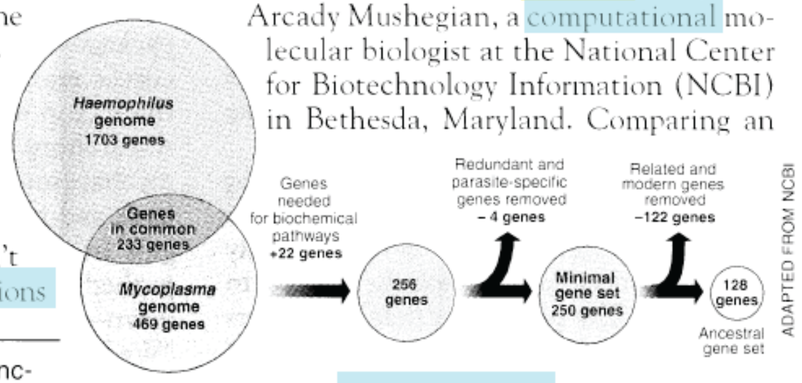
Topic modeling

Documents Exhibit Multiple Topics

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

(Figure courtesy Prof. David Blei)

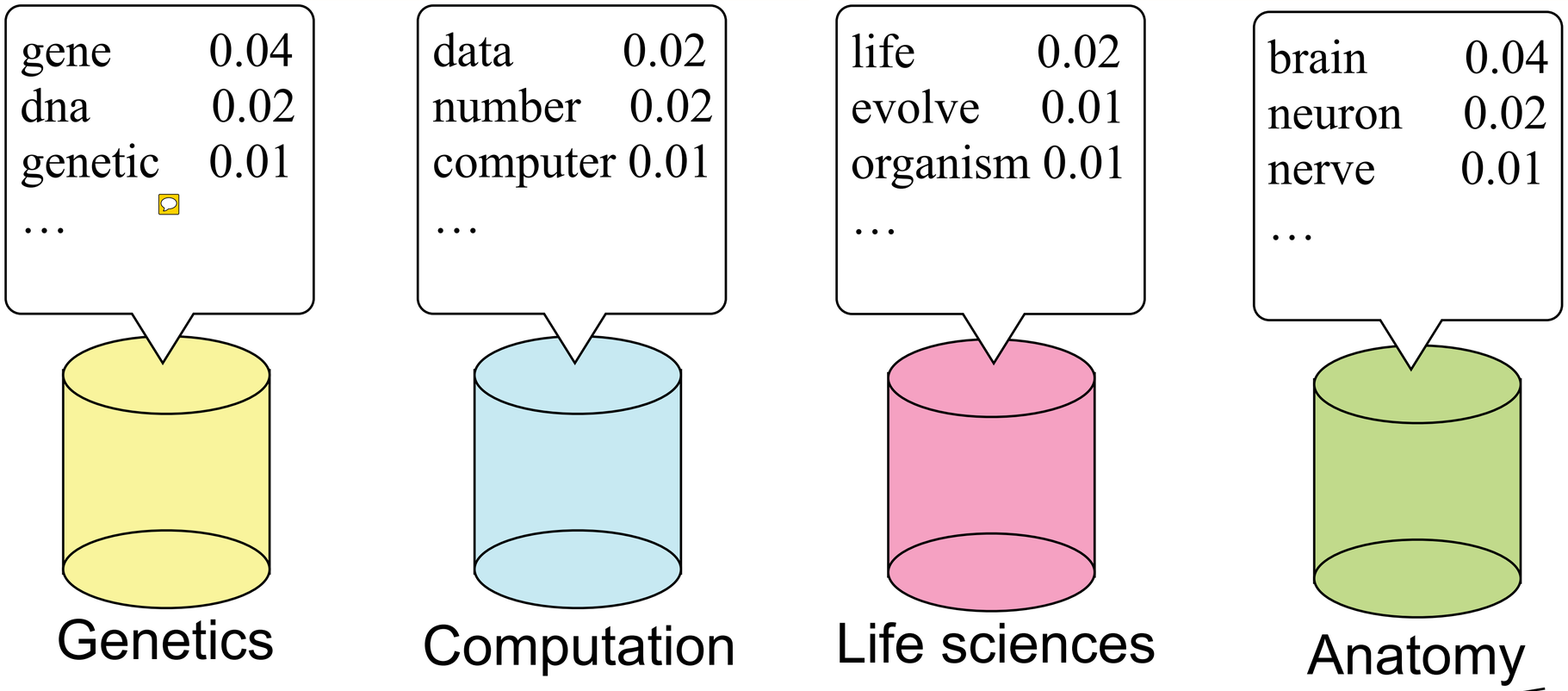
Latent Dirichlet Allocation
(Blei et al., '03)

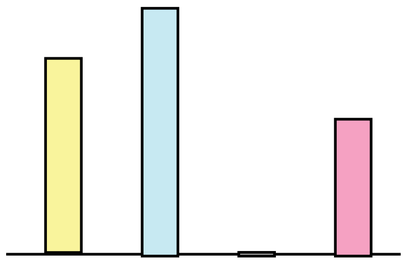
Topic 1: Genetics
gene, sequence, genome, …

Topic 2: Computation
number, computer, analysis, …

Topic 3: Life sciences
life, survive, organism, …

Intuition: Documents as a mixture of topics





What is Topic Modeling?

- A course-level analysis of what's in a text collection
- Topic : the subject (theme) of a discourse
- Topics are represented as a word distribution
- A document is assumed to be a mixture of topics

More examples of topics

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

(Figure courtesy Prof. David Blei)

What is Topic Modeling? (2)

- What's known:
  - The text collection or corpus
  - Number of topics
- What's not known:
  - The actual topics
  - Topic distribution for each document

What is Topic Modeling? (3)

- Essentially, text clustering problem
  - Documents and words clustered simultaneously

- Different topic modeling approaches available
  - Probabilistic Latent Semantic Analysis (PLSA) [Hoffman '99]
  - Latent Dirichlet Allocation (LDA) [Blei, Ng, and Jordan, '03]