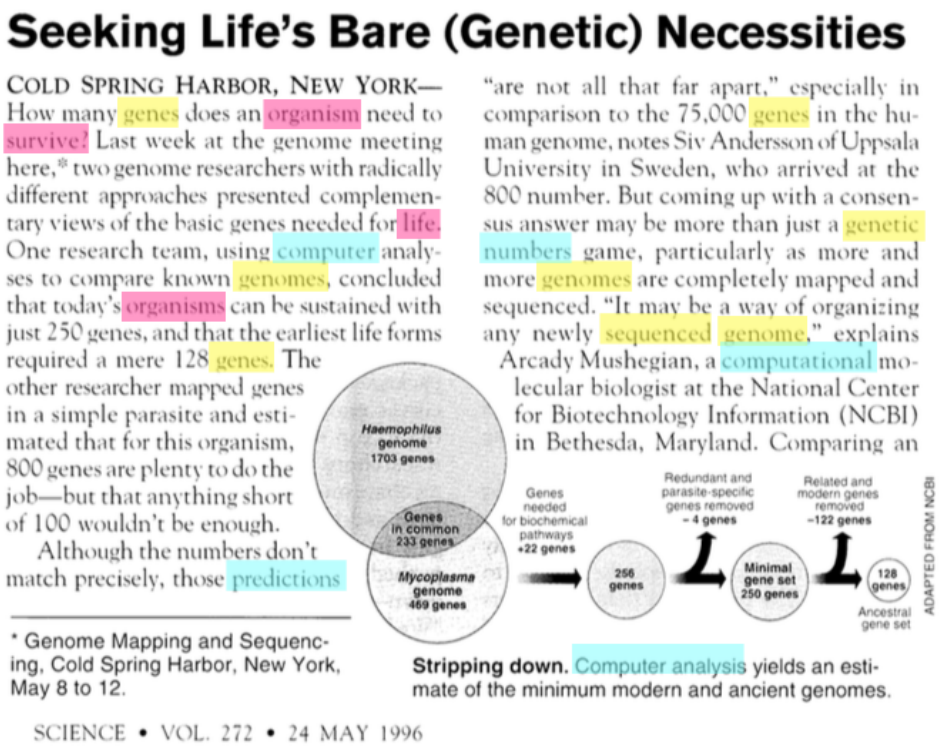
Latent Dirichlet Allocation (LDA)

Rebecca C. Steorts
Predictive Modeling and Data Mining: STA 521

October 2015

- Recall what we did in information retrieval.
- Review this.

Intuition behind LDA



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.

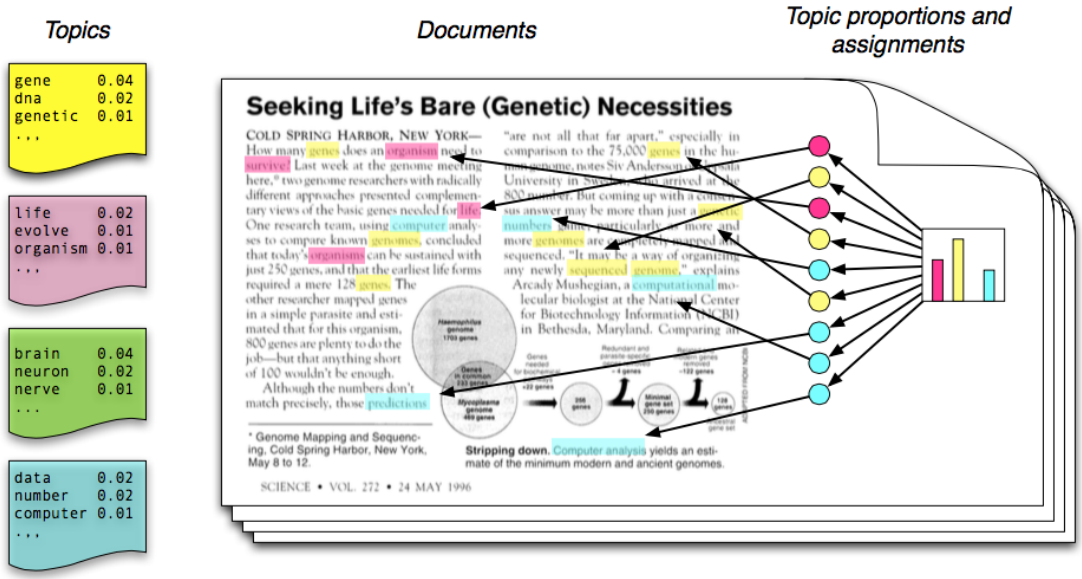
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Simple intuition: Documents exhibit multiple topics.

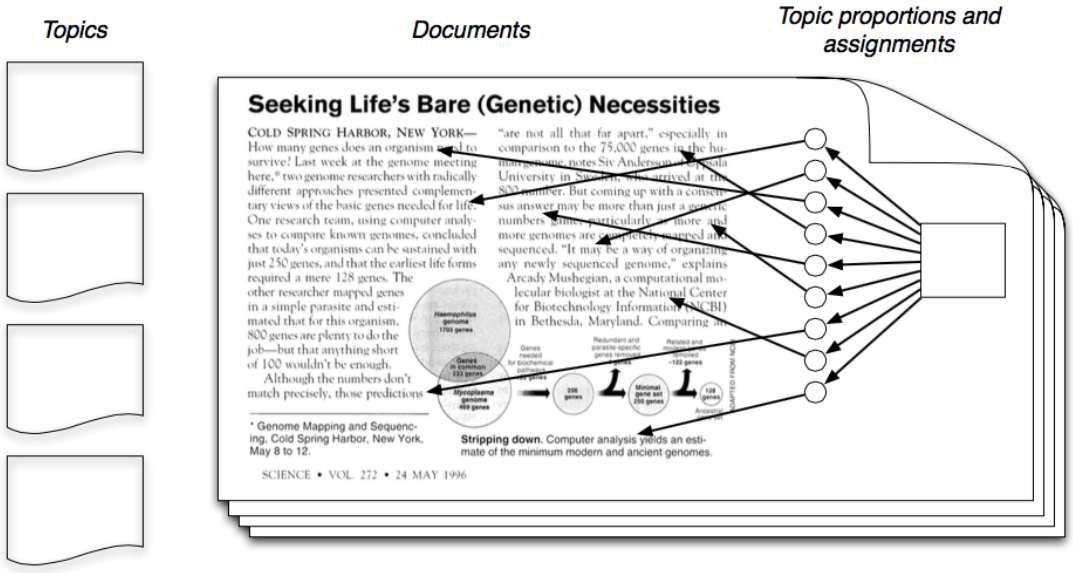
Figure 1: Simple intuition: Documents exhibit multiple topics

Probabilitistic Model



- Each document is a random mixture of corpus-wide topics.
- Each word is drawn from one of those topics.

Probabilistic Model



- We ONLY observe the documents.
- Our goal is to infer the underlying topic structure.

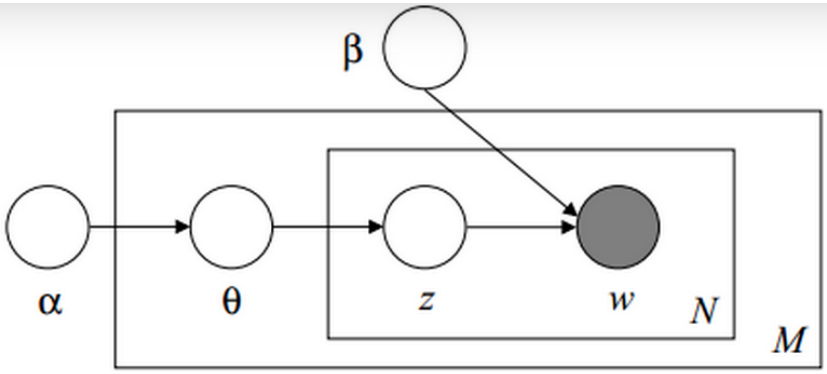
Probabilistic Model

- Observations come from a generative model that include hidden or latent (unknown, random) variables.
- Infer the hidden structure using posterior inference.
- Situate new data into the estimated model. How does a query or new document fit into the estimated topic structure?

Notation

- word $1, \ldots, V$
- document: $\boldsymbol{w} = (w_1, \ldots, w_N)$ which is a sequence of N words
- corpus: $D = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_M)$ collection of M documents

Probabilitistic Model



$$N \sim Poisson(\eta) \tag{1}$$

$$\theta \sim Dir(\alpha) \tag{2}$$

For each of N words w_n :

$$z_n(topic) \sim Multinomial(\theta)$$

$$w_n(word) \sim P(w_n \mid z_n, \beta)$$

Does this model make sense?

- N: total number of words (Poisson seems reasonable).
- θ: is the parameter from the Multinomial.
- What about the topics?

Note: if you're not familiar with the Dirichlet distribution, please go look up some basic facts about it.

LDA in R

```
install.packages(c("RTextTools","topicmodels"))
library(RTextTools)
library(topicmodels)
```

LDA in R

- This dataset contains headlines from front-page NYTimes articles.
- We will take a random sample of 1000 articles.

```
data(NYTimes)
data <- NYTimes[sample(1:3100,size=1000,replace=FALSE),]
```

I love that DTM

- Our text data consists of the Title and Subject columns of the NYTimes data.
- We will be removing numbers, stemming words, and weighting the DocumentTermMatrix by term frequency.

```
matrix <- create_matrix(cbind(as.vector(data$Title),
as.vector(data$Subject)), language="english",
removeNumbers=TRUE, stemWords=TRUE, weighting=weightTf)
```

Perform LDA

- First, determine the number of topics in the dataset.

```
k <- length(unique(data$Topic.Code))
lda <- LDA(matrix, k)
```

Results of LDA

- View the results most likely topic per document.

```
terms(lda)
Topic 1  "campaign"  Topic 2  "kill"       Topic 3  "elect"
Topic 13 "republican"Topic 14 "aid"        Topic 15 "set"
Topic 19 "iraq"      Topic 20 "bush"       Topic 21 "citi"
Topic 25 "basebal"   Topic 26 "court"      Topic 27 "war"
topics(lda)
```