universität
wien

Topic Modeling

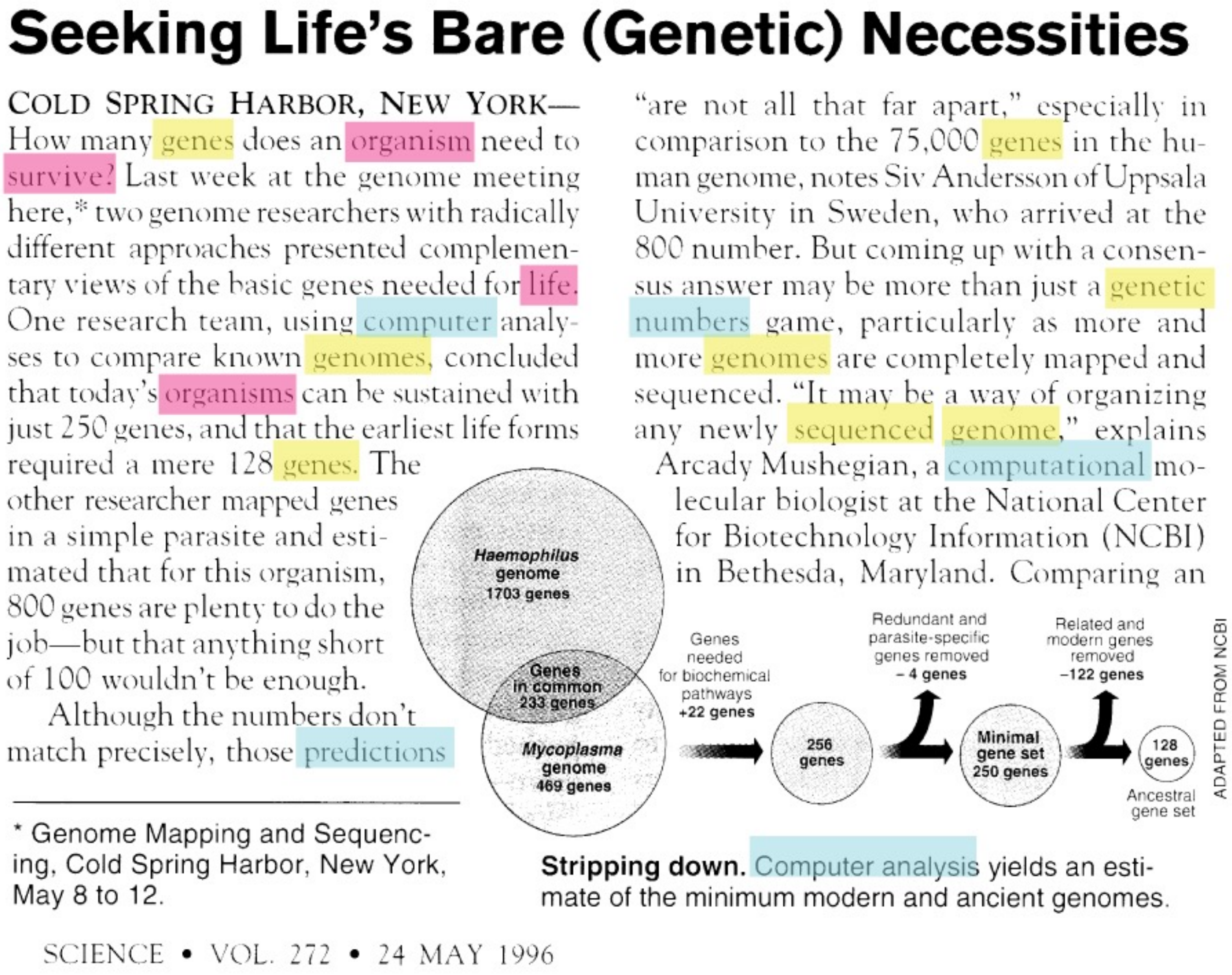
Gabriel Viehhauser

Topic Modeling

- Method to extract topics (thematic clusters?) from a large number of documents

- Based on distributional semantics (John Rupert Firth: You shall know a word by the company it keeps)

- An (old, but still great) example: http://signsat40.signsjournal.org/



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

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed − 4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

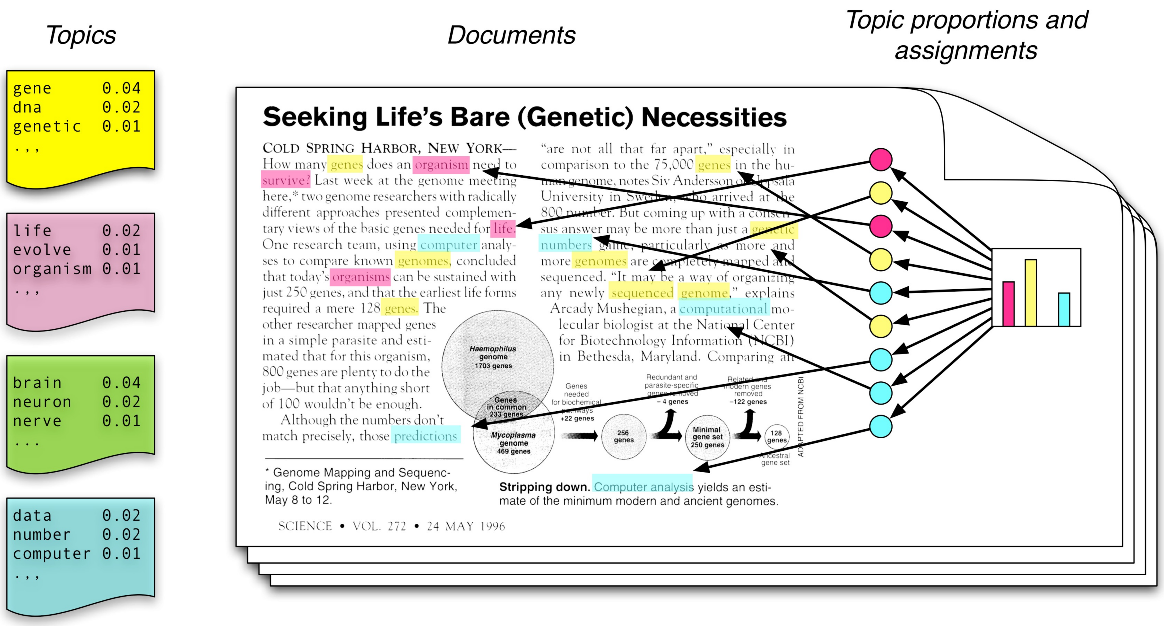
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

aus: David Blei: Probalistic Topic Models. August 22, 2011

Generative model for LDA



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—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

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed – 4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

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

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

- -> **Latent Dirichlet Allocation** (LDA)

aus: David Blei: Probalistic Topic Models. August 22, 2011

LDA – Gibbs Sampling

- Start with random assignment of words to topics.
- For each document, go through each word w and consider:
  - the proportion of words in the document d that are assigned to topic t. If a lot of words from d are in t, it is more likely, that w is in t.
  - How often does word w appear in topic t elsewhere? If w appears in t very often, this instance of w is most likely part of t.
- Update the probability

  p(word w with topic t) = p(topic t | document d) * p(word w | topic t)

- Example of one step:
  Topic 1 = nature, Topic 2 = city.
  We assigned „tree“ to t1, „building“ and „car“ to t2
  We see a document „The tree is in front of the building and behind a car“
  -> probability for „tree“ in t1 will decrease, because „tree“ normally is in t1, but in this document, there are only t2-words. On the other hand, probability for „tree“ in t2 will increase.

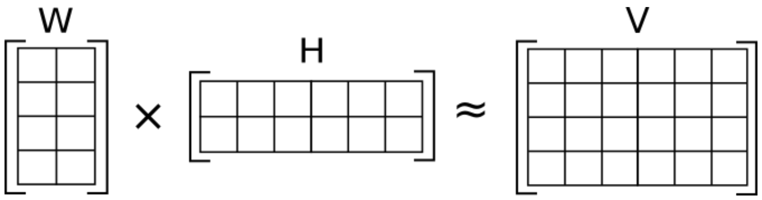
https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2

Topic Modeling - Workflow

- Tokenize Texts
- Preprocessing: Exclude Stopwords, proper names or certain POS
- Create Document-Term-Matrix
- Determine number of topics
- Let computer calculate the topics in several iterations
- Inspect model and adapt parameters
- Interpret and name topics.

Nonnegative matrix factorization (NMF)

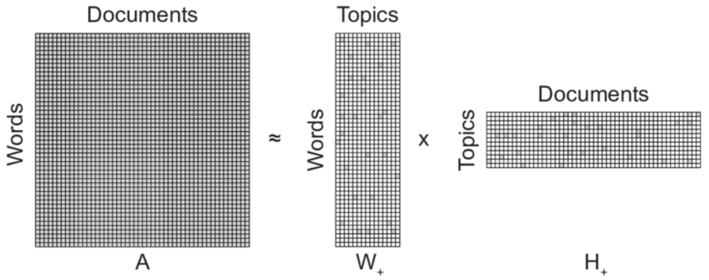
- automatically extracts sparse and meaningful features from a set of nonnegative data vectors.



Quwertyus, https://commons.wikimedia.org/wiki/File:NMF.png

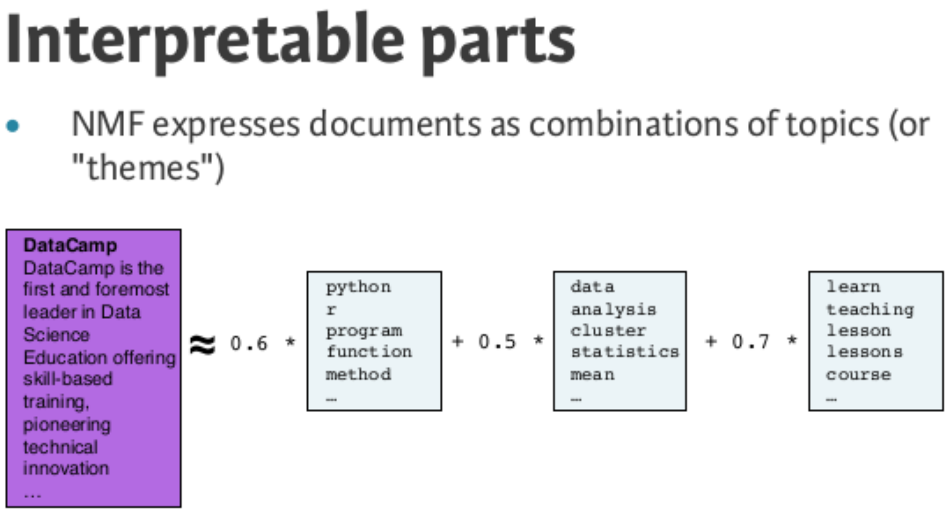
Nonnegative matrix factorization (NMF)

- *„NMF is designed for discovering interpretable latent components in high-dimensional unlabeled data such as the set of documents described by the counts of unique words. NMF uncovers major hidden themes by recasting the term-document matrix A into the product of two other matrices, one matrix representing the relationships between words and topics and another representing the relationship between topics and documents in the latent topic space‟*



Kuang, Da & Brantingham, P. & Bertozzi, Andrea. (2017). Crime Topic Modeling. Crime Science. 6. 10.1186/s40163-017-0074-0.

Nonnegative matrix factorization (NMF)



Interpretable parts

- NMF expresses documents as combinations of topics (or "themes")

DataCamp
DataCamp is the first and foremost leader in Data Science Education offering skill-based training, pioneering technical innovation
...

≈ 0.6 *

```
python
r
program
function
method
...
```

+ 0.5 *

```
data
analysis
cluster
statistics
mean
...
```

+ 0.7 *

```
learn
teaching
lesson
lessons
course
...
```

https://goldinlocks.github.io/Non-negative-matrix-factorization/