OPEN DATA SCIENCE CONFERENCE

Boston | May 1 - 4 2018

#ODSC

@ODSC

#ODSC

@ODSC

Topic Modeling:
From *doc-term matrix* to *Latent Dirichlet Allocation*

The Plan for Topic Modeling

- Motivate topic modeling
- Outline the evolution from word-space models to probabilistic models
- Discuss Latent Semantic Analysis and Latent Dirichlet Allocation
- Implement all of these models using sklearn and gensim

Topic Modeling: Goals

Goal: automatically organize, understand, search & summarize (lots of) text

- Capture semantic information beyond individual words
- Discover hidden topics or themes across documents
- Annotate documents accordingly
- Use annotations to manage, summarize, search and recommend content

From Bags of Words to Latent Topics

Model	Year	Description
Vector Space Model	1975	Documents as vectors in word space
Latent Semantic Analysis	1988	Capture semantic term-document relationship through dimensionality reduction of the word space
Probabilistic LSA	1999	Words generated a topic, documents as mix of topics
Latent Dirichlet Allocation	2003	Adds generative process for documents: three-level hierarchical, Bayesian model

Document-Term Matrix

Text to Numbers

Bag of n Terms

	Term 1	Term 2	. . .	Term n-1	Term n
Doc 1	0	1	1	1	0
Doc 2	0	1	0	0	0
Doc 3	2	0	3	0	0
⋮	.	.	.	.	.
	.	.	.	.	.
Doc m-2	1	0	2	0	0
Doc m-1	0	0	1	0	0
Doc m	0	1	0	0	1

Term Weights

m Documents as vectors in Term Space

Vector Space Model: In Practice

Challenge: large # of unique terms, but each doc only contains a small subset

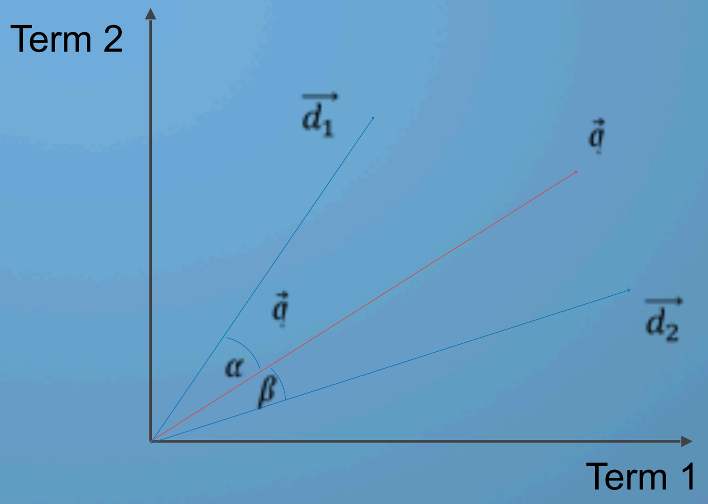
- Remove less informative (stop) terms: too high or too low in frequency
- Consolidate terms using stemming or lemmatization

Similarity Query

- Query vector $\vec{q}$
- Compare documents $\vec{d_1}, \vec{d_2}$

=> Max. cosine similarity [0, 1]

$$\cos(\alpha) = \frac{\vec{d_1}\vec{q}}{\left\|\vec{d_1}\right\|\left\|\vec{q}\right\|}$$



Vector Space Model: Limitations

- **Curse of Dimensionality**: inaccurate distance metrics, overfitting
- **Sparse vectors**: noisy similarity measure
- **Loss of context**: bag of words model ignores word order
- **Loss of semantics**: similarity of words does not capture synonymy & polysemy

How to model topics or themes that represent semantic content and facilitate more productive interaction with text content?

Linear Algebra: Latent Semantic Indexing

- **Goal**: find latent topics by decomposing the term-document matrix
- **Solution**: reduce dimensionality via (Truncated) Singular Value Decomposition
- **Assumption**: best lower-rank approximation using K<N singular values & vectors

Document-Topic Similarity | Concept Strength | Term-Concept Similarity

N Terms

$$\text{M Docs}\ \begin{pmatrix} & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & \end{pmatrix} = \begin{pmatrix} & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & \end{pmatrix}\begin{bmatrix} & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & \end{bmatrix}\begin{pmatrix} & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & \end{pmatrix}$$

Singular Vectors | Singular Values (diagonal) | Singular Vectors (transposed)

U | Σ | V^T

M x K | K x K | K x N

Latent Semantic Indexing: Pros & Cons

Pros:

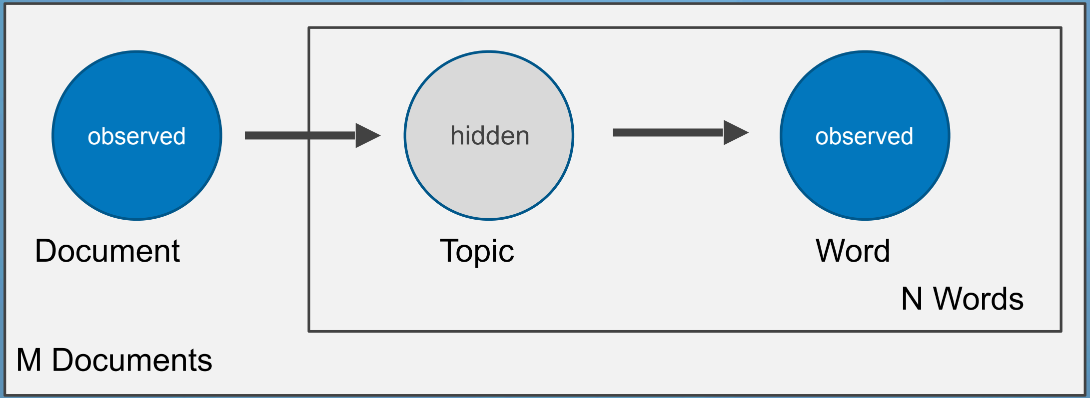
- **Dimensionality Reduction**: helps address curse, removes noise
- **Context Space**: captures some semantics, clustering of docs & terms

Cons:

- **Hard to interpret**: topics as word vectors with positive & negative entries
- **No probabilistic model**: harder to evaluate fit, select number of dimensions

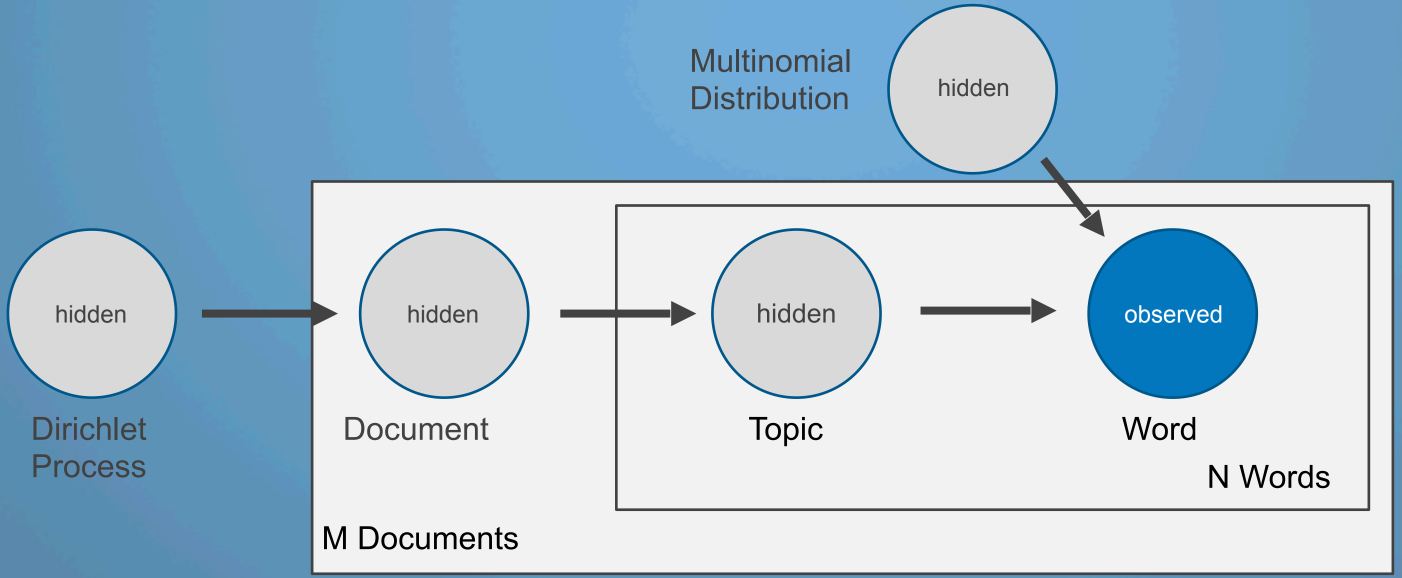
Probabilistic Models: probabilistic LSI

- **Goal**: model the origination of documents and terms based on topics
- **Solution**: Generative model with topics as latent (hidden) variables
- **Assumption**: words sampled from topics, and docs are a (given) mix of topics
- **Model:** Estimate parameters to maximize data likelihood using EM algorithm

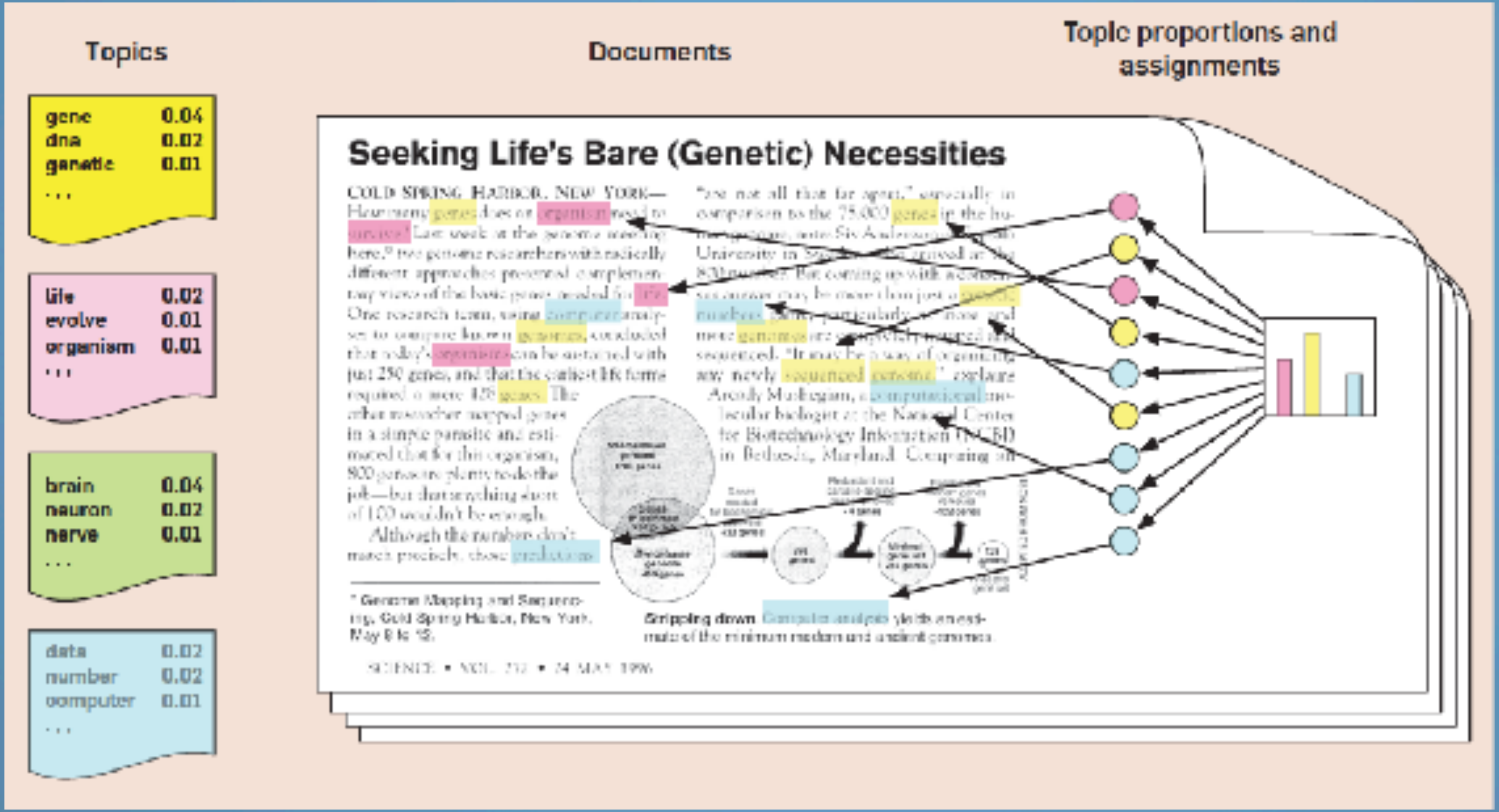


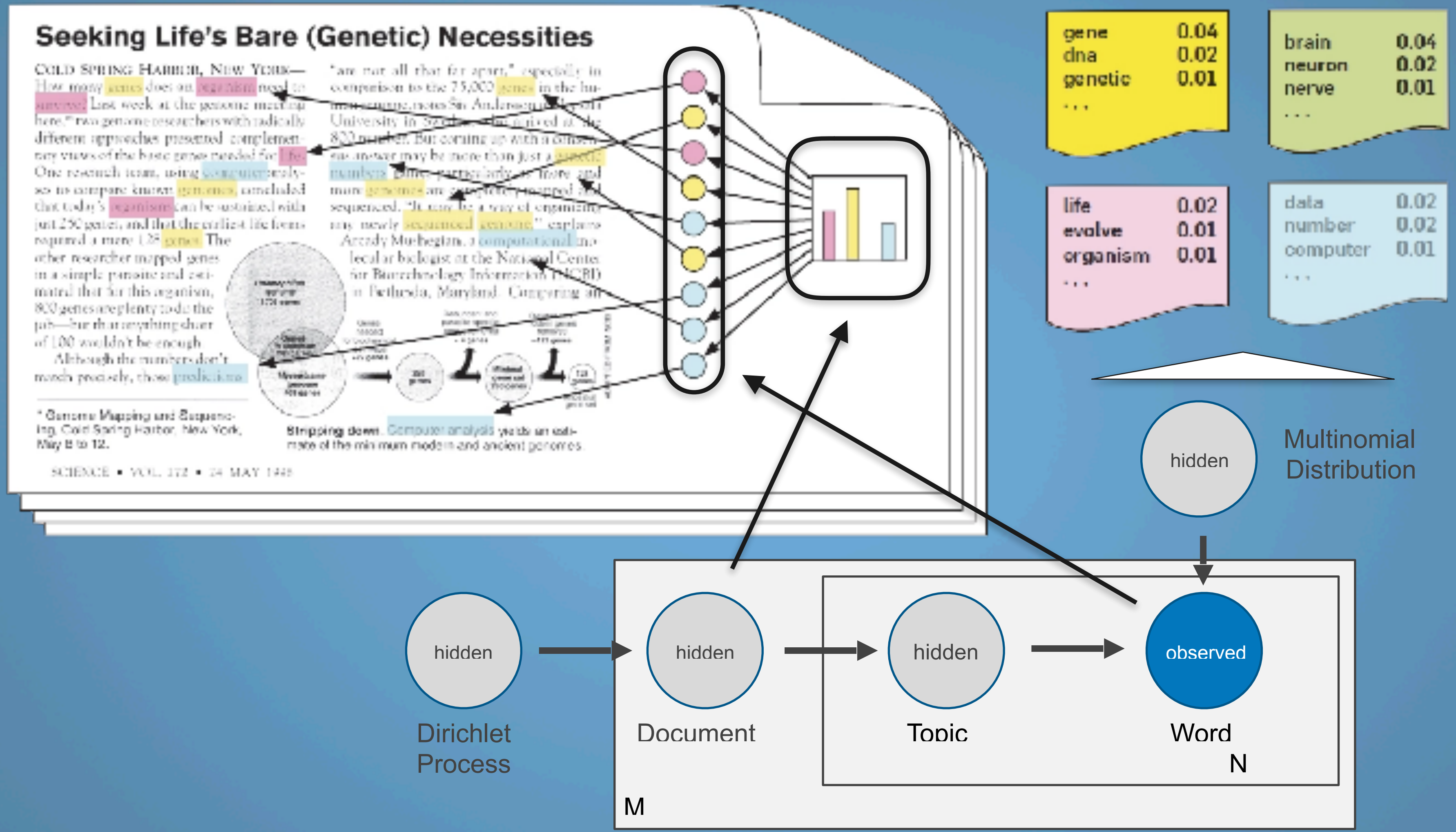
Probabilistic Models: Latent Dirichlet Allocation

- **Goal**: extend probabilistic model to document layer
- **Solution**: sample topics for documents using Dirichlet process
- **Assumption**: three-level model: number of words, mix of topics, word choice



LDA: From Topics to Documents to Written Text - and back




Seeking Life's Bare (Genetic) Necessities
gene 0.04
dna 0.02
genetic 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
data 0.02
number 0.02
computer 0.01
...
hidden
Multinomial Distribution
hidden
Dirichlet Process
hidden
Document
hidden
Topic
observed
Word
N
M

LDA: Pros & Cons

Pros:

- **Meaningful Topics**: tends to produce topics that humans can relate to
- **Fully generative**: can assign topics to new documents
- **Extensible**: use metadata, apply to image data, hierarchical topics

Resources

- Topic Modeling:
  - https://github.com/stefan-jansen/topic-modeling