

# OPEN DATA SCIENCE CONFERENCE



@ODSC

Boston | May 1 - 4 2018



Stefan Jansen  
Applied AI



@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 to probabilistic models
- Discuss Latent Semantic Analysis and Latent Dirichlet Allocation
- Implement all of these models using sklearn and gensim
- Evaluate topic models using pyLDAvis & topic coherence

## Topic Modeling: Goals

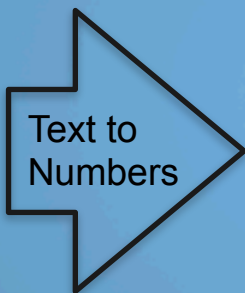
**Goal:** automatically organize, understand, search & summarize (lots of) text

- 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



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

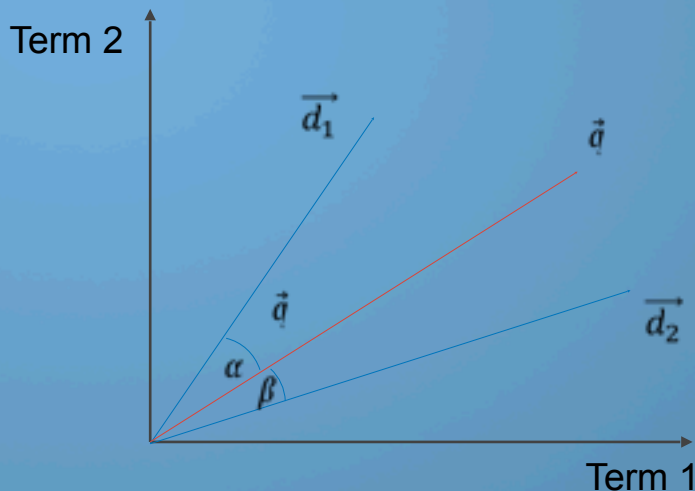
- Remove less informative (stop) terms: too high or too low in frequency
- Consolidate terms using stemming or lemmatization

### Similarity Query $\vec{q}$

- Query vector  $\vec{q}$
- Compare documents  $\vec{d}_1, \vec{d}_2$

=> Max. cosine similarity [0, 1]

$$\cos(\alpha) = \frac{\vec{d}_1 \cdot \vec{q}}{\|\vec{d}_1\| \|\vec{q}\|}$$



## 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:** word representation 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

$$\begin{array}{ccccccc}
 & \text{N Terms} & & \text{Document-Topic Similarity} & \text{Concept Strength} & \text{Term-Concept Similarity} & \\
 \text{M Docs} & \begin{pmatrix} & \dots & \\ \vdots & \ddots & \vdots \\ & \dots & \end{pmatrix} & = & \begin{pmatrix} & \dots & \\ \vdots & \ddots & \vdots \\ & \dots & \end{pmatrix} & \begin{bmatrix} & \dots & \\ \vdots & \ddots & \vdots \\ & \dots & \end{bmatrix} & \begin{pmatrix} & \dots & \\ \vdots & \ddots & \vdots \\ & \dots & \end{pmatrix} \\
 & & & \text{Singular Vectors} & \text{Singular Values (diagonal)} & \text{Singular Vectors (transposed)} & \\
 & & & \mathbf{U} & \mathbf{\Sigma} & \mathbf{V}^T & \\
 & & & \text{M} \times \text{K} & \text{K} \times \text{K} & \text{K} \times \text{N} & 
 \end{array}$$

# Latent Semantic Indexing: Pros & Cons

## Pros:

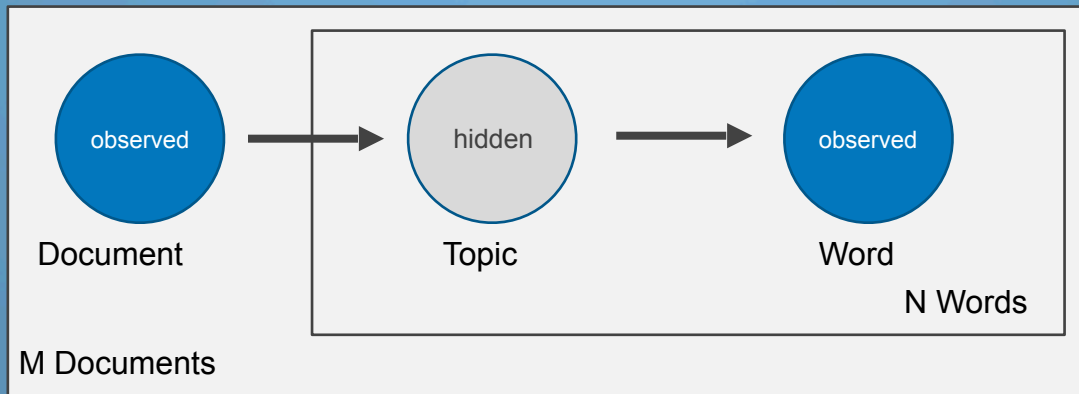
- **Dimensionality Reduction:** helps address curse, removes noise
- **Context Space:** captures some semantics, clustering of docs & terms

## Cons:

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

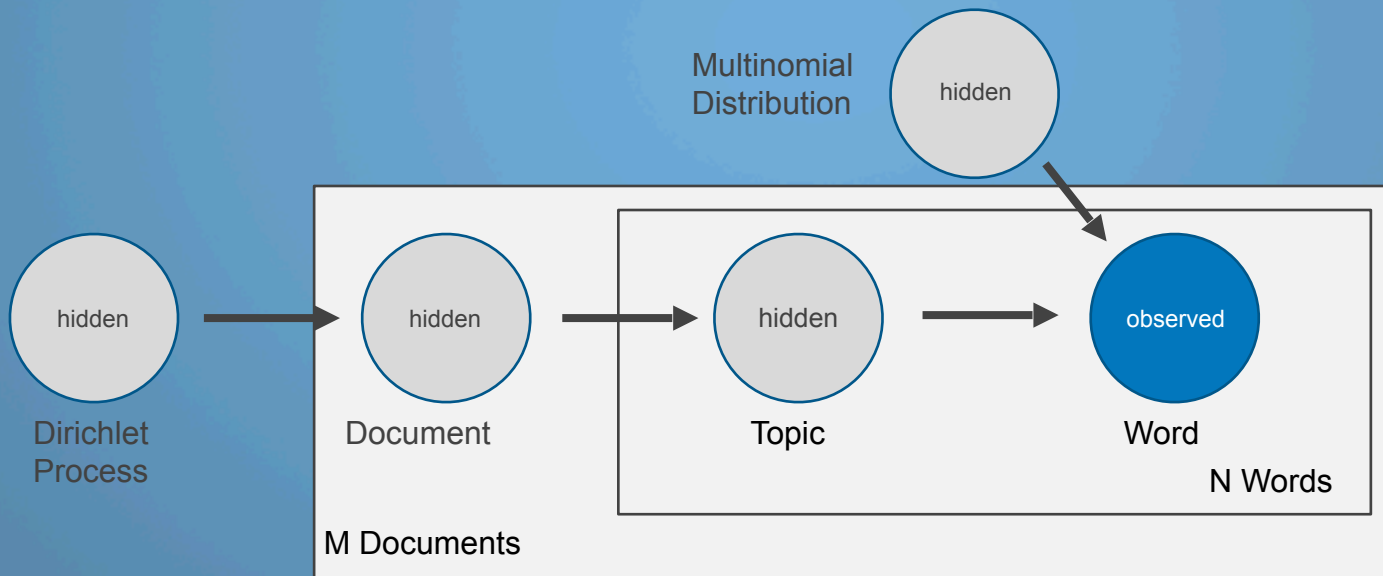
## Probabilistic Models: probabilistic LSI

- **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

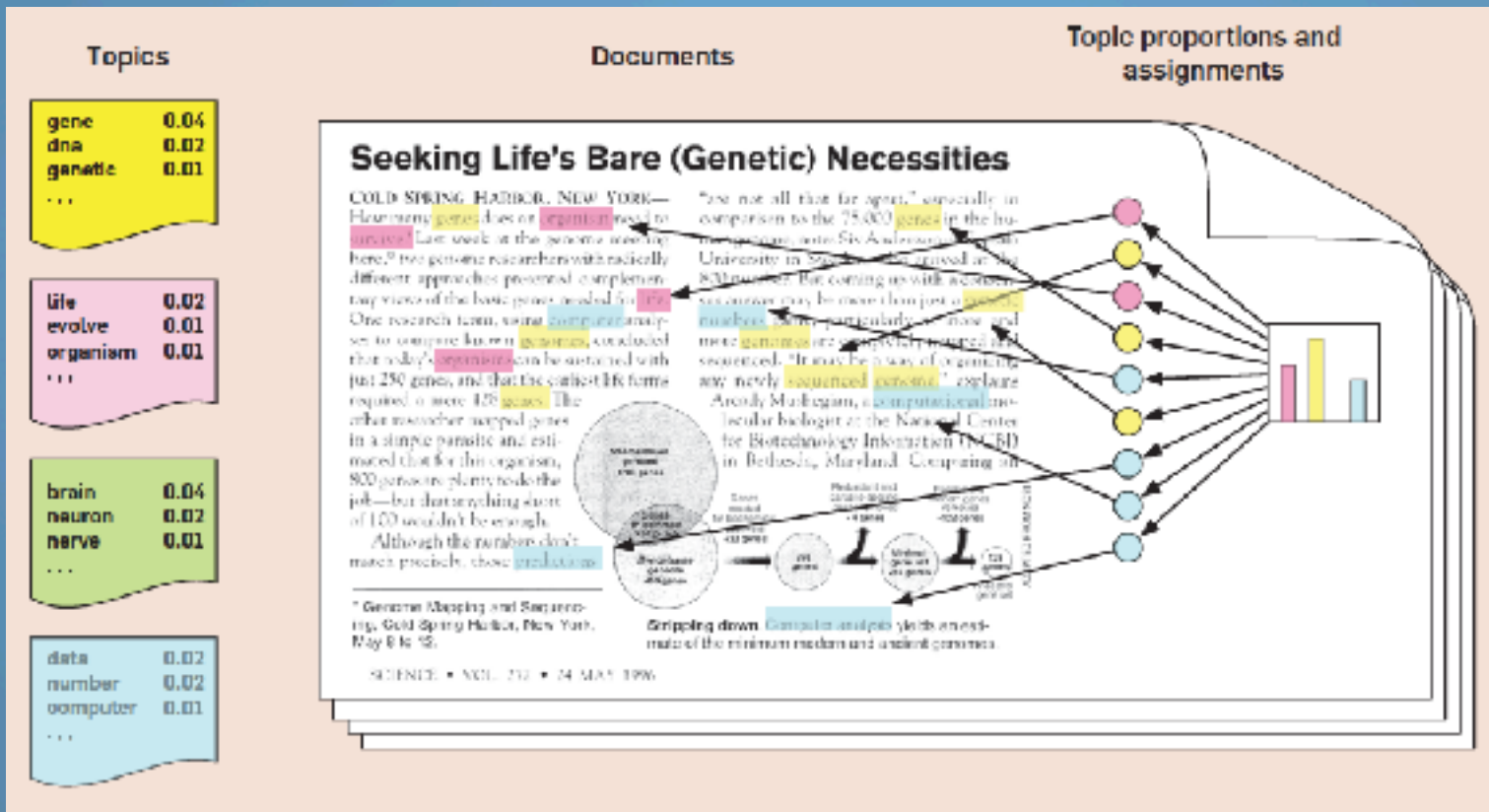


# Probabilistic Models: Latent Dirichlet Allocation

- **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

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 **sequencing** to compare known **genomes**, concluded that today's **genome** can be sustained with just 290 genes, and that the earliest life forms required a mere 125 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 San Anderson, a biologist at the University of Seattle, who chaired at the 800 meeting. But coming up with a consensus answer may be more than just a **genetic** matter. Some, particularly in **more** and **more** genomes are completely mapped and sequenced, "it may be a way of organizing our newly **sequenced** genome," explains Arcady Mushegian, a **computational** and **molecular biologist** at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing all



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

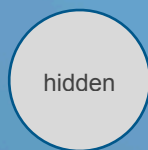
SCIENCE • VOL. 172 • 24 MAY 1998

gene	0.04
dna	0.02
genetic	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

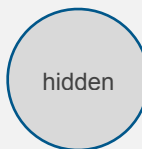
life	0.02
evolve	0.01
organism	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

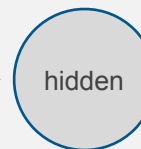


Dirichlet Process

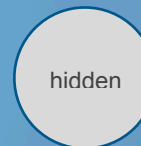
M



Document



Topic



Multinomial Distribution



Word N

# 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>