Information School
UNIVERSITY OF WISCONSIN–MADISON

Topic Modeling

The Information School, UW-Madison

Text Normalization

1. Tokenizing (segmenting words)

2. Normalizing word formats

3. Segmenting sentences

The US is a big nation. Americans love the U.S.A. a lot. They like to drive their cars around the country. They measure speed in m.p.h and not km.p.h.

Text Normalization

1. **Tokenizing (segmenting words)**

2. Normalizing word formats

3. Segmenting sentences

‘The’, ‘US’, ‘is’, ‘a’, ‘big’, ‘nation’, ‘.’, ‘Americans’, ‘love’, ‘the’, ‘U’, ‘.’, ‘S’, ‘.’, ‘A’, ‘.’, ‘a’, ‘lot’, ‘.’, ‘They’, ‘like’, ‘to’, ‘drive’, ‘their’, ‘cars’, ‘around’, ‘the’, ‘country’, ‘.’, ‘They’, ‘measure’, ‘speed’, ‘in’, ‘m’, ‘.’, ‘p’, ‘.’, ‘h’, ‘and’, ‘not’, ‘km’, ‘.’, ‘p’, ‘.’, ‘h’, ‘.’

Text Normalization

1. Tokenizing (segmenting words)
2. **Normalizing word formats**
3. Segmenting sentences

‘The’, **‘US’**, ‘is’, ‘a’, ‘big’, ‘nation’, ‘.’, ‘Americans’, ‘love’, ‘the’, **‘U’, ‘.’, ‘S’, ‘.’, ‘A’**, ‘.’, ‘a’, ‘lot’, ‘.’, ‘They’, ‘like’, ‘to’, ‘drive’, ‘their’, ‘cars’, ‘around’, ‘the’, ‘country’, ‘.’, ‘They’, ‘measure’, ‘speed’, ‘in’, ‘m’, ‘.’, ‘p’, ‘.’, ‘h’, ‘and’, ‘not’, ‘km’, ‘.’, ‘p’, ‘.’, ‘h’, ‘.’

Text Normalization

1. Tokenizing (segmenting words)
2. **Normalizing word formats**
3. Segmenting sentences

‘The’, **‘US’**, ‘is’, ‘a’, ‘big’, ‘nation’, ‘.’, ‘Americans’, ‘love’, ‘the’, **‘US’**, ‘.’, ‘a’, ‘lot’, ‘.’, ‘They’, ‘like’, ‘to’, ‘drive’, ‘their’, ‘cars’, ‘around’, ‘the’, ‘country’, ‘.’, ‘They’, ‘measure’, ‘speed’, ‘in’, ‘m’, ‘.’, ‘p’, ‘.’, ‘h’, ‘and’, ‘not’, ‘km’, ‘.’, ‘p’, ‘.’, ‘h’, ‘.’

Text Normalization

1. Tokenizing (segmenting words)

2. Normalizing word formats

3. **Segmenting sentences**

‘The’, ‘US’, ‘is’, ‘a’, ‘big’, ‘nation’, ‘.’, ‘Americans’, ‘love’, ‘the’, ‘US’, ‘.’, ‘a’, ‘lot’, ‘.’, ‘They’, ‘like’, ‘to’, ‘drive’, ‘their’, ‘cars’, ‘around’, ‘the’, ‘country’, ‘.’, ‘They’, ‘measure’, ‘speed’, ‘in’, ‘m’, ‘.’, ‘p’, ‘.’, ‘h’, ‘and’, ‘not’, ‘km’, ‘.’, ‘p’, ‘.’, ‘h’, ‘.’

What is topic modeling?

- Unsupervised methods to discover "topics" in a corpus.
- In most popular methods, topics are represented as word distributions and are learned from word co-occurrence information (and thus are corpus dependent).

A topic possibly related to "**air travel**"

Word	Prob
plane	0.082
airport	0.075
crash	0.048
flight	0.032
safety	0.028
aircraft	0.024
passenger	0.023

A topic possibly related to "**space shuttle**"

Word	Prob
space	0.101
shuttle	0.081
mission	0.042
astronauts	0.027
launch	0.026
station	0.024
nasa	0.020

A topic possibly related to "**Kobe earthquake**"

Word	Prob
building	0.098
city	0.087
people	0.068
rescue	0.042
buildings	0.038
kobe	0.031
victims	0.028

Some example topics learned from the TDT-1 corpus (adapted from Hofmann, 1999).

Core Assumptions

1. Documents discussing similar topics will use a similar group of words
2. Topics can be discovered by identifying groups of words in a corpus that frequently occur together

Structure of a Document

1. A document is a probability distribution over a set of topics
2. Topics are probability distribution over words

Harry Potter: Distribution of Topics

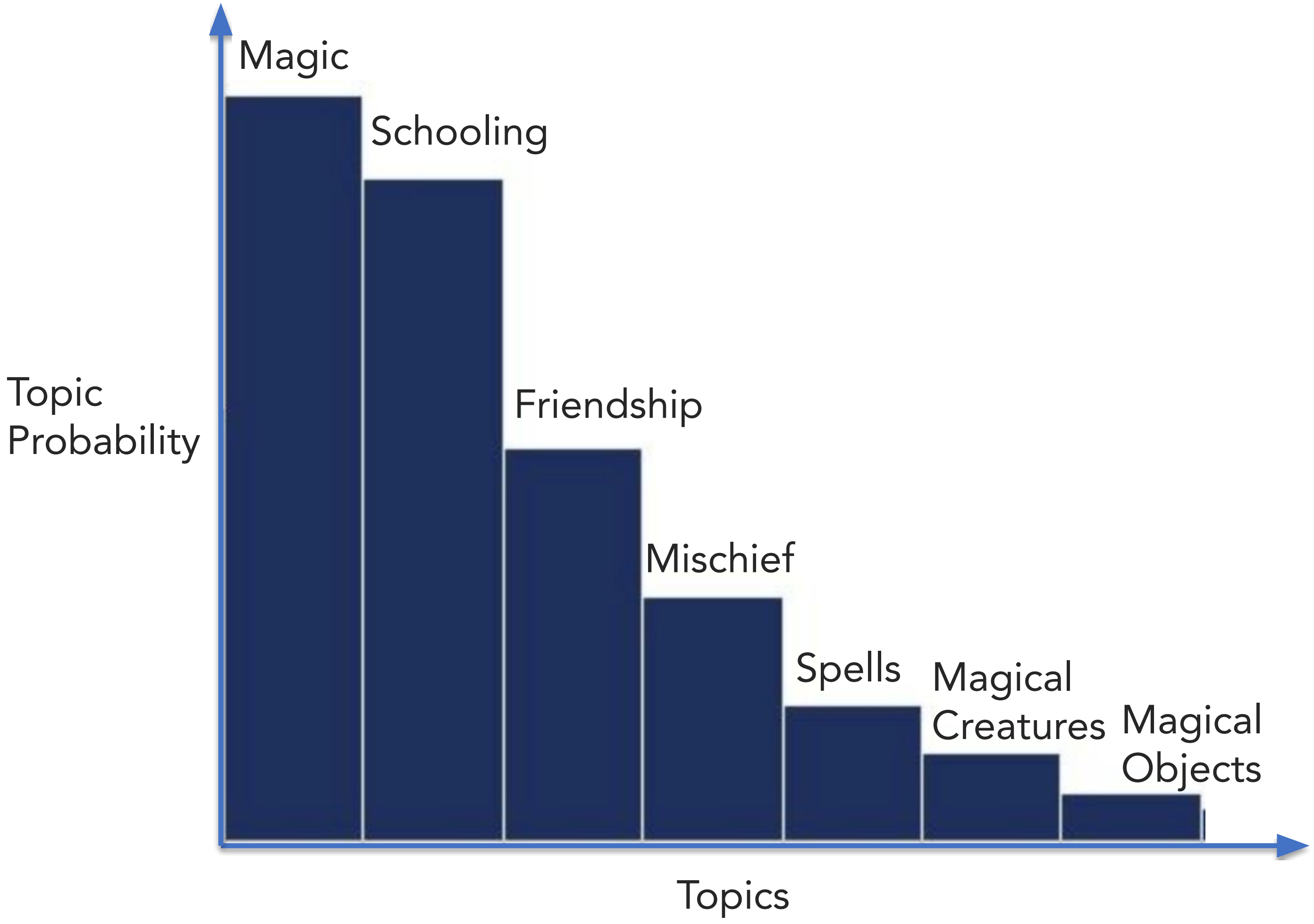
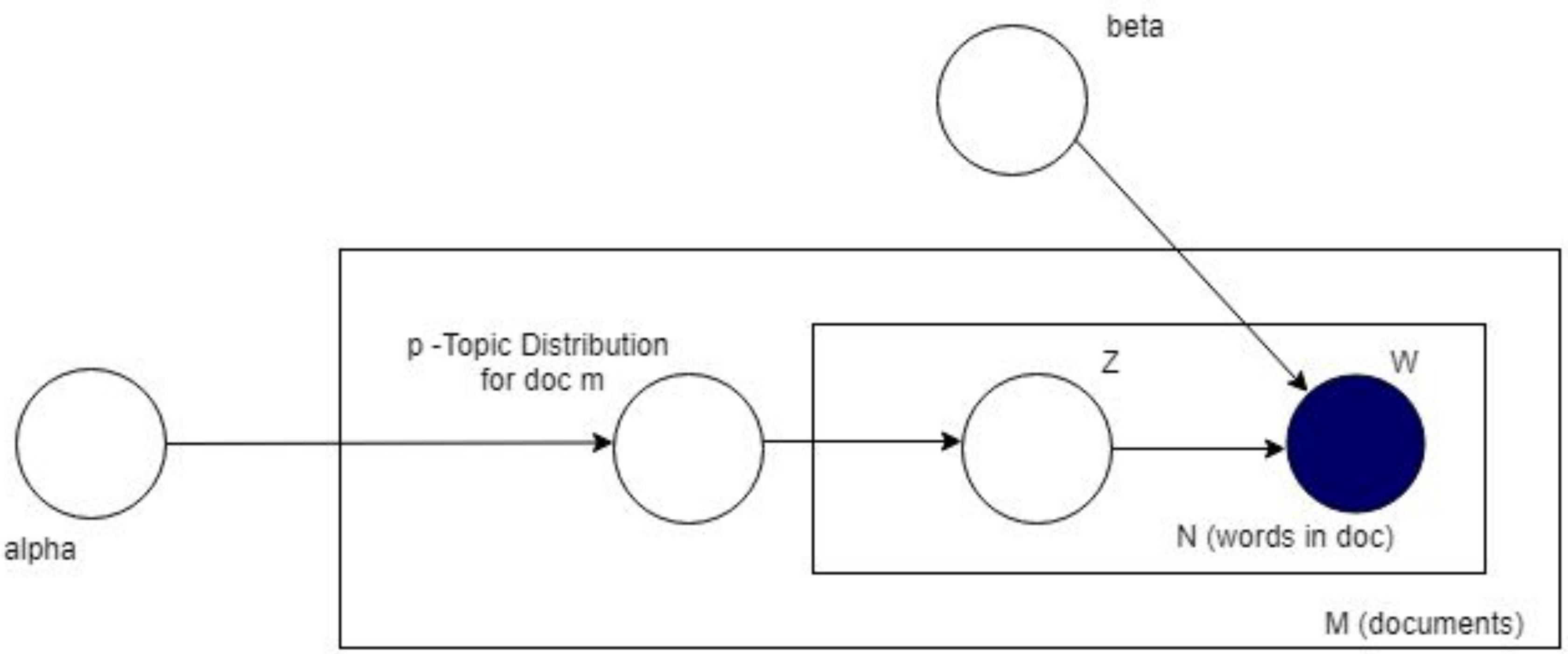


Plate Notation



Generative Process

LDA assumes that new documents are created in the following way -

1. Determine the number if words in a document

2. Choose a topic mixture for a document (i.e., 40% Topic A, 30% Topic B, 20 %Topic C, 10% Topic D)

3. Generate the words in the document by:

    1. First pick a word based on the document's distribution above

    2. Next pick a word based on the topic's distribution

Working Backwards

LDA works backwards from the generative process -

1. Suppose you have a set of documents

2. You want LDA to learn the topic representation of K-topics (in each doc) and the word distribution of each topic
3. LDA backtracks from the doc level to identify topics that are likely to have generated set of documents.

Working Backwards

Randomly assign each word in each document to one of the K topics.

For each document d:

Assume that all topic assignments except for the current one are correct.

Calculate two proportions:

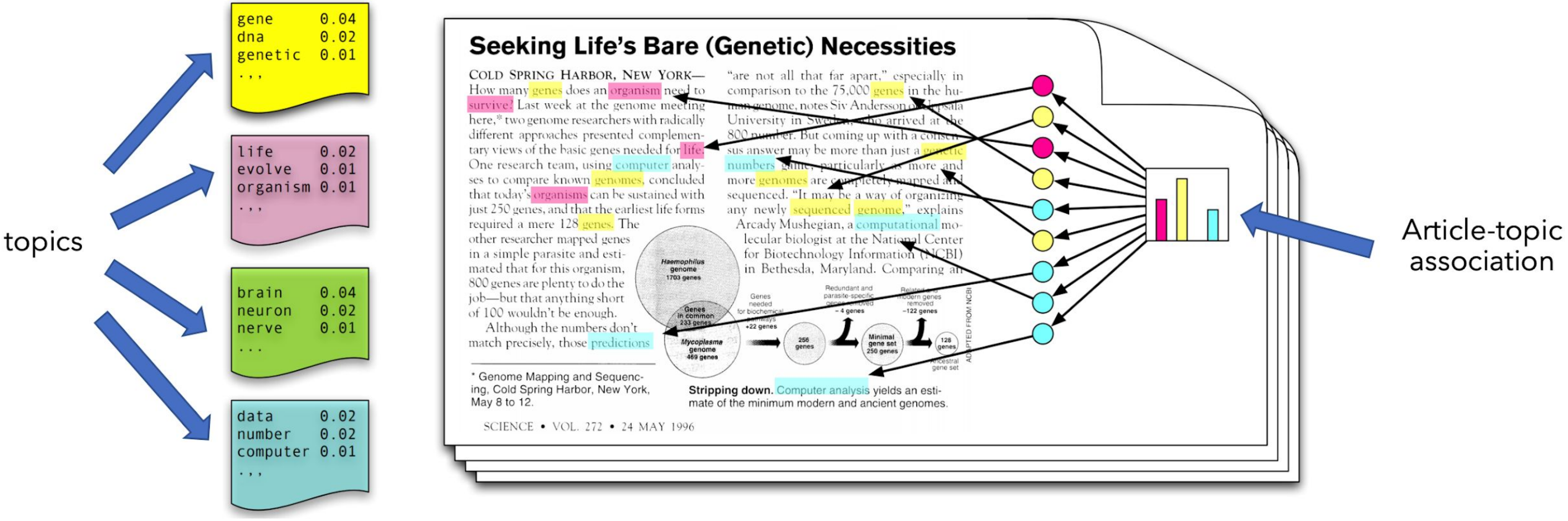
1. Proportion of words in document d that are currently assigned to topic t = p(topic t | document d)
2. Proportion of assignments to topic t over all documents that come from this word w = p(word w | topic t)

Multiply those two proportions and assign w a new topic based on that probability. p(topic t | document d) * p(word w | topic t)

Eventually we'll reach a steady state where assignments make sense

Topic Modeling: Typical Inputs & Outputs

- Inputs: the corpus; the number of topics
- Outputs: topics (word distributions); article-topic association
  - Note that topic labels are not a part of the outputs (but you can select words to label topics)



Example figure from Blei (2012).

Results from Sugimoto et al. (2011)

TABLE 5. Extended latent Dirichlet allocation results for 1990–1999 (741 dissertations).

	Topic 4a	Topic 4b	Topic 4c	Topic 4d	Topic 4e
Labels	Model development	Library outreach	Information seeking behavior	Library management	Information retreival
Words	research	library	information	variables	search
	study	libraries	work	study	users
	data	study	study	significant	user
	model	services	access	relationship	system
	analysis	academic	seeking	satisfaction	online
	process	librarians	sources	characteristics	searches
	identified	public	research	level	experience
	developed	support	personal	performance	searching
	development	data	resources	results	systems
	problem	respondents	related	found	task
	factors	professional	people	research	computer
	framework	university	individuals	significantly	interaction
	interviews	staff	environment	perceived	searchers
	based	education	behavior	factors	tasks
	approach	community	data	number	retrieval
	understanding	questionnaire	questions	analysis	browsing
	design	institutions	providers	perceptions	results
	studies	programs	human	academic	number
	environment	provided	professionals	relationships	types
	purpose	national	survey	related	participants