

Topic Modelling

COMP90042

Natural Language Processing

Lecture 20



THE UNIVERSITY OF
MELBOURNE

Making Sense of Text

- English Wikipedia: 6M articles
- Twitter: 500M tweets per day
- New York Times: 15M articles
- arXiv: 1M articles
- What can we do if we want to learn something about these document collections?

Questions

- What are the less popular topics on Wikipedia?
- What are the big trends on Twitter in the past month?
- How do the themes/topics evolve over time in New York Times from 1900s to 2000s?
- What are some influential research areas?

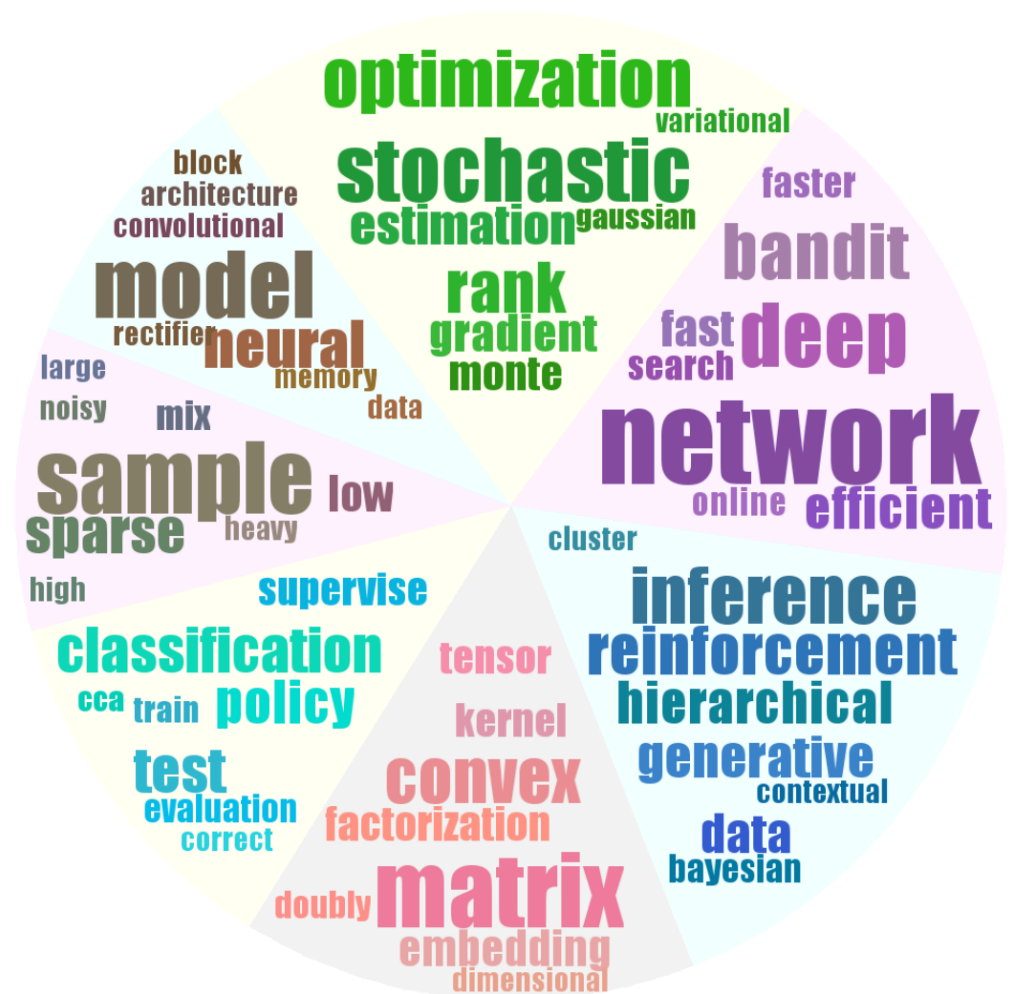
Topic Models To The Rescue

- Topic models learn common, overlapping themes in a document collection
- Unsupervised model *Just*
 - ▶ No labels; input is just the documents!

What Is A Topic?

- A set of words
- Collectively describes a concept or subject
- Words of a topic typically appear in the same set of documents in the corpus

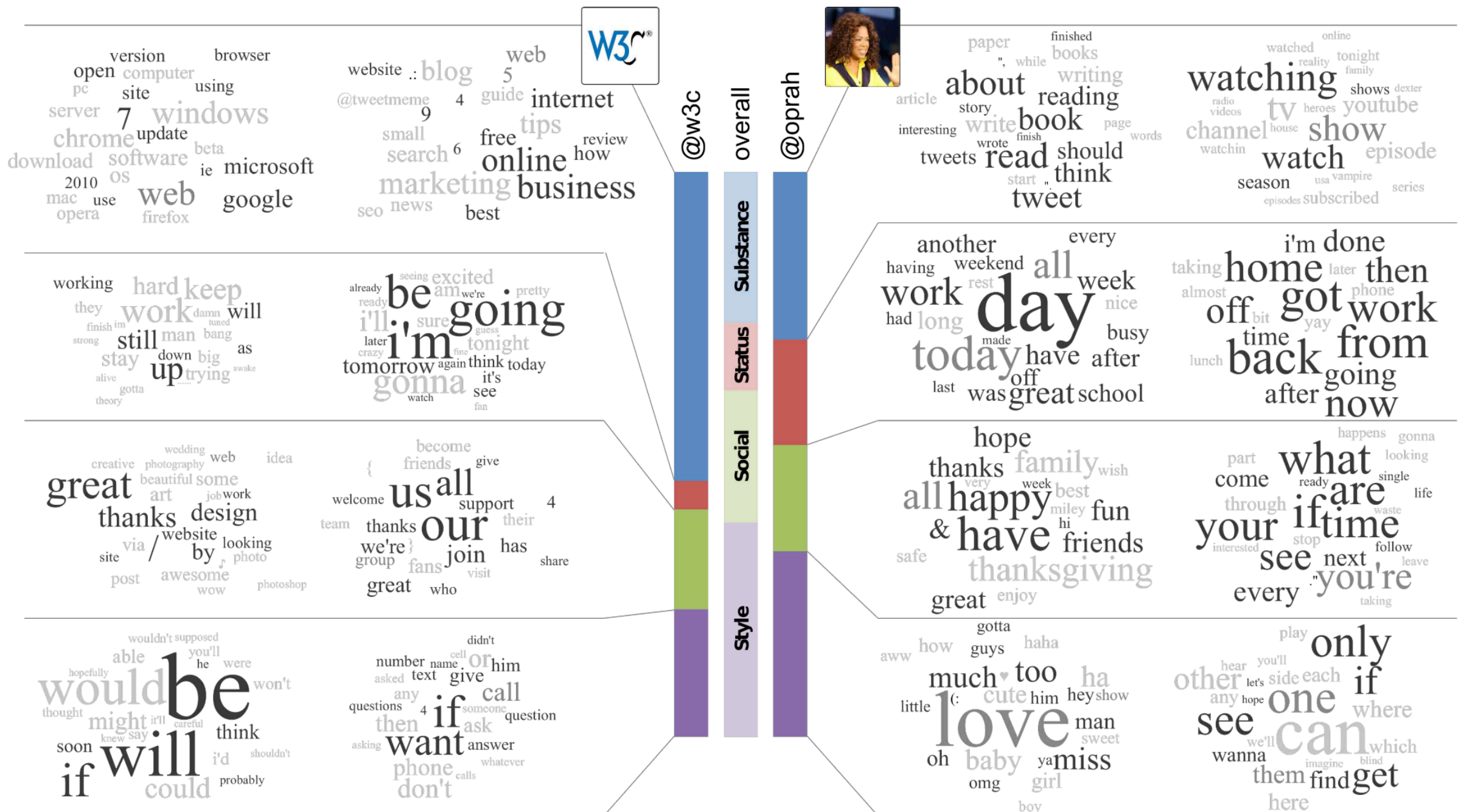
words tends to appear in same category of documents



Wikipedia Topics



Twitter Topics



New York Times Topics

music
band
songs
rock
album
jazz
pop
song
singer
night

book
life
novel
story
books
man
stories
love
children
family

art
museum
show
exhibition
artist
artists
paintings
painting
century
works

game
knicks
nets
points
team
season
play
games
night
coach

show
film
television
movie
series
says
life
man
character
know

theater
play
production
show
stage
street
broadway
director
musical
directed

clinton
bush
campaign
gore
political
republican
dole
presidential
senator
house

stock
market
percent
fund
investors
funds
companies
stocks
investment
trading

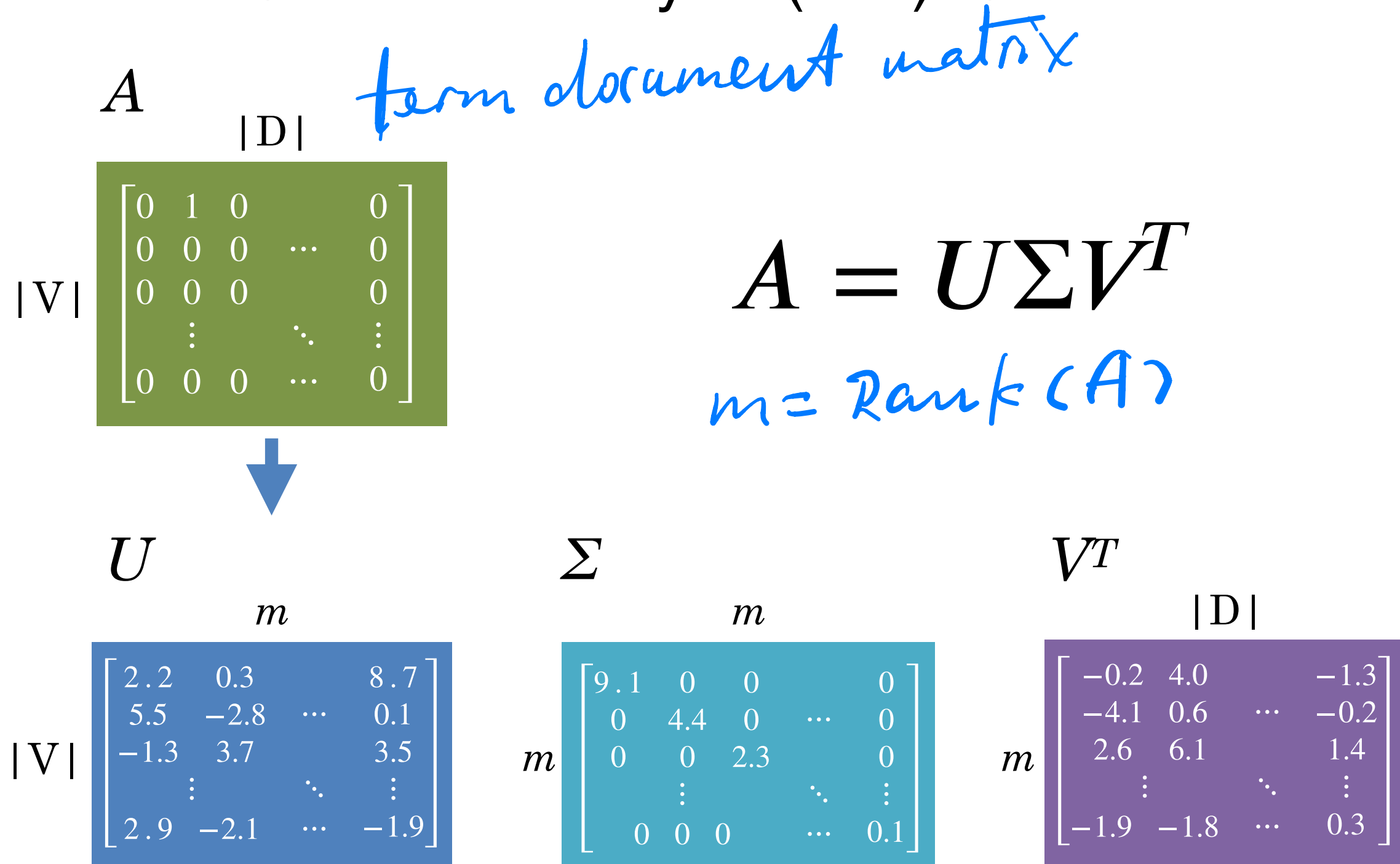
restaurant
sauce
menu
food
dishes
street
dining
dinner
chicken
served

budget
tax
governor
county
mayor
billion
taxes
plan
legislature
fiscal

A Brief History of Topic Models

What Is A Topic Model?

- Latent Semantic Analysis (L10): SVD+Truncate



$$A = U\Sigma V^T \quad \text{LSA: Truncate}$$

top k.

U
 m k

$|V|$

$$\begin{bmatrix} 2.2 & 0.3 & \dots & -2.4 & \dots & 8.7 \\ 5.5 & -2.8 & \dots & 1.1 & \dots & 0.1 \\ -1.3 & 3.7 & & 4.7 & & 3.5 \\ \vdots & & \ddots & \vdots & \ddots & \vdots \\ 2.9 & -2.1 & \dots & -3.3 & \dots & -1.9 \end{bmatrix}$$


$|V|$

$$\begin{bmatrix} 2.2 & 0.3 & \dots & -2.4 \\ 5.5 & -2.8 & \dots & 1.1 \\ -1.3 & 3.7 & & 4.7 \\ \vdots & & \ddots & \vdots \\ 2.9 & -2.1 & \dots & -3.3 \end{bmatrix}$$

Word Vector

pick top 10 words.

V^T

$|D|$

some words.

$$\begin{matrix} m \\ k \end{matrix} \begin{bmatrix} -0.2 & 4.0 & \dots & -1.3 \\ -4.1 & 0.6 & \dots & -0.2 \\ 2.6 & 6.1 & & 1.4 \\ \vdots & & \ddots & \vdots \\ -1.9 & -1.8 & \dots & 0.3 \end{bmatrix}$$


k

topic distribution

$$\begin{bmatrix} -0.2 & 4.0 & \dots & -1.3 \\ -4.1 & 0.6 & \dots & -0.2 \\ 2.6 & 6.1 & & 1.4 \\ \vdots & & \ddots & \vdots \end{bmatrix}$$

V_k^T

$|D|$

Topic Distribution

Document

Issues

- Positive and negative values in the U and V^T
- Difficult to interpret

U_k

topic

$|V|$

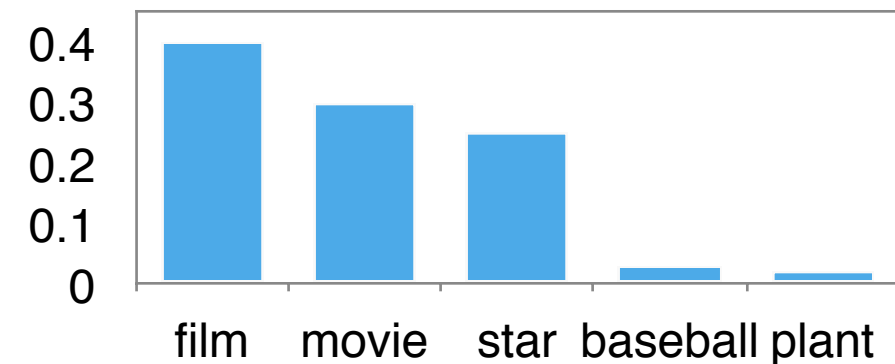
2.2	0.3	...	-2.4
5.5	-2.8	...	1.1
-1.3	3.7	...	4.7
⋮	⋮	⋮	⋮
2.9	-2.1	...	-3.3

← Topic

Probabilistic LSA

- Based on a probabilistic model

Topic 1

Word distribution
for a topic

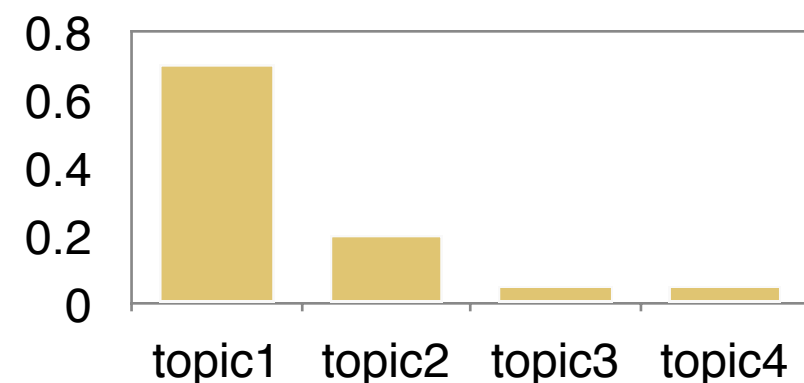
$$\begin{aligned}
 P(w, d) &= P(w | d)P(d) \\
 &= P(d) \sum_T P(w | t)P(t | d)
 \end{aligned}$$

Joint probability of a
word and a document

Number of topics

Topic distribution
for a document

Document 1



Issues

- No more negative values!
- PLSA can learn topics and topic distribution for documents in the train corpus
- But it is unable to infer topic distribution on **new documents**
- PLSA needs to be re-trained for new documents

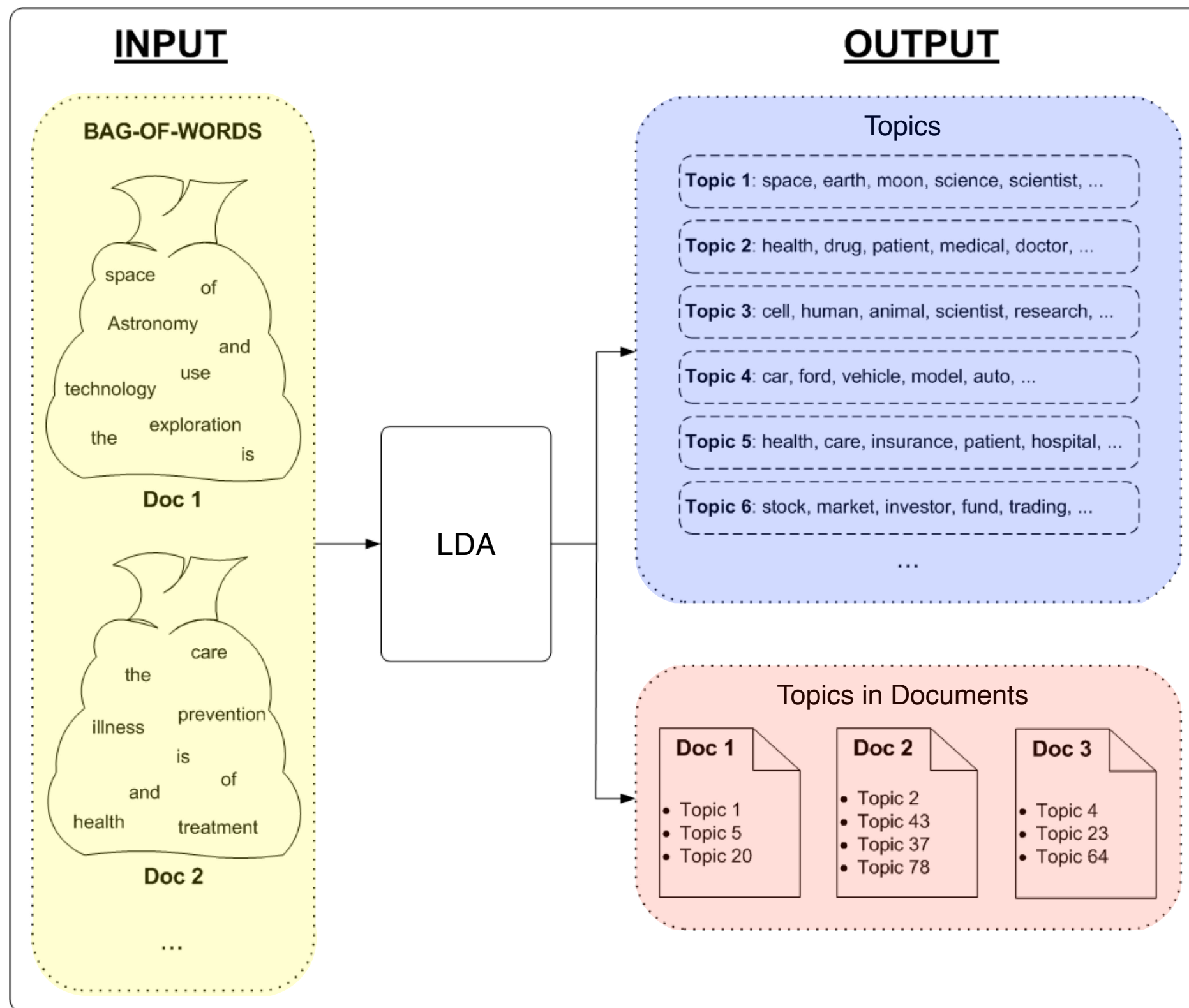
Latent Dirichlet Allocation

- Introduces a prior to the document-topic and topic-word distribution
- Fully generative: trained LDA model can infer topics on unseen documents!
- LDA is a Bayesian version of PLSA

Latent Dirichlet Allocation

LDA

- Core idea: assume each document contains **a mix of topics**
- But the topic structure is **hidden (latent)**
- LDA infers the topic structure given the observed words and documents
- LDA produces soft clusters of documents (based on topic overlap), rather than hard clusters

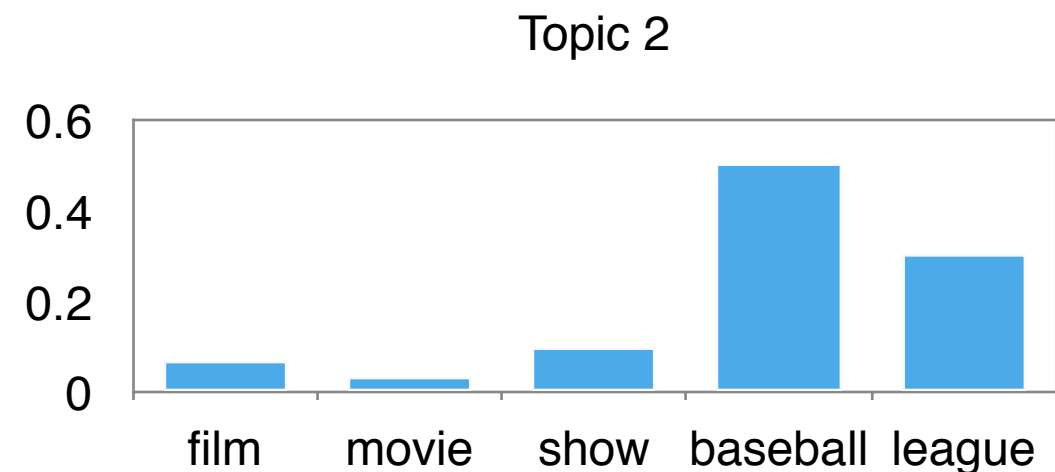
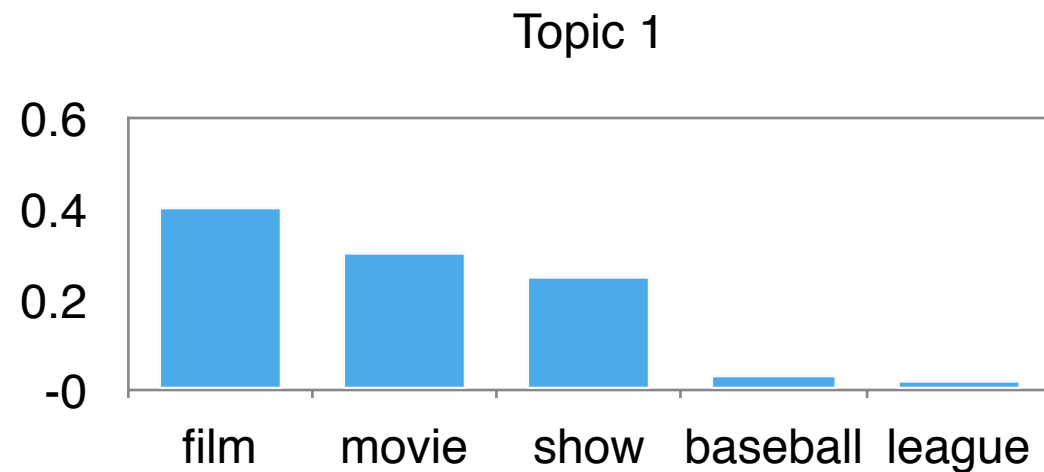


Input

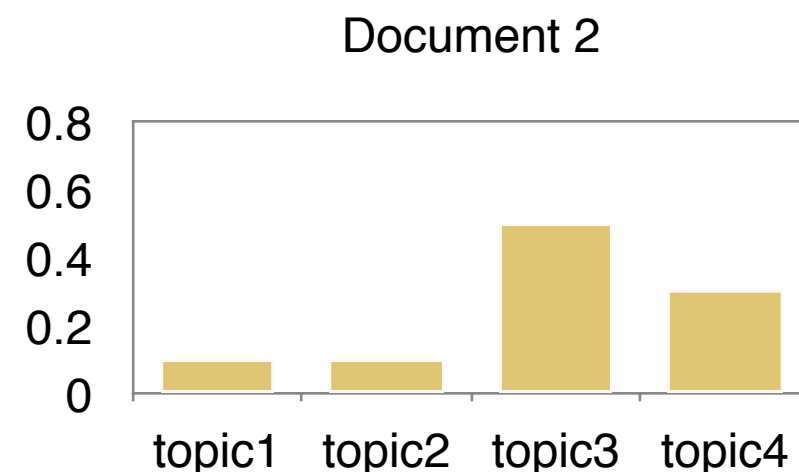
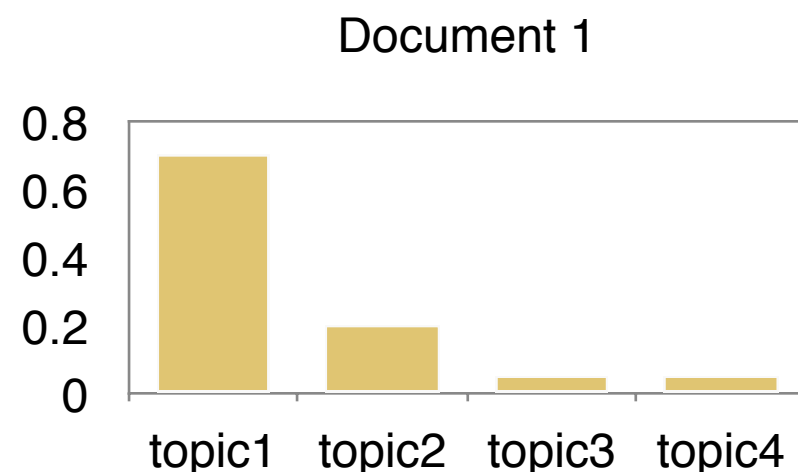
- A collection of documents
- Bag-of-words
- Good preprocessing practice:
 - ▶ Remove stopwords
 - ▶ Remove low and high frequency word types
 - ▶ Lemmatisation

Output

- **Topics:** multinomial distribution over words in each topic



- **Topics in documents:** multinomial distribution over topics in each document



Learning

- How do we learn the latent topics?
- Two main family of algorithms:
 - ▶ Variational methods
 - ▶ Sampling-based methods

Sampling Method (Gibbs)

1. Randomly assign topics to all tokens in documents

doc₁	mouse: t₁	cat: t₃	rat: t₂	chase: t₁	mouse: t₃
doc₂	scroll: t₁	mouse: t₃	scroll: t₃	scroll: t₂	click: t₂
doc₃	tonight: t₂	baseball: t₁	tv: t₂	exciting: t₁	

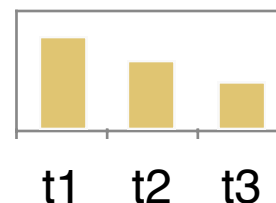
2. Collect topic-word and document-topic co-occurrence statistics based on the assignments

	mouse	cat	scroll	tv	...
t₁	1	0	1	0	
t₂	0	0	1	1	
t₃	2	1	1	0	

	t ₁	t ₂	t ₃
d₁	2	1	2
d₂	1	2	2
...			

3. Go through every word token in corpus and sample a new topic:

$$\triangleright P(t_i) \propto P(t_i | w)P(t_i | d)$$



Need to de-allocate the current topic assignment and update the co-occurrence matrices before sampling

4. Repeat until convergence

When Do We Stop?

- Train until convergence
- Convergence = model probability of training set becomes stable
- How to compute model probability?

$$\log P(w_1, w_2, \dots, w_m) = \log \sum_{j=0}^T P(w_1 | t_j) P(t_j | d_{w_1}) + \dots + \log \sum_{j=0}^T P(w_m | t_j) P(t_j | d_{w_m})$$

► $m = \text{\#word tokens}$

Based on the topic-word
co-occurrence matrix

Based on the document-topic
co-occurrence matrix

Infer Topics For New Documents

1. Randomly assign topics to all tokens in new/test documents

testdoc ₁	tiger: t_2	cow: t_1	cat: t_3	tiger: t_3	
testdoc ₂	football: t_2	live: t_2	men: t_2	fun: t_3	soccer: t_1
testdoc ₃	news: t_1	cnn: t_3	tonight: t_1		

2. Update document-topic matrix based on the assignments; but use the trained topic-word matrix

from trained topic model

→

	mouse	cat	scroll	tv	...
t_1	1	0	1	0	
t_2	0	0	1	1	
t_3	2	1	1	0	

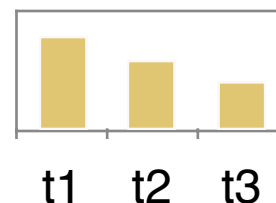
→

	t_1	t_2	t_3
td ₁	1	1	2
td ₂	1	3	1
...			

← new matrix

3. Go through every word in the test documents and sample topics:

$$\triangleright P(t_i) \propto P(t_i | w)P(t_i | d)$$



4. Repeat

Hyper-Parameters

- T : number of topic

attorney
death
charges
judge
authorities

A word cloud for a broad topic (Low T) with words in red. The most prominent words are 'attorney' and 'charges'. Other words include 'death', 'judge', and 'authorities'.

economic
economy
company
market
sales

A word cloud for a broad topic (Low T) with words in red. The most prominent words are 'economy' and 'market'. Other words include 'economic', 'company', and 'sales'.

Low T (<10): broad topics

prison judge
manning
attorney
classified
bradley

A word cloud for a specific topic (High T) with words in blue. The most prominent words are 'attorney' and 'bradley'. Other words include 'prison', 'judge', 'manning', and 'classified'.

media
oprah money
winfrey
franken
network

A word cloud for a specific topic (High T) with words in blue. The most prominent words are 'oprah' and 'winfrey'. Other words include 'media', 'money', 'franken', and 'network'.

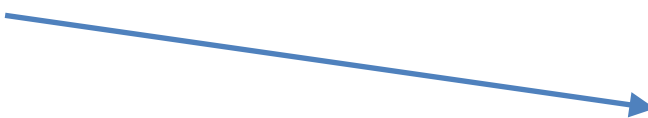
High T (100+): fine-grained, specific topics

Hyper-Parameters

- β : prior on the topic-word distribution
- α : prior on the document-topic distribution
- Analogous to k in add- k smoothing in N -gram LM
- Pseudo counts when computing:

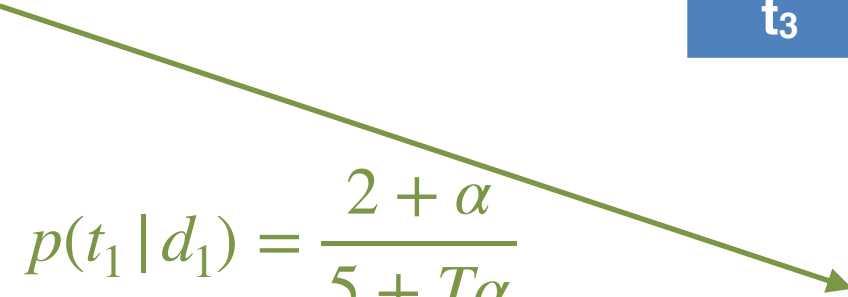
► $p(w | t): \beta$

► $p(t | d): \alpha$





	mouse	cat	scroll	tv	...
t ₁	1	0	1	0	
t ₂	0	0	1	1	
t ₃	2	1	1	0	

$$p(t_1 | d_1) = \frac{2 + \alpha}{5 + T\alpha}$$



	t ₁	t ₂	t ₃
d ₁	2	1	2
d ₂	1	2	2
...			

Hyper-Parameters

- High prior values \rightarrow flatter distribution 
- ▶ a very very large value would lead to a uniform distribution
- Low prior values \rightarrow peaky distribution 
- β : generally small (< 0.01)
 - ▶ Large vocabulary, but we want each topic to focus on specific themes
- α : generally larger (> 0.1)
 - ▶ Multiple topics within a document

Evaluation

How To Evaluate Topic Models?

- Unsupervised learning → no labels
- Intrinsic evaluation:
 - ▶ model logprob / perplexity on test documents

$$L = \prod_w \sum_t P(w | t) P(t | d_w)$$

$$\text{ppl} = \exp \frac{-\log L}{m} \leftarrow \text{total number of word tokens in test documents}$$

Issues with Perplexity

- More topics = better (lower) perplexity
- Smaller vocabulary = better perplexity
 - ▶ Perplexity not comparable for different corpora, or different tokenisation/preprocessing methods
- Does not correlate with human perception of topic quality
- Extrinsic evaluation the way to go:
 - ▶ Evaluate topic models based on downstream task

Topic Coherence

- A better intrinsic evaluation method
- Measure how **coherent** the generated topics

food,
farmers,
rice,
farm,
agriculture

simply
give unionist
choice
count
i.e.

- A good topic model is one that generates more coherent topics

Word Intrusion

- Idea: inject one random word to a topic
- {farmers, farm, food, rice, agriculture}
↓
{farmers, farm, food, rice, **cat**, agriculture}
- Ask users to guess which is the **intruder word**
- Correct guess → topic is coherent
- Try guess the intruder word in:
 - ▶ {choice, count, village, i.e., simply, unionist}
- Manual effort; does not scale

Estimate Coherence Automatically?

- PMI!
- Compute pairwise PMI of top- N words in a topic

$$\text{PMI}(t) = \sum_{j=2}^N \sum_{i=1}^{j-1} \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}$$

- Given topic: {farmers, farm, food, rice, agriculture}
- Sum logPMI for all word pairs in the topic:
 - ▶ $\log\text{PMI}(\text{farmers, farm}) + \log\text{PMI}(\text{farmers, food})$
 $+ \dots + \log\text{PMI}(\text{rice, agriculture})$

Variants

- Normalised PMI

$$\text{NPMI}(t) = \sum_{j=2}^N \sum_{i=1}^{j-1} \frac{\log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)}$$

- Conditional probability

$$\text{LCP}(t) = \sum_{j=2}^N \sum_{i=1}^{j-1} \log \frac{P(w_i, w_j)}{P(w_i)}$$

- Good correlation with human perception of topic coherence
- Better correlation if we use a **different corpus** to estimate PMI

i.e. don't use corpus that we run the topic model on

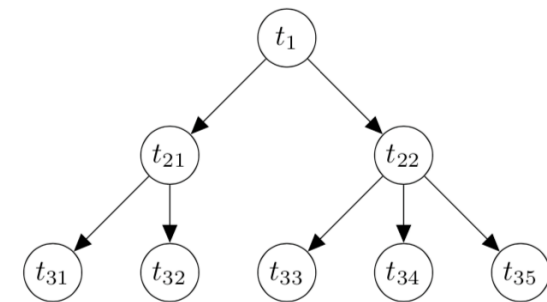
PMI Examples

Topic	PMI	NPMI
cell hormone insulin muscle receptor	0.61	0.59
electron laser magnetic voltage wavelength	0.54	0.52
magnetic neutrino particle quantum universe	0.55	0.55
album band music release song	0.37	0.56
college education school student university	0.38	0.57
city county district population town	0.34	0.52

Improvements

Topic Model Variants

- Use phrases or n-grams instead of words
- Learn hierarchical topics
- Non-parametric models
 - ▶ #topics automatically learned
- Supervised models
 - ▶ Takes into account document labels



Topic Labelling

vmware server virtual oracle update → virtualisation

church archway building window gothic → church architecture

investigation fbi official department federal → criminal investigation

rate population prevalence study incidence → mortality rate

- Use wikipedia article titles as labels
- Measure distance between a label and topic words based on document embeddings and word embeddings

A Final Word

- Topic model: an unsupervised model for learning latent concepts in a document collection
- LDA: a popular topic model
 - ▶ Learning
 - ▶ Hyper-parameters
- How to evaluate topic models?
 - ▶ Topic coherence