Natural Language Processing

Dr. Karen Mazidi

Part Four: Documents

Topics

- topic modeling

Quizzes

- Q: IE and more

Homework

- no homework on this material

Enron example

- 5 topics from a 25-topic model fit on Enron e-mails and the 5 most probable words from each topic

Topic	Terms
3	trading financial trade product price
6	gas capacity deal pipeline contract
9	state california davis power utilities
14	ferc issue order party case
22	group meeting team process plan

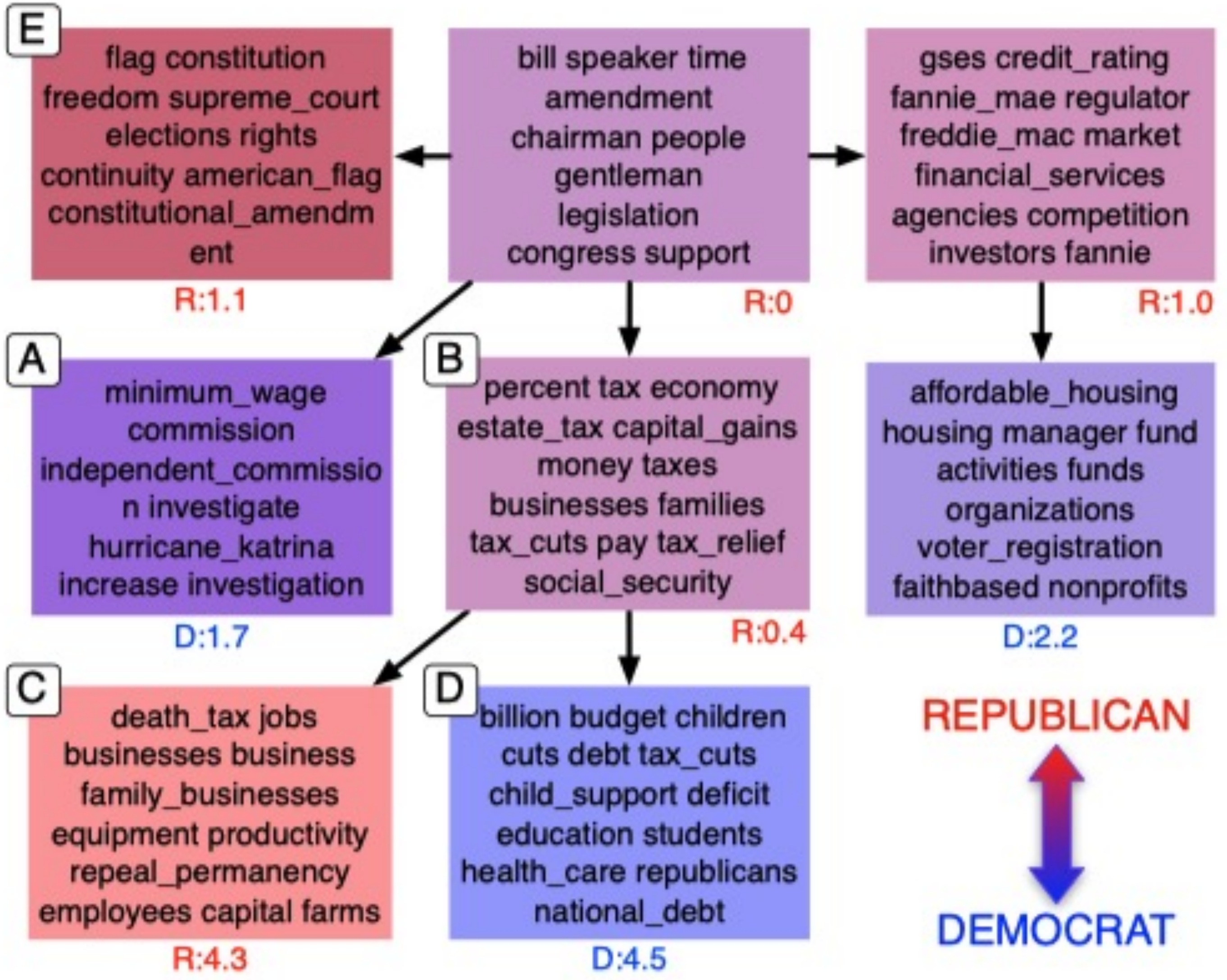
See Kaggle notebook:
https://www.kaggle.com/jesbin/topic-modeling-enron-email-dataset/notebook

Enron example

- Topic modeling included a topic with word ‘California’ even though this document did not contain that word
- The doc references SDG&E, a California energy company
- No domain expert needed!

> Yesterday, SDG&E filed a motion for adoption of an electric procurement cost recovery mechanism and for an order shortening time for parties to file comments on the mechanism. The attached email from SDG&E contains the motion, an executive summary, and a detailed summary of their proposals and recommendations governing procurement of the net short energy requirements for SDG&E’s customers. The utility requests a 15-day comment period, which means comments would have to be filed by September 10 (September 8 is a Saturday). Reply comments would be filed 10 days later.

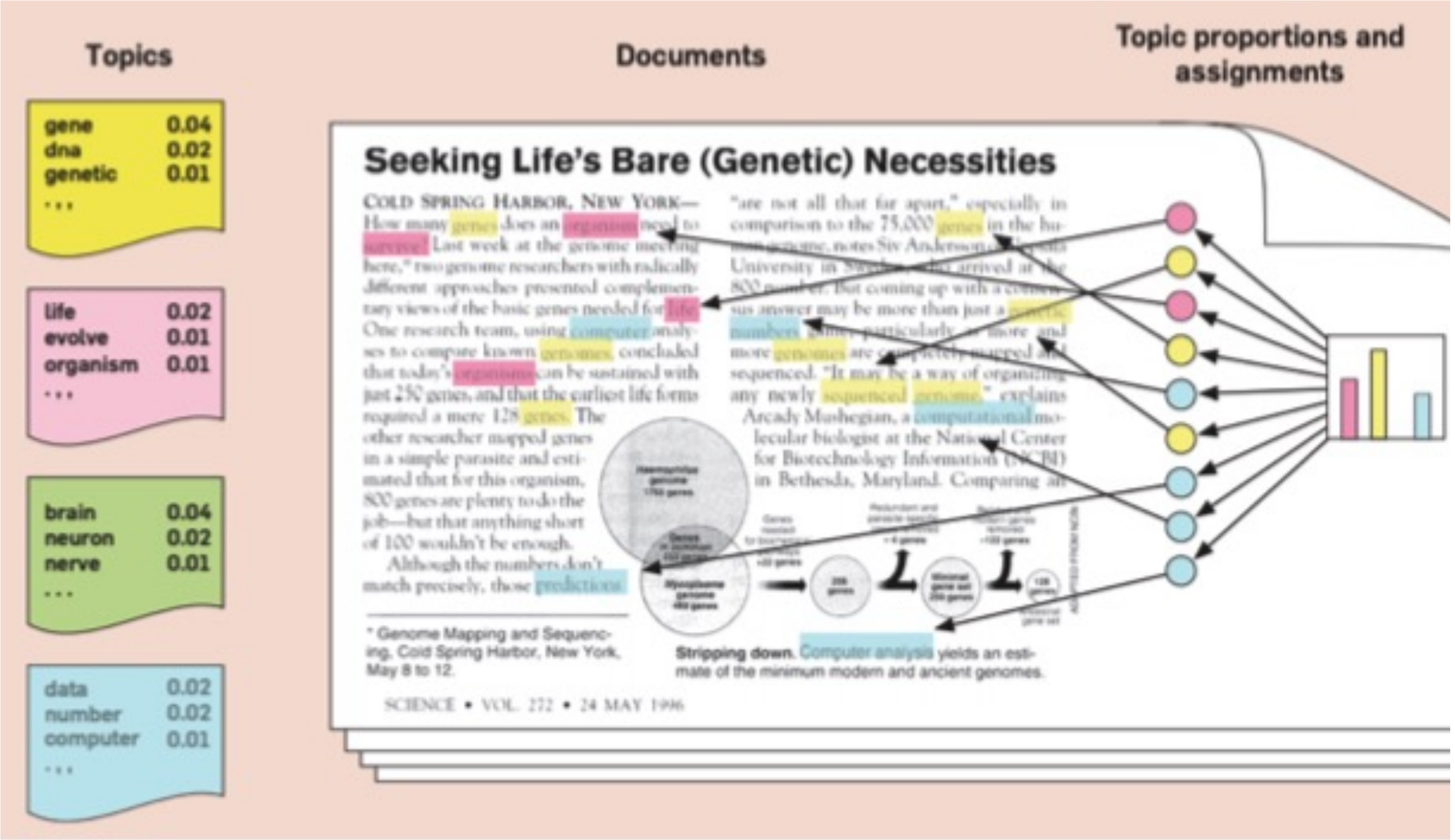
Congressional floor topics example



Topic modeling

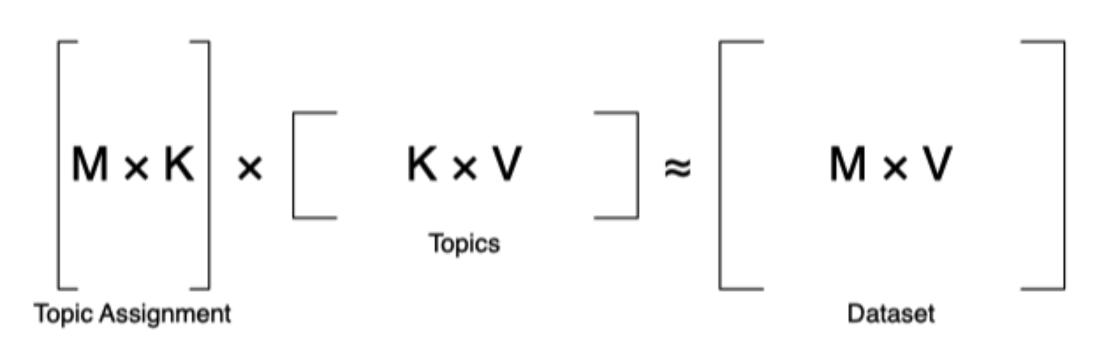
- Topic modeling defines a <u>topic</u> as a set of words
- A topic is a multinomial distribution over words
- Topic modeling defines a <u>document</u> as a mixture of topics
- These two are discovered simultaneously:
  - Topics in the corpus
  - Which topics are in which documents

Big picture example



- 4 topics – each of which is a set of words
- The document is a mixture of these topics

Big picture



- M documents
- K topics
- V vocabulary
- KxV connects topics to a jumbled ‘bag of words’
- MxK links topics to individual documents

Distributions of words

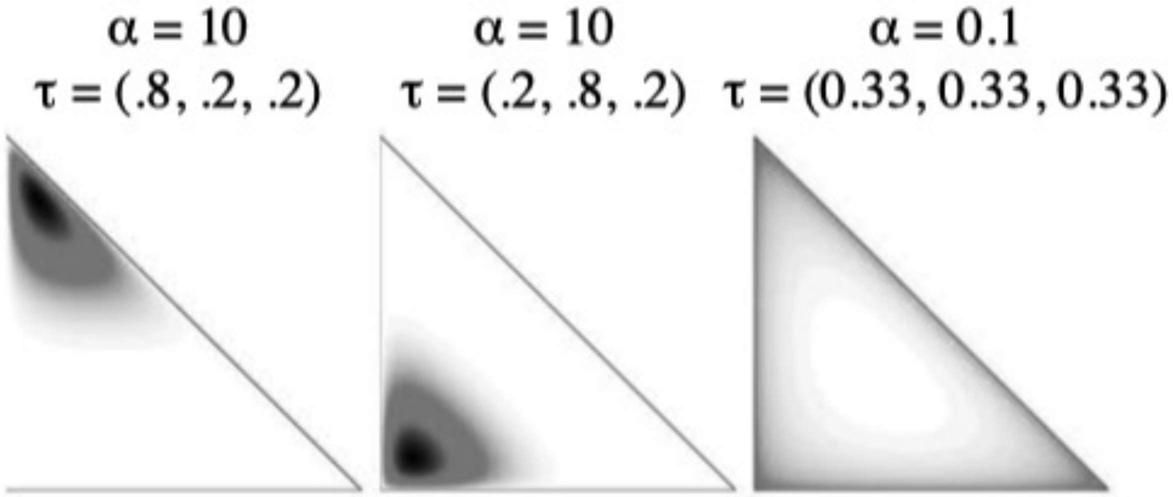
- Common distributions used in topic modeling

Distribution	Density	Example Parameters	Example Draws
Gaussian	$\frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$\mu = 2, \sigma^2 = 1.1$	$x = 2.21$
Discrete	$\prod_i \phi_i^{\mathbb{1}[w=i]}$	$\phi = \begin{bmatrix} 0.1 \\ 0.6 \\ 0.3 \end{bmatrix}$	$w = 2$
Dirichlet	$\frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1}$	$\alpha = \begin{bmatrix} 1.1 \\ 0.1 \\ 0.1 \end{bmatrix}$	$\theta = \begin{bmatrix} 0.8 \\ 0.15 \\ 0.05 \end{bmatrix}$

- Documents are combinations of discrete symbols – tokens
- Topics are discrete (multinomial) distributions over words
- Some words have higher probability than others

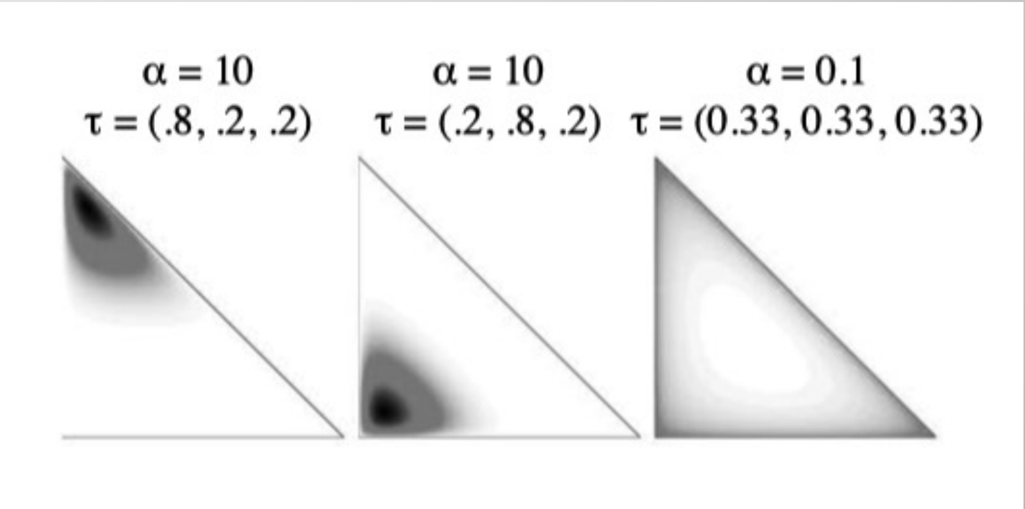
Dirichlet distributions

- Produce probability vectors that can be used as the parameters of discrete distributions
- Like the Gaussian, Dirichlet has parameters that are analogous to the mean and variance
- The base measure, tau, is the expected value of the Dirichlet distribution
- The concentration parameter, alpha, controls how far away individual samples are from the base



Dirichlet distributions

- The base measure, tau, is the expected value of the Dirichlet distribution
- The concentration parameter, alpha, controls how far away individual samples are from the base
  - If alpha is large, samples will be close to tau
  - If alpha is small, samples will become sparse (only a few values have high probability and others are small)



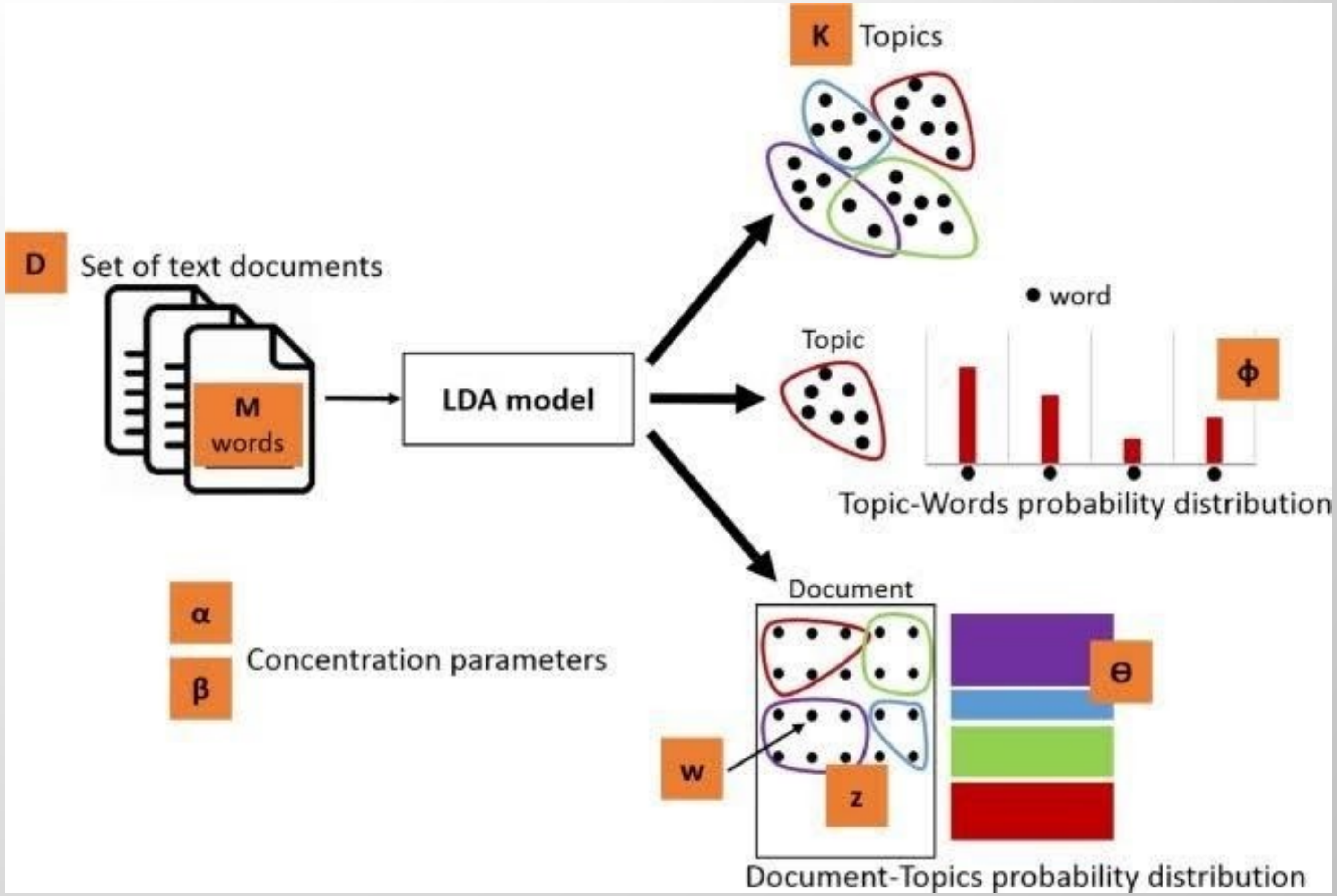
LDA

- LDA latent Dirichlet allocation is a common technique
- LDA is a generative probabilistic model with both observed and hidden variables combined in a joint probability distribution
- LDA speculates on how the documents could have been created from the distributions of topics and words

Generating topics

- User specifies K as the number of topics
- Each of the K topics is drawn from a Dirichlet distribution with a uniform base distribution

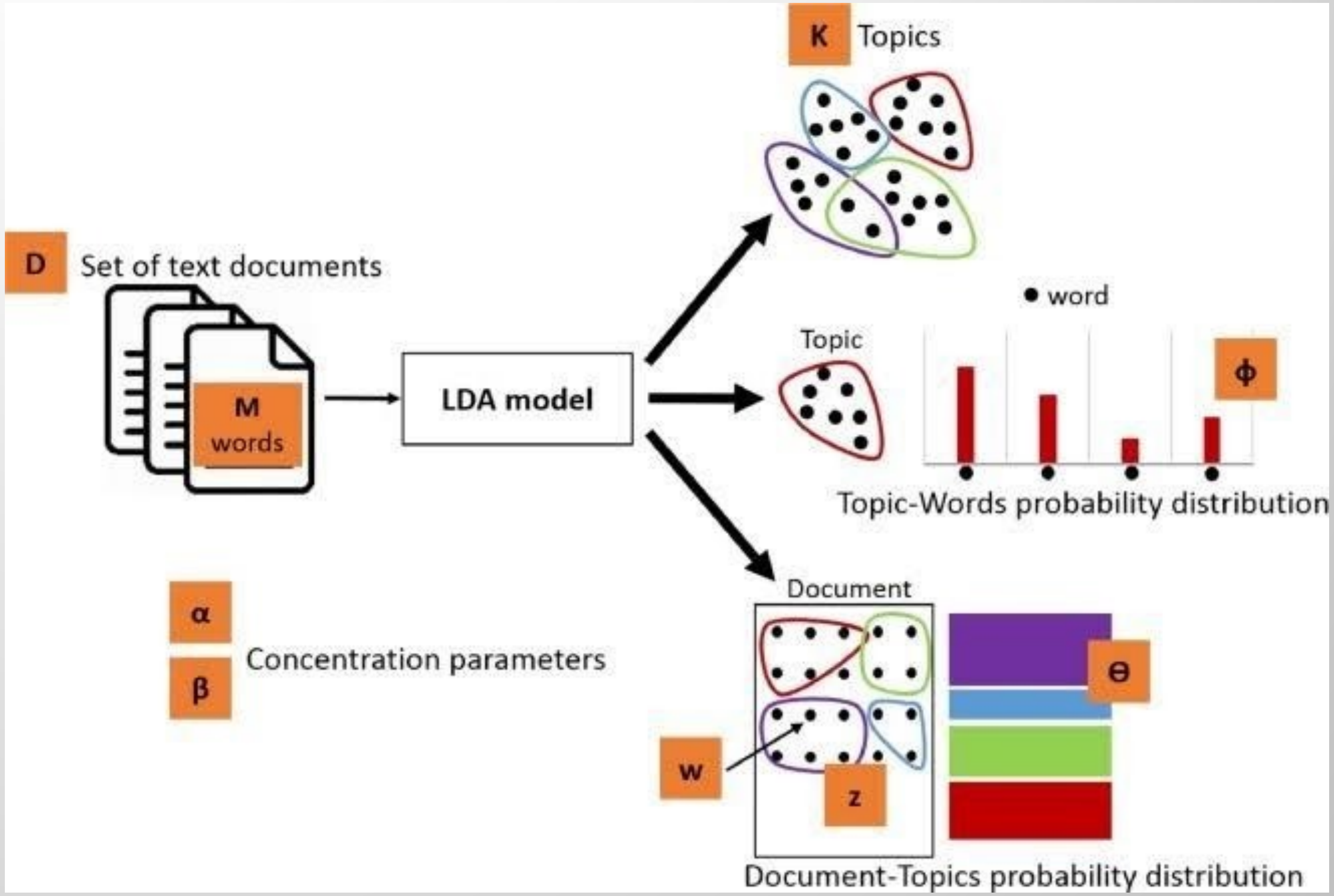
$$\lambda\colon\ \phi_k \sim \mathrm{Dir}(\lambda \boldsymbol{u})$$



Document allocations

- Each document is a distribution over topics
- The concentration parameter, alpha, ensures that each document is only about a few topics

$$\theta_d \sim \mathrm{Dir}(\alpha \boldsymbol{u})$$



Words in context

- For each word n in document d, choose a topic assignment z

$$z_{d,n} \sim \mathrm{Discrete}(\theta_d)$$

- The assignment of a word to a topic is a random variable
- A word can be assigned to different topics in the same document

The math

- LDA is a generative probabilistic model
- The posterior is the conditional distribution of the hidden variables (topics), given the observed words

$$p(\beta_{1:k}, \theta_{1:D}, z_{1:D} | w_{1:D}) = \frac{p(\beta_{1:k}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})} \tag{15.1}$$

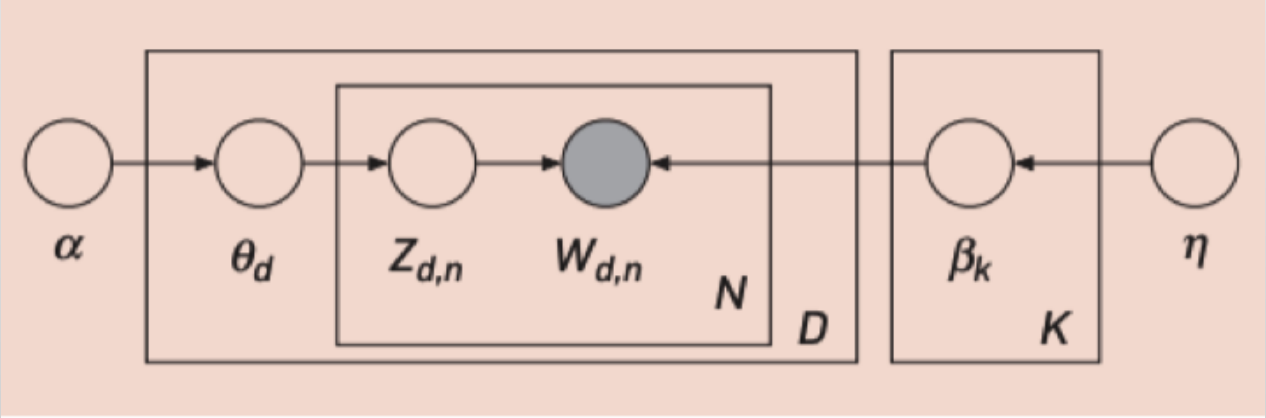
where betas are topics 1:K, thetas are topic proportions for each topic in each document, z represents topic assignment for a given word in a document, and the observed words are w.

Computation

- The distribution cannot be computed directly, so sampling techniques are used
- Gibbs sampling is a Monte Carlo Markov Chain (MCMC) technique that starts with the variables at random values
- Iteratively, holding all variables constant but one:
  - Repeatedly sample the data to get an estimate of that variable
- Repeat the process for each variable until convergence

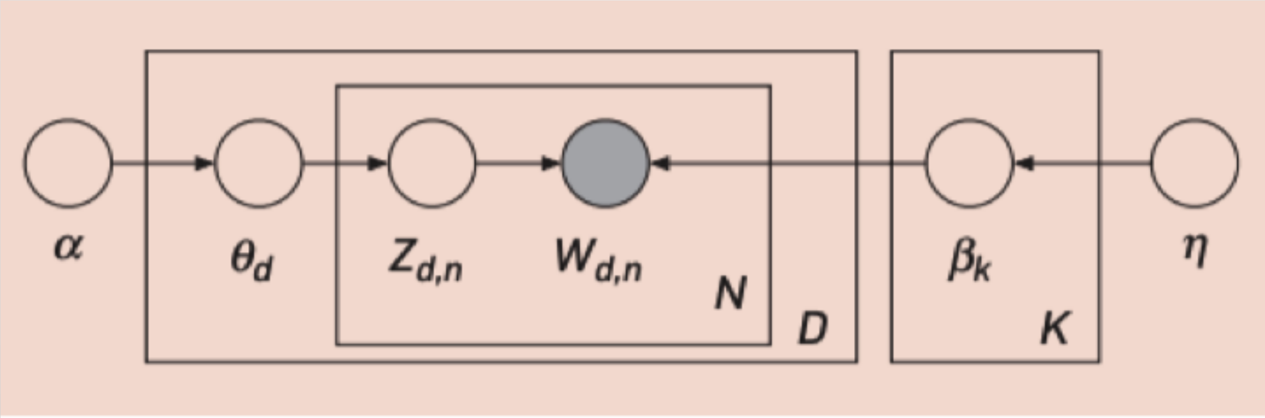
Graphical model

- Shaded node: observed data, the words
- Unshaded nodes are hidden variables
- A ‘plate’ indicates replication



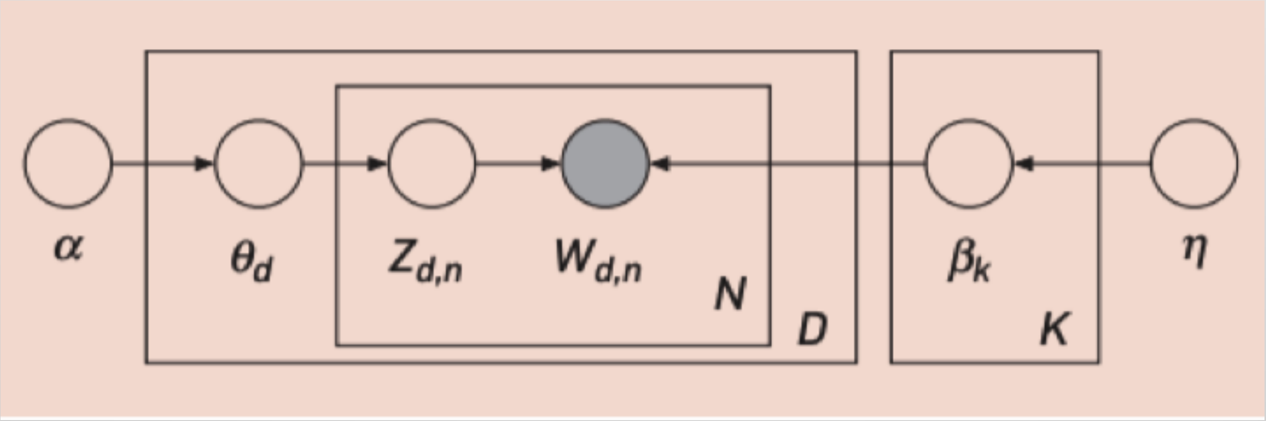
Graphical model

- Plate N represents collections of words in docs
- Plate D represents the documents in the corpus
- Plat K represents the topics



Graphical model

- Parameter alpha represents document-topic density
  - The higher the alpha, the more topics per document
- Parameter eta represents topic-word density
  - The higher the eta, the more words per topic



LDA tips

- Preprocessing: lower case, remove non-alpha and stop words, a custom stop word list may help
- Lemmatizing helps also
- Try using just nouns and adjectives
- Number of topics, alpha, eta are chosen beforehand
  - Try default settings, then experiment
- LDA is hungry for data. A small corpus won't get good results, as we shall see.
- LDA needs to see words co-occurring in many instances in order to learn that they are related

LDA evaluation

- Visual inspection of the topics can be informative
- If the same word appears in many topics, then k is probably too large
- The metric coherence is often used
- Coherence measures how much words in the topic tend to occur together in documents
- Coherence ranges from 0 to 1, below .5 is not good

LDA or LSI

- LSI (latent semantic indexing) is sometimes used instead of LDA
- LSI is a dimensionality reduction technique, reducing similar words to indexes
- The dimensionality reduction is called SVD Singular Value Decomposition
- LSI is generally faster to train
- LDA often gets better results
- Both techniques use a bag-of-words input matrix

Labeling

- Several approaches to labeling topics have been explored:
  - Internal labeling: extract prominent phrases from the topic and compare how consistent it's context is with the topic distribution
  - Supervised approach: trained from labeled data
  - Using knowledge bases: a topic's words should be consistent with the label's children in an ontology

Gensim Example

- 4 small texts are used to demonstrate the code
- Texts were lower cased, tokenized, stopwords and non-alpha tokens were removed
- Each doc is a list of tokens
- Each token is mapped to an id number in a dict

Open-source Python library: https://radimrehurek.com/gensim/

Gensim

- Using LDA

Code 15.5.1 — LDA. Building the topic model

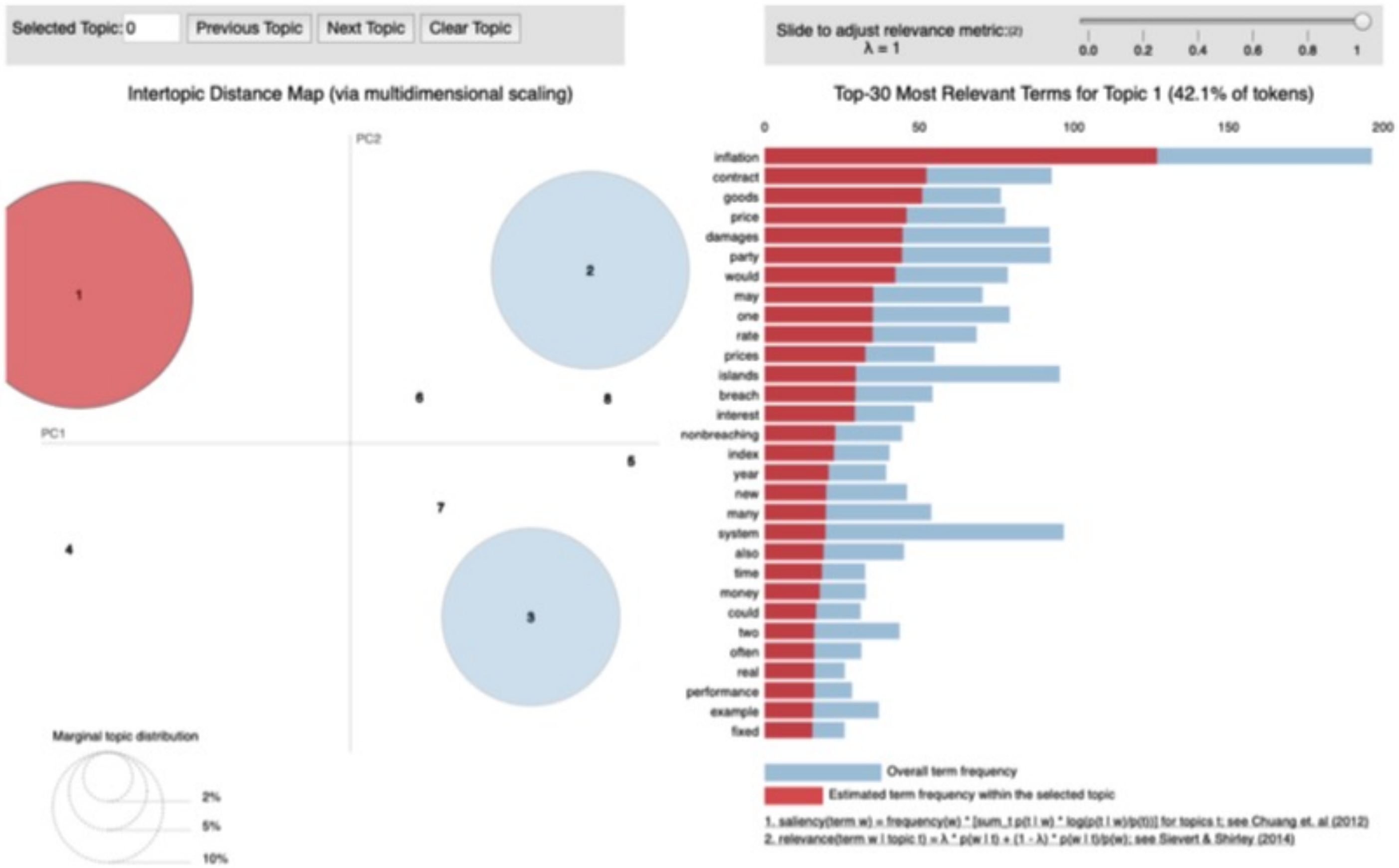
```
from gensim import models, corpora
NUM_TOPICS = 8

# the dictionary maps words to id numbers
dictionary = corpora.Dictionary(preprocessed_docs)

lda_model = models.LdaModel(corpus=corpus,
         num_topics=NUM_TOPICS, id2word=dictionary)
```

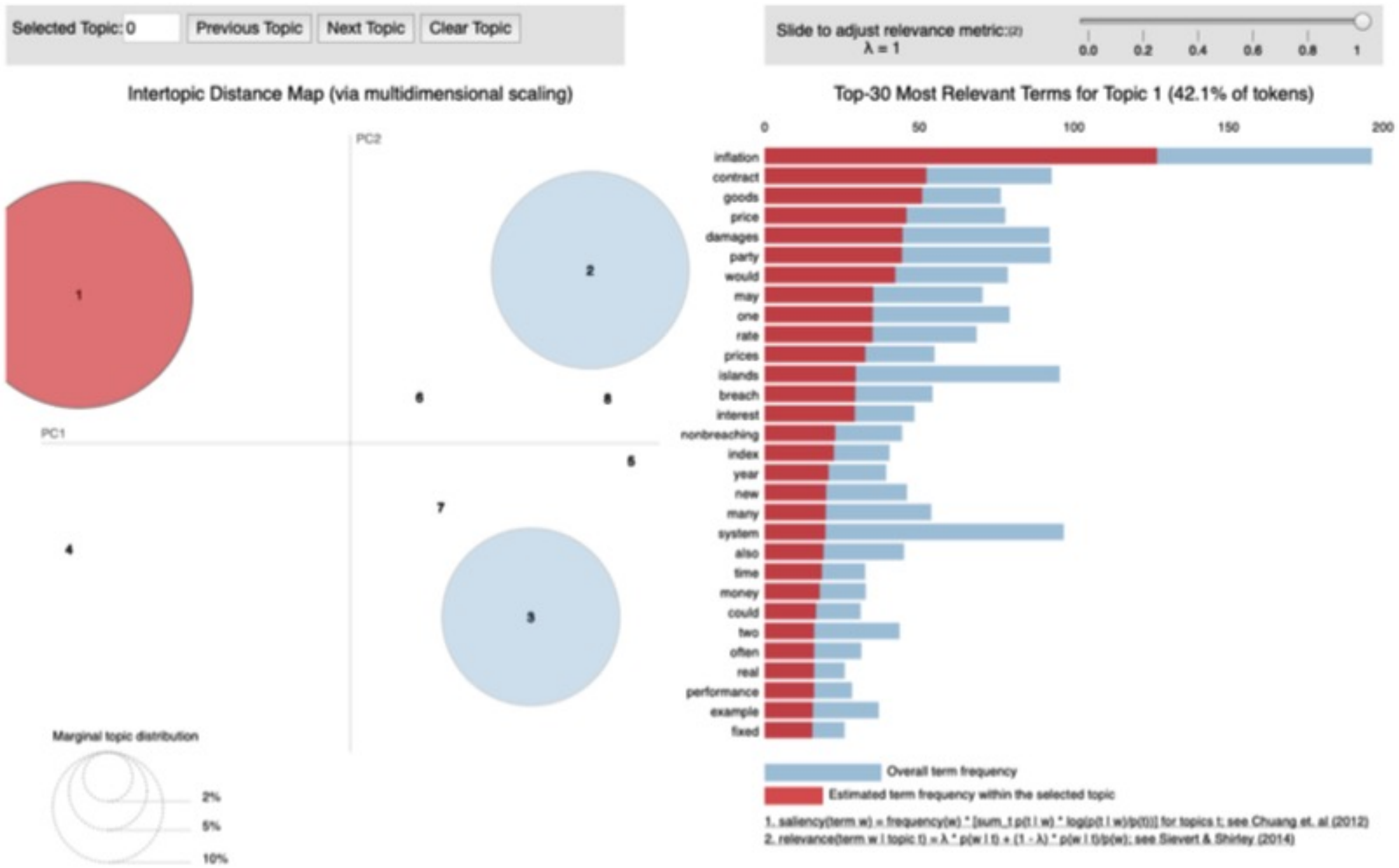
Visualization

- Library pyLDAvis
- Left: numbers represent topics, the bigger the balloon the more important the topic
- Balloons should be well-separated
- Overlapping balloons indicate too many topics



Visualization

- Hover over a balloon, it changes to red, the important words in that topic appear on the right
- Sliding relevance metric upper right



Essential points to note

- Topic modeling has received a lot of attention in recent years
- Pros: unsupervised way of learning about a corpus
- Cons: hard to tell what was learned, often little correspondence to human evaluation of topics

To Do

- Quiz on IE and more

Next class

Discuss chatbot project