Subsection 2

Topic modeling (very very very briefly)

Topics

Given a corpus of documents, what do they talk about?

- talks about $\rightarrow$ *topic*

Probabilistic model

Assume stochastic document building process:

- there exist k topics
- a topic is a distribution over words
- a topic is assigned to the document according to a known probability (a document may exhibit multiple topics)
- a word in a document is drawn according to topic and document-topic assignment

Words order does not matter!

Probabilistic model

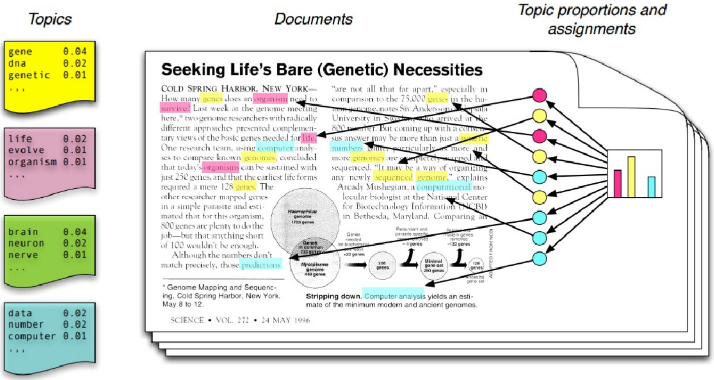

Image from `https://www.cs.princeton.edu/~blei/topicmodeling.html`

Probabilistic graphical model

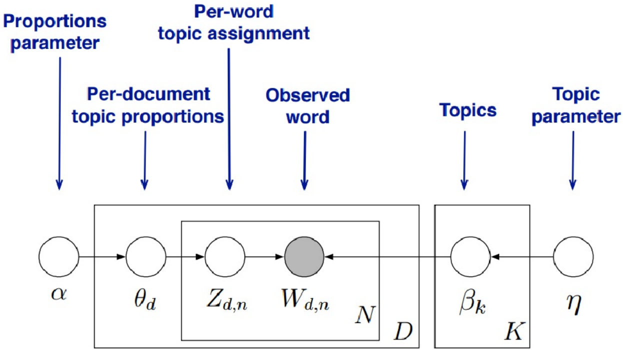

Image from https://www.cs.princeton.edu/~blei/topicmodeling.html

- nodes are random variables
- edges are dependencies
- shaded nodes are observed
- boxes are repeated variables

Latent Dirichlet allocation (LDA)

A way for inferring distributions/assignments from observed values!
(*Posterior inference*)

Given K (parameter),

- for each topic, compute words distribution
- for each document, compute topic "distribution"

- **Latent** refers to the unknown random variables
- **Dirichlet** is the distribution assumed for topics and words
- **Allocation** of words to topics and topics to documents

LDA internals

(Just a coarse overview)

While inferring posterior, try to (both):

- associate each document with as few topics as possible
- associate each topic with as few words as possible

Conflicting goals, which results in finding (and putting in the same topics) words which often co-occur

LDA output

- For each document of the corpus, a vector in $[0,1]^K$ where i-th value is "how much the document exhibits i-th topic"
  - reasonable values for the number of topics K is some tens (10–50) (**Q:** how to choose the right value for a problem?)
- For each topic, a vector $[0,1]^V$ where the i-th value is "how much the i-th word (on V words) is associated with the topic"
  - how to visualize/understand a topic? Select its most likely words

Visualize topics

Image from `https://www.cs.princeton.edu/~blei/topicmodeling.html`

LDA as a building block

- corpus visualization
- document similarities
- ...
- document $\rightarrow \mathbb{R}^K$

LDA: document $\to \mathbb{R}^K$

How to apply to new data?

- assume everything is known (i.e., already computed on the corpus)
- just infer the posterior of topic assignment for the new document

Lab: Topics in poetry

KEY POINTS:
1) HOW TO CHOOSE K
2) PREPROCESSING ?
3) HOW TO PRESENT OUTPUT

What's modern and renaissance poetry about?

Data: https://www.kaggle.com/ultrajack/modern-renaissance-poetry

In R:

- package `topicmodels`
- functions `LDA(train.data, k)`, `posterior(lda.model, test.data)`