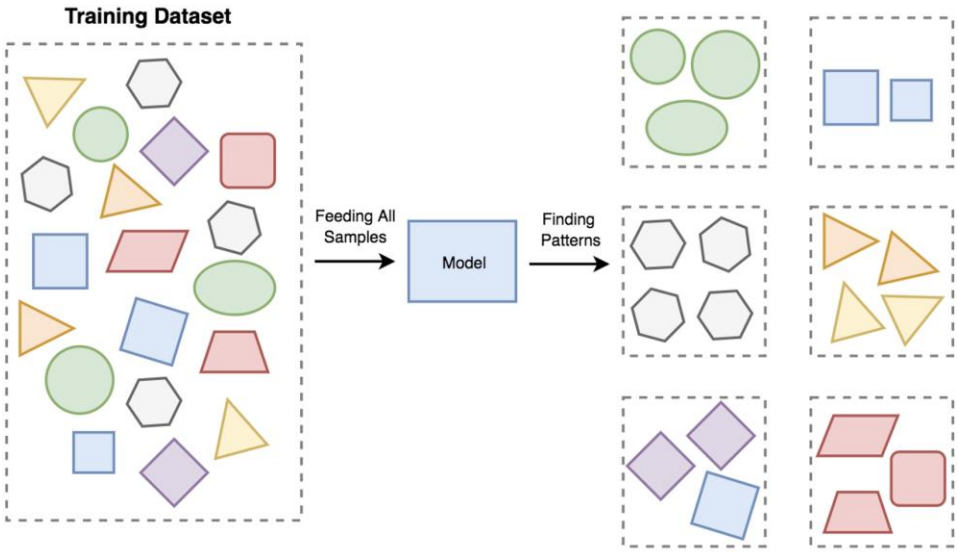
Topic Modeling

Ayoub Bagheri

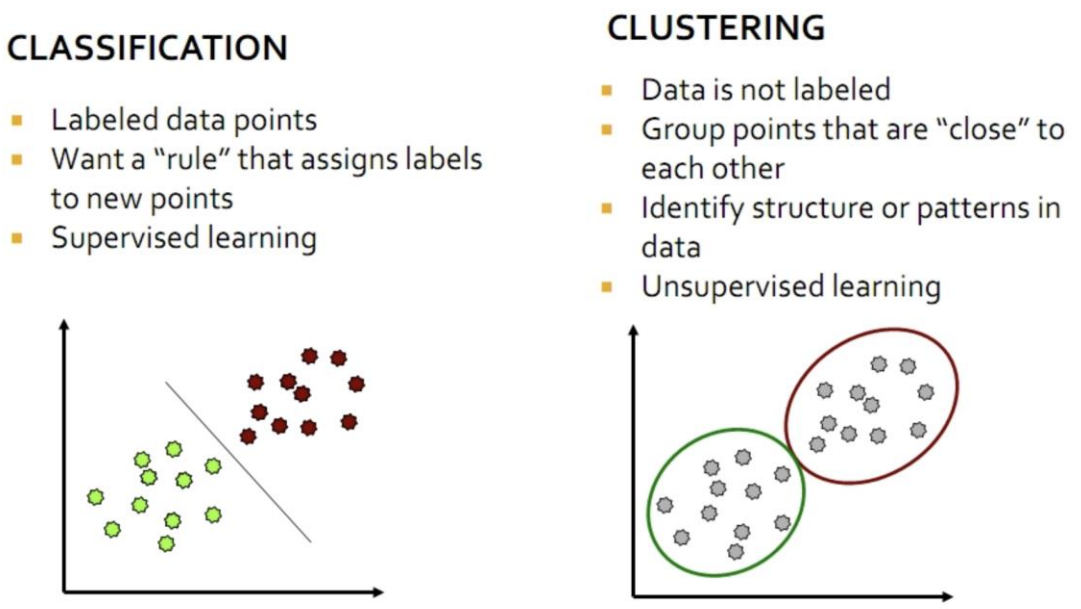
Lecture plan

1. Text clustering
2. Probabilistic topic modeling
3. Latent Dirichlet allocation

Unsupervised learning

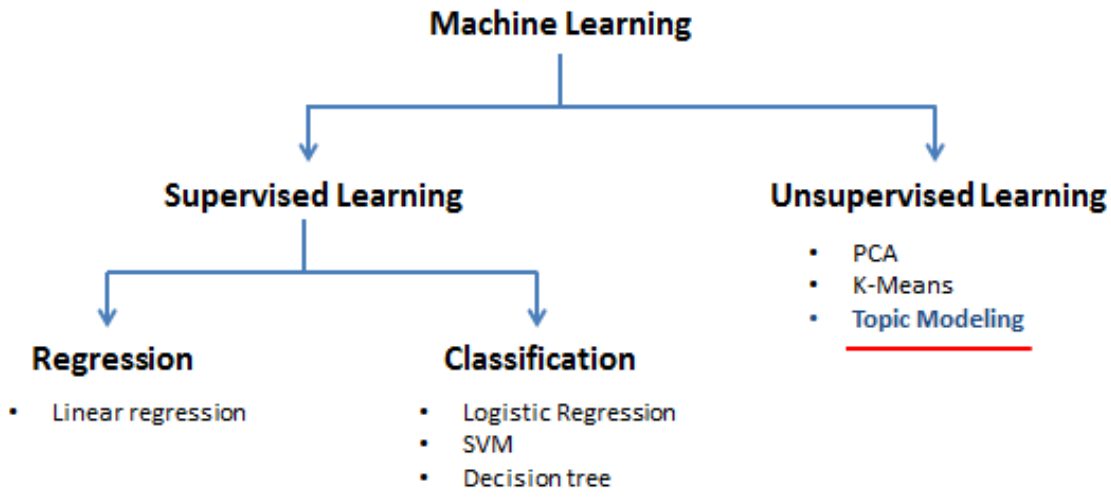


Clustering versus classification



Topic Modeling

Topic modeling



https://thinkinfi.com/

Topic models

- Three concepts: words, topics, and documents

- Documents are a collection of words and have a probability distribution over topics
- Topics have a probability distribution over words
- Model:
  - Topics made up of words used to generate documents

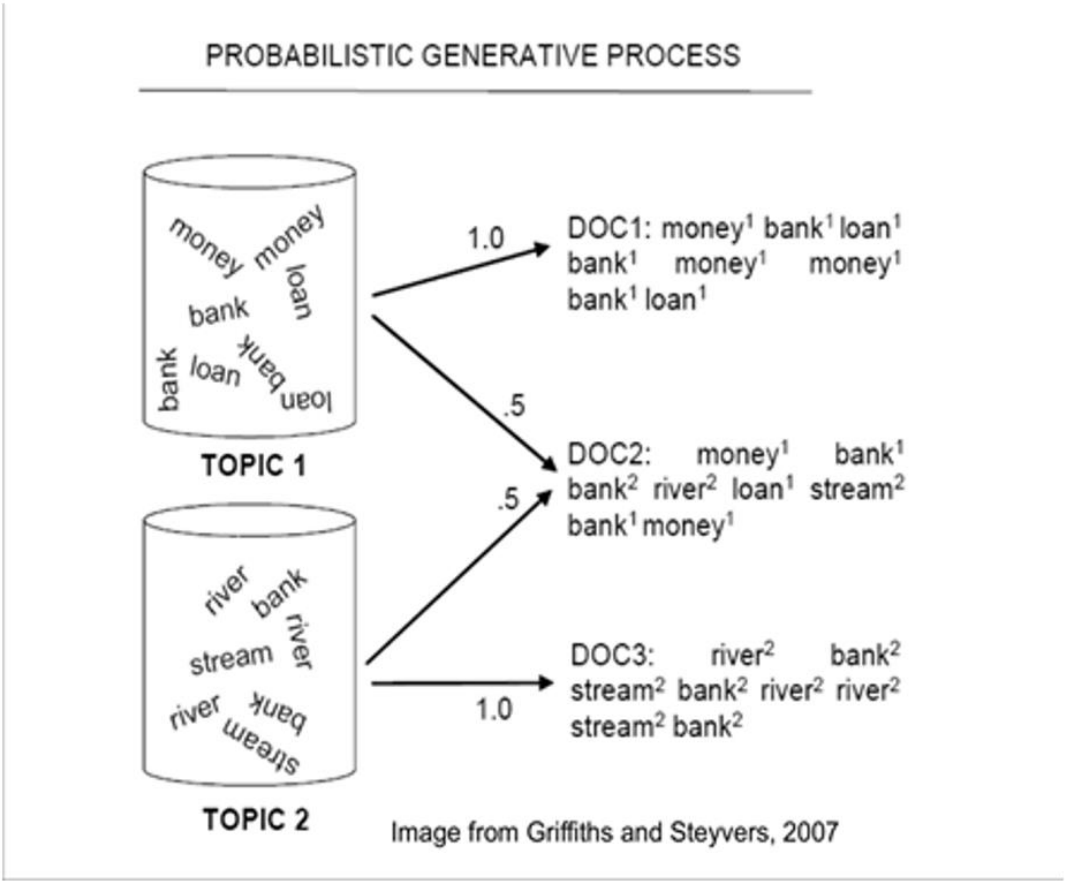
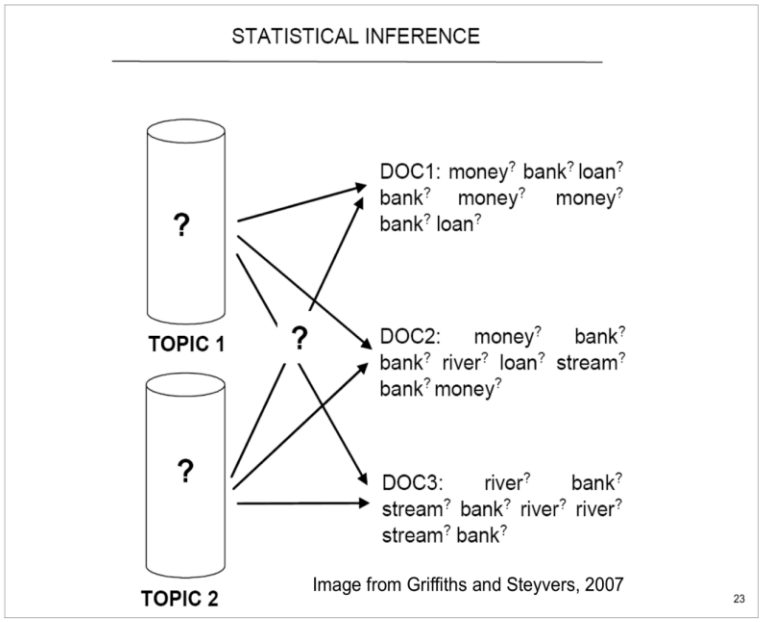
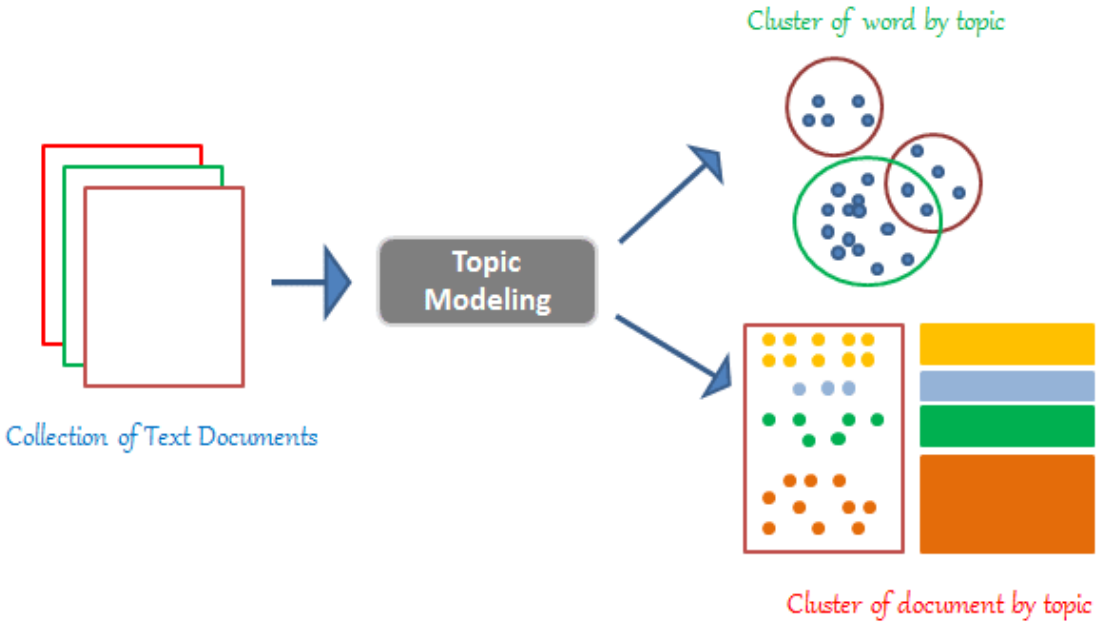


Image from Griffiths and Steyvers, 2007

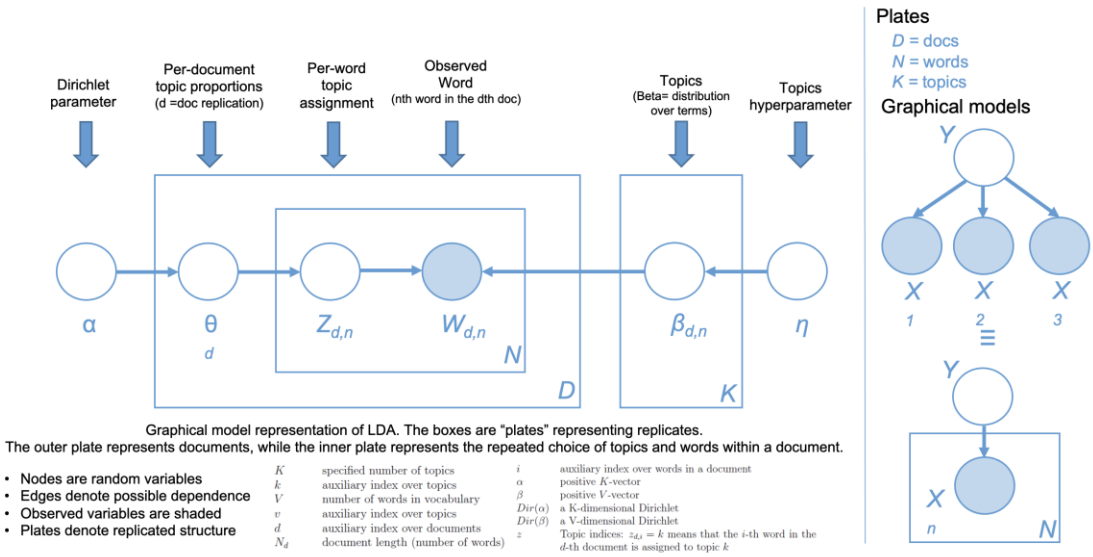
Topic models | Reality: Documents observed, infer topics



Topic models



LDA graphical model

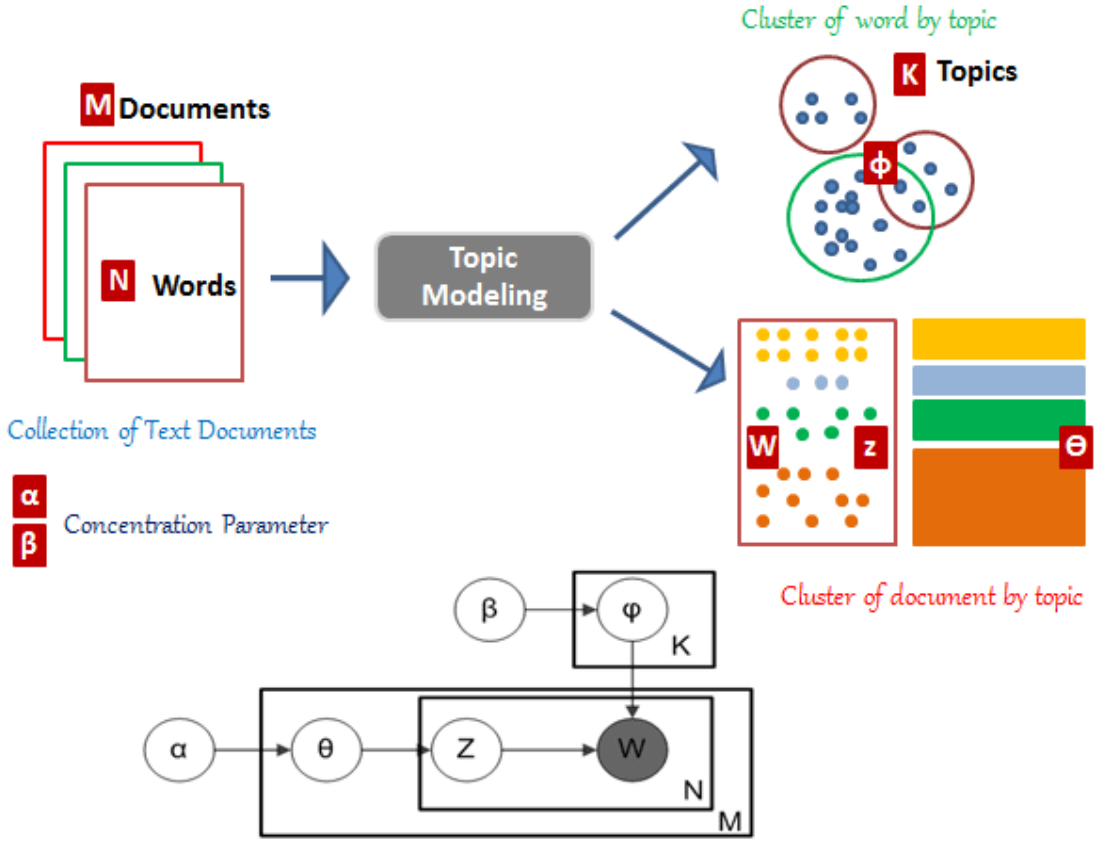


Graphical model representation of LDA. The boxes are "plates" representing replicates.
The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

- Nodes are random variables
- Edges denote possible dependence
- Observed variables are shaded
- Plates denote replicated structure

K	specified number of topics	i	auxiliary index over words in a document
k	auxiliary index over topics	α	positive K-vector
V	number of words in vocabulary	β	positive V-vector
v	auxiliary index over topics	$Dir(\alpha)$	a K-dimensional Dirichlet
d	auxiliary index over documents	$Dir(\beta)$	a V-dimensional Dirichlet
N_d	document length (number of words)	z	Topic indices: $z_{d,i} = k$ means that the i-th word in the d-th document is assigned to topic k

LDA



Probabilistic modeling

1. Treat data as observations that arise from a generative probabilistic process that includes hidden variables: For documents, the hidden variables reflect the thematic structure of the collection.

2. Infer the hidden structure using posterior inference: What are the topics that describe this collection?

3. Situate new data into the estimated model: How does this query or new document fit into the estimated topic structure?

Example

What is latent Dirichlet allocation? It's a way of automatically discovering topics that these sentences contain.

Suppose you have the following set of sentences:

- I like to eat broccoli and bananas.
- I ate a banana and spinach smoothie for breakfast.
- Chinchillas and kittens are cute.
- My sister adopted a kitten yesterday.
- Look at this cute hamster munching on a piece of broccoli.

Example

Given these sentences and asked for 2 topics, LDA might produce something like:

- Sentences 1 and 2: 100% Topic A
- Sentences 3 and 4: 100% Topic B
- Sentence 5: 60% Topic A, 40% Topic B
- Topic A: 30% broccoli, 15% bananas, 10% breakfast, 10% munching, … (at which point, you could interpret topic A to be about food)
- Topic B: 20% chinchillas, 20% kittens, 20% cute, 15% hamster, … (at which point, you could interpret topic B to be about cute animals)

How does LDA perform this discovery?

LDA training

- Go through each document, and randomly assign each word in the document to one of the K topics.
- Notice that this random assignment already gives you both topic representations of all the documents and word distributions of all the topics (albeit not very good ones).
- So to improve on them, for each document d…
- Go through each word w in d…

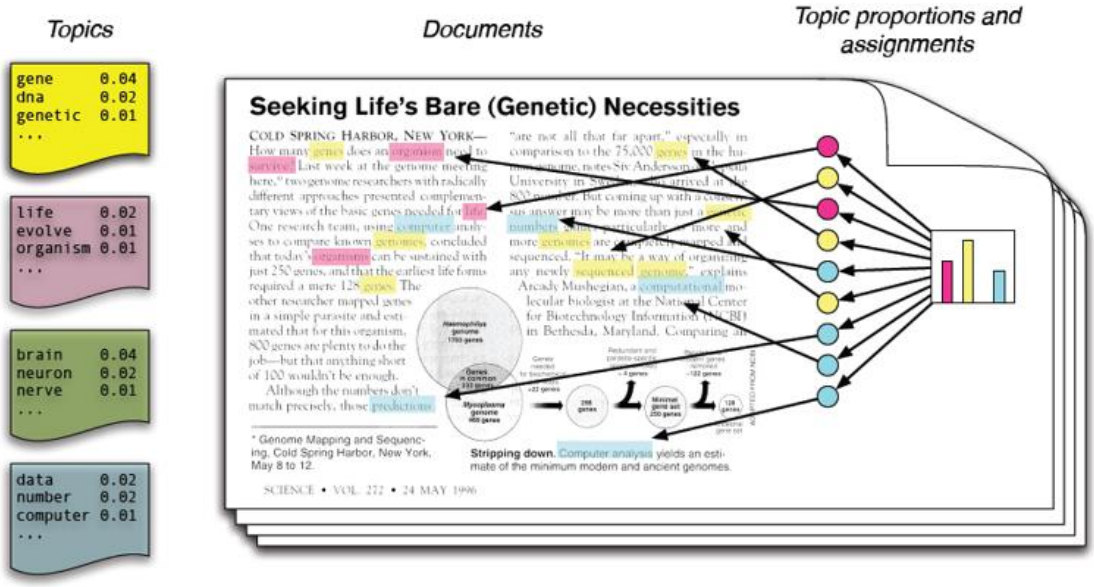
LDA training

- And for each topic t, compute two things:
  - p(topic t | document d) = the proportion of words in document d that are currently assigned to topic t, and
  - p(word w | topic t) = the proportion of assignments to topic t over all documents that come from this word w.
- Reassign w a new topic, where we choose topic t with probability p(topic t | document d) * p(word w | topic t)
- In other words, in this step, we're assuming that all topic assignments except for the current word in question are correct, and then updating the assignment of the current word using our model of how documents are generated.

LDA training

- After repeating the previous step a large number of times, you'll eventually reach a roughly steady state where your assignments are pretty good.
- Use these assignments to estimate the topic mixtures of each document (by counting the proportion of words assigned to each topic within that document) and the words associated to each topic (by counting the proportion of words assigned to each topic overall).

LDA: Identifying structure in text



In R

```
library(topicmodels)

LDA(data,
```

```
    k = 5,
    method= "Gibbs",
    control = list(seed = 321))
```

Cluster Validation

Desirable properties of clustering algorithms

- Scalability
  - Both in time and space
- Ability to deal with various types of data
  - No/less assumption about input data
  - Minimal requirement about domain knowledge
- Interpretability and usability

What is a good clustering?

- Internal criterion: A good clustering will produce high quality clusters in which:
  - the intra-class (that is, intra-cluster) similarity is high
  - the inter-class similarity is low
  - The measured quality of a clustering depends on both the document representation and the similarity measure used

Cluster validation

- Criteria to determine whether the clusters are meaningful
  - Internal validation
    - Stability and coherence
  - External validation
    - Match with known categories

Internal validation

- Coherence
  - Inter-cluster similarity v.s. intra-cluster similarity
  - Davies–Bouldin index
    - $DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$ ← Evaluate every pair of clusters

- where k is total number of clusters, σ_i is average distance of all elements in cluster i from the cluster center, $d(c_i, c_j)$ is the distance between cluster centroid c_i and c_j.

We prefer smaller DB-index!

External criteria for clustering quality

- Quality measured by its ability to discover some or all of the hidden patterns or latent classes in gold standard data
- Assesses a clustering with respect to ground truth … requires labeled data
- Assume documents with C gold standard classes, while our clustering algorithms produce K clusters, $\omega_1, \omega_2, \ldots, \omega_K$ with n_i members.

Summary

Summary

- Text clustering
- In clustering, clusters are inferred from the data without human input (unsupervised learning)
- Topic modeling

Practical 6