Text Clustering

Applied Text Mining

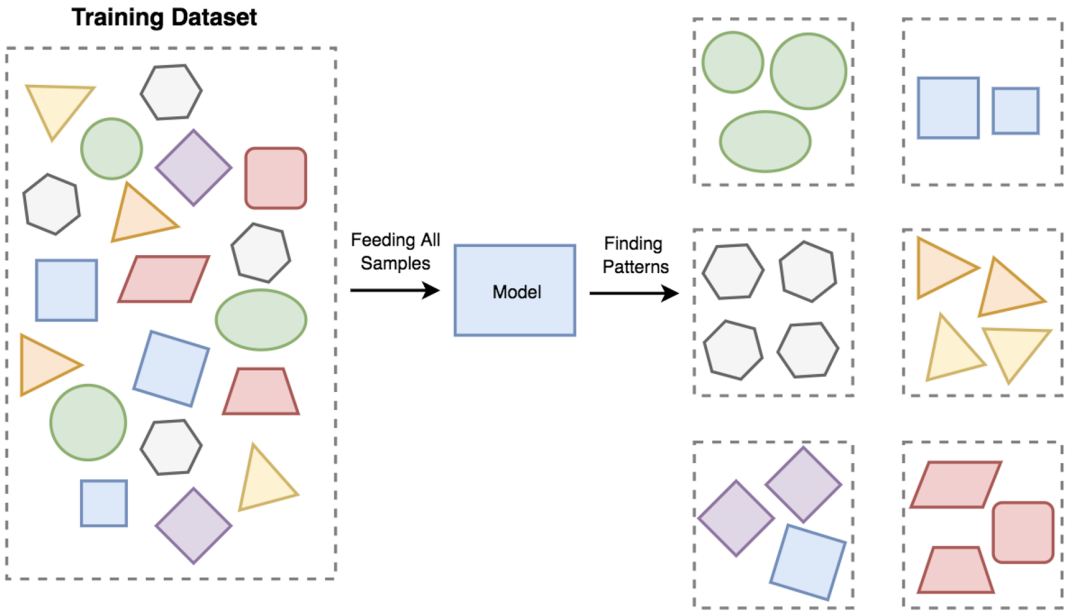
Ayoub Bagheri

Contents

Lecture's plan

1. What is text clustering?
2. What are the applications?
3. How to cluster text data?

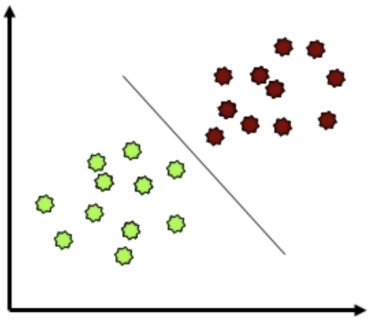
Unsupervised learning



Clustering versus classification

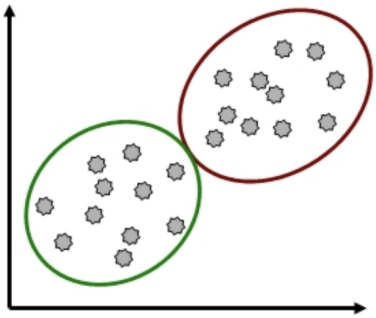
CLASSIFICATION

- Labeled data points
- Want a "rule" that assigns labels to new points
- Supervised learning



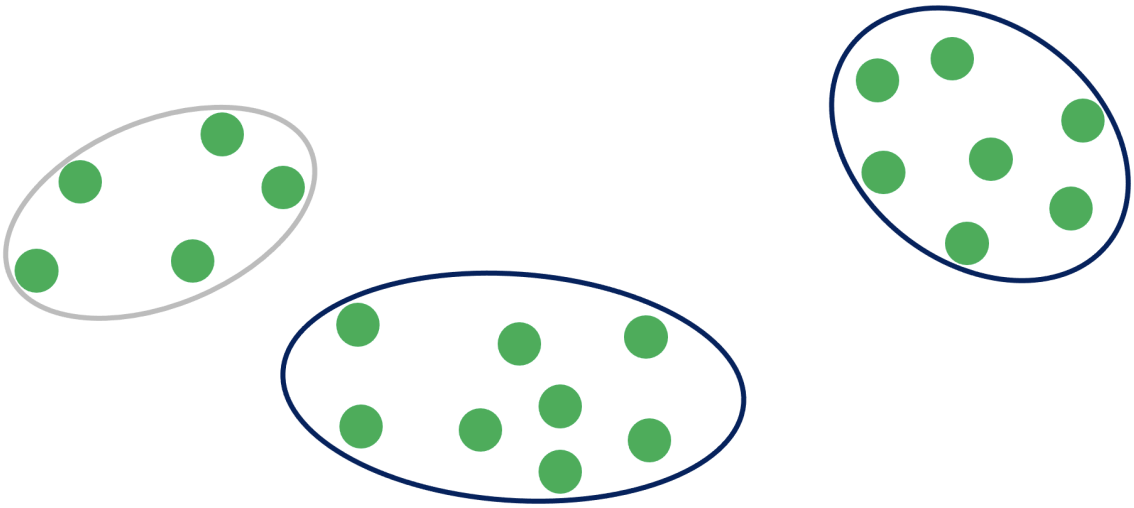
CLUSTERING

- Data is not labeled
- Group points that are "close" to each other
- Identify structure or patterns in data
- Unsupervised learning



Clustering

- Clustering: the process of grouping a set of objects into clusters of similar objects
- Discover "natural structure" of data
  - What is the criterion?
  - How to identify them?
  - How to evaluate the results?



Question

Which one is not a text clustering task?

- Grouping Trump's tweets and finding the main topics
- Finding similar patterns (demands) in customer reviews
- Grouping scientific articles
- Detection of heart failure (0 or 1) using discharge summaries

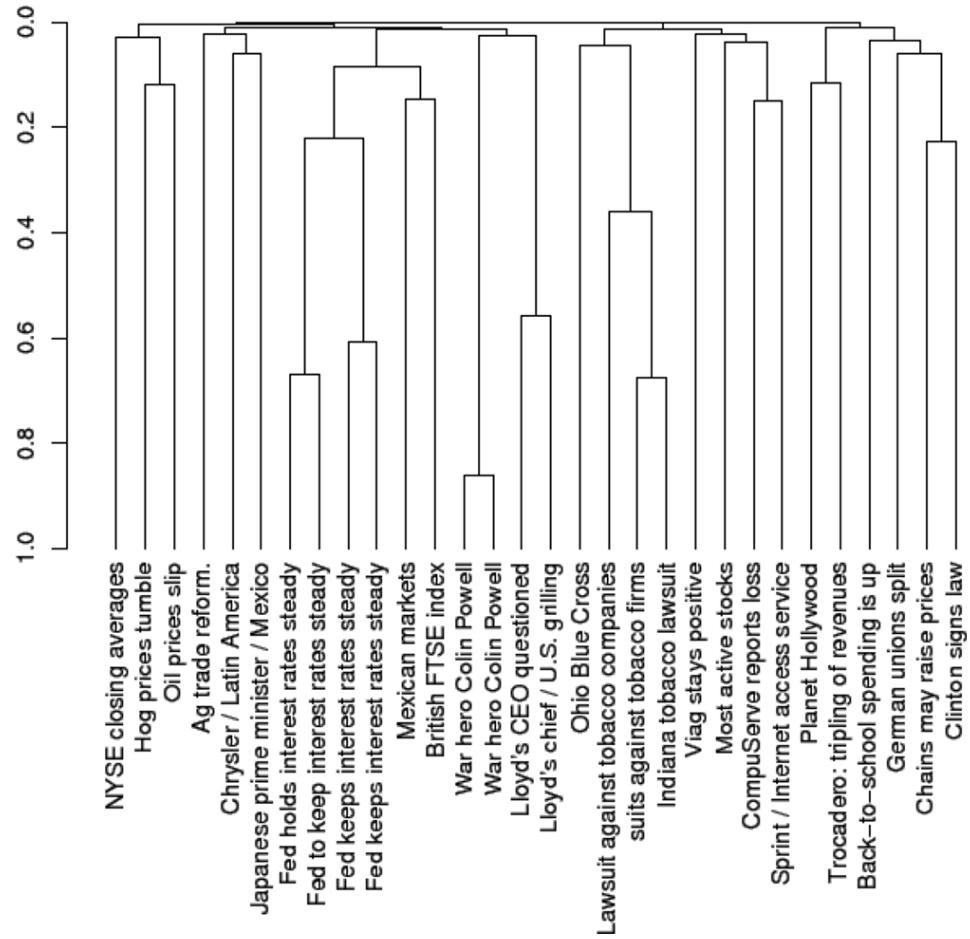
Please go to www.menti.com and use the code 7338 2184

Clustering

- Basic criteria
  - high intra-cluster similarity
  - low inter-cluster similarity
- No (little) supervision signal about the underlying clustering structure
- Need similarity/distance as guidance to form clusters

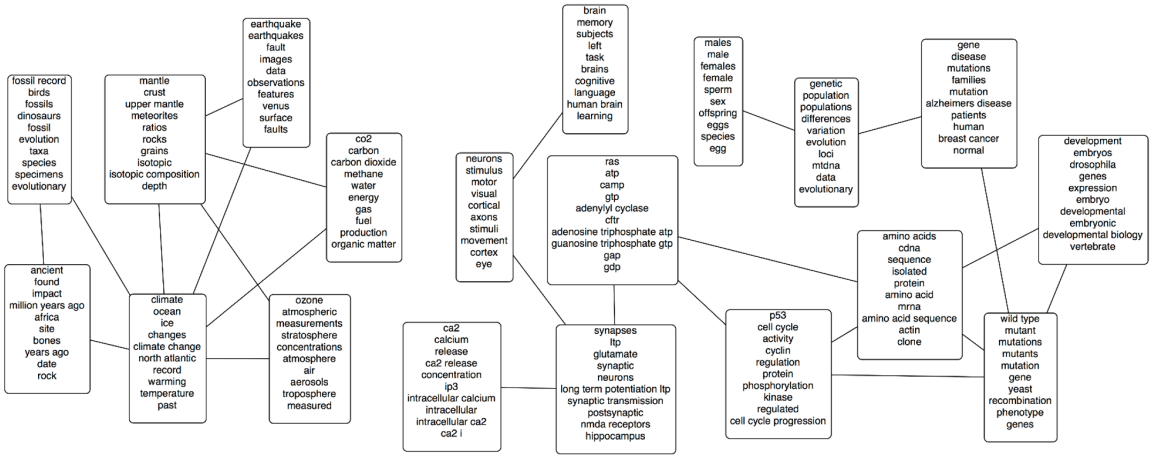
Applications of text clustering

- Organize document collections
  - Automatically identify hierarchical/topical relation among documents



Applications of text clustering

- Topic modeling
  - Grouping words into topics



Clustering algorithms

Clustering algorithms

- Partitional clustering
- Hierarchical clustering
- Topic modeling

Hard versus soft clustering

- Hard clustering: Each document belongs to exactly one cluster
  - More common and easier to do
- Soft clustering: A document can belong to more than one cluster.

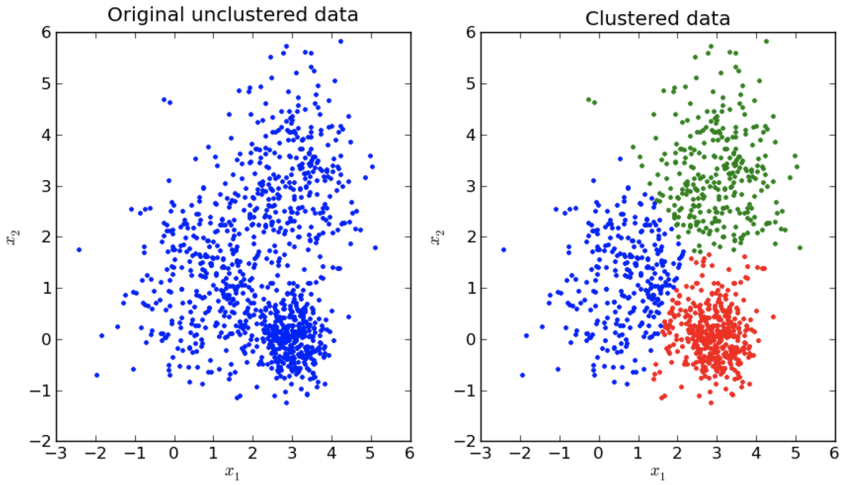
Partitional clustering

Partitional clustering algorithms

- Partitional clustering method: Construct a partition of n documents into a set of K clusters
- Given: a set of documents and the number K
- Find: a partition of K clusters that optimizes the chosen partitioning criterion
  - Globally optimal
    - Intractable for many objective functions
    - Ergo, exhaustively enumerate all partitions
  - Effective heuristic methods: K-means and K-medoids algorithms

Partitional clustering algorithms

- Typical partitional clustering algorithms
  - k-means clustering
    - Partition data by its closest mean



K-Means algorithm

- Assumes documents are real-valued vectors.
- Clusters based on centroids of points in a cluster, c:

$$\vec{\mu}(c) = \frac{1}{|c|} \sum_{\vec{a} \in c} \vec{x}$$

- Reassignment of instances to clusters is based on distance to the current cluster centroids.

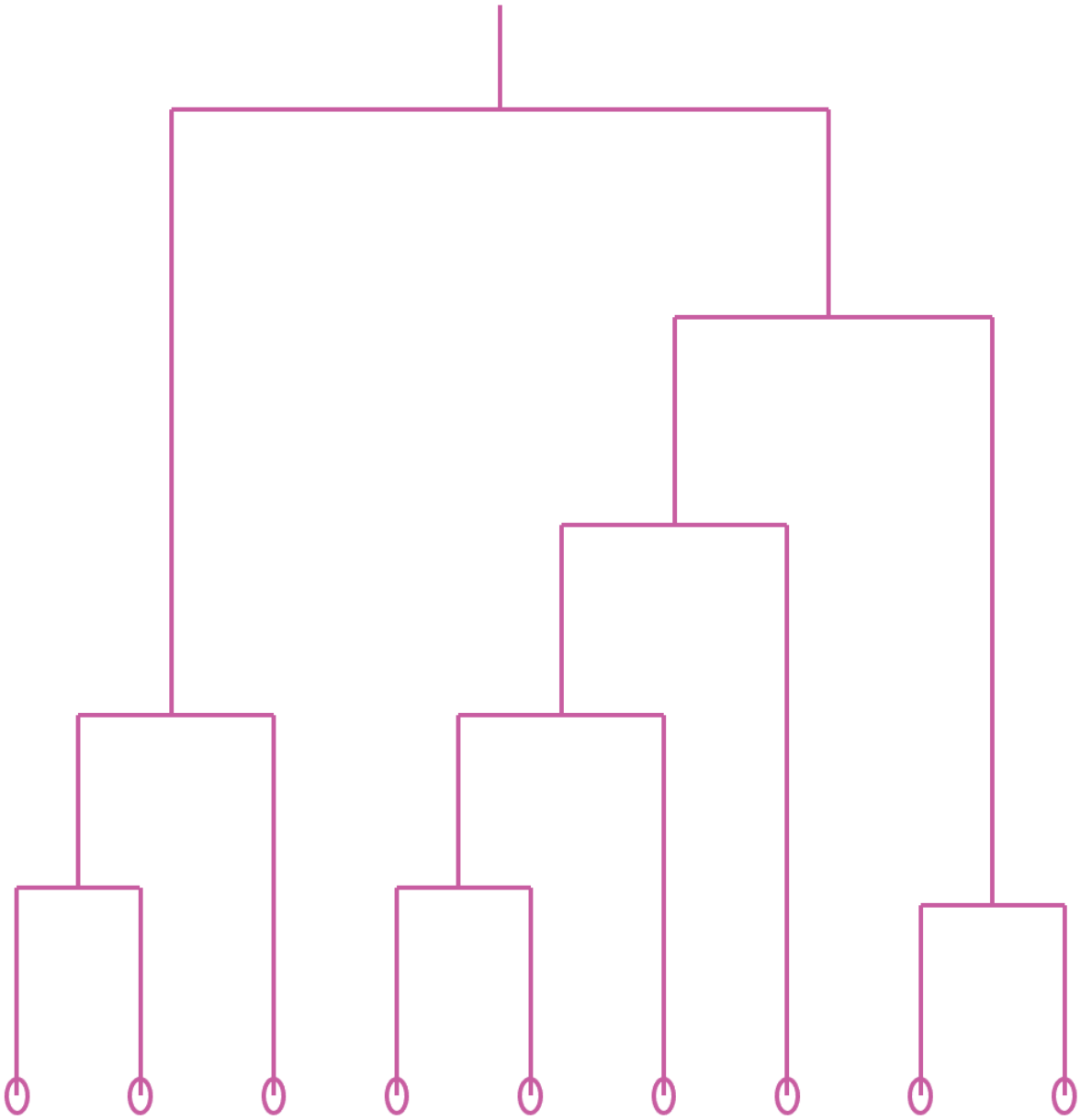
K-Means algorithm

- Select K random docs $\{s_1, s_2, \ldots s_K\}$ as seeds.
- Until clustering converges (or other stopping criterion):
  - For each doc d_i:
    - Assign d_i to the cluster c_j such that $dist(x_i, s_j)$ is minimal.
  - (Next, update the seeds to the centroid of each cluster)
  - For each cluster cj
    - $s_j = \mu(c_j)$

Hierarchical Clustering

Dendrogram: Hierarchical clustering

- Build a tree-based hierarchical taxonomy (dendrogram) from a set of documents.
- Clustering obtained by cutting the dendrogram at a desired level: each connected component forms a cluster.

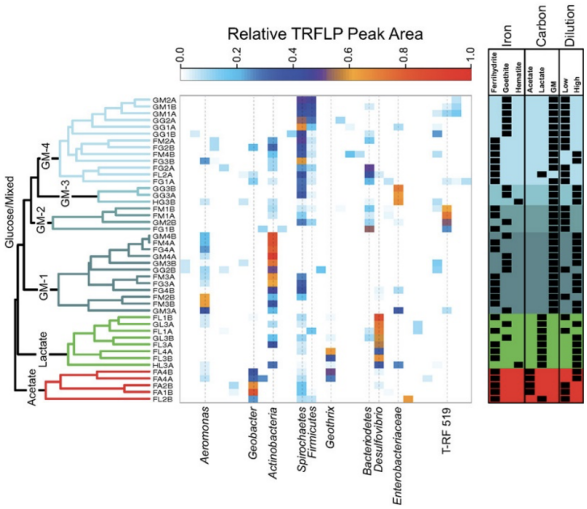


Clustering algorithms

- Typical hierarchical clustering algorithms

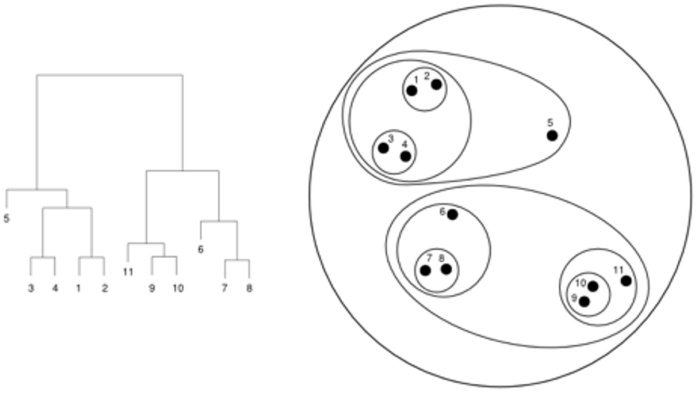
- Bottom-up agglomerative clustering
  - Start with individual objects as separated clusters
  - Repeatedly merge closest pair of clusters

Most typical usage: gene sequence analysis



Clustering algorithms

- Typical hierarchical clustering algorithms
  - Top-down divisive clustering
    - Start with all data as one cluster
    - Repeatedly splitting the remaining clusters into two



Hierarchical Agglomerative Clustering (HAC)

- Starts with each doc in a separate cluster
  - then repeatedly joins the closest pair of clusters, until there is only one cluster.
- The history of merging forms a binary tree or hierarchy.

Closest pair of clusters

- Many variants to defining closest pair of clusters (linkage methods):
  - Single-link
    - Similarity of the most cosine-similar
  - Complete-link
    - Similarity of the "furthest" points, the least cosine-similar
  - Centroid
    - Clusters whose centroids (centers of gravity) are the most cosine-similar
  - Average-link
    - Average cosine between pairs of elements
  - Ward's linkage
    - Ward's minimum variance method, much in common with analysis of variance (ANOVA)
    - The distance between two clusters is computed as the increase in the "error sum of squares" (ESS) after fusing two clusters into a single cluster.

Topic Modeling

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

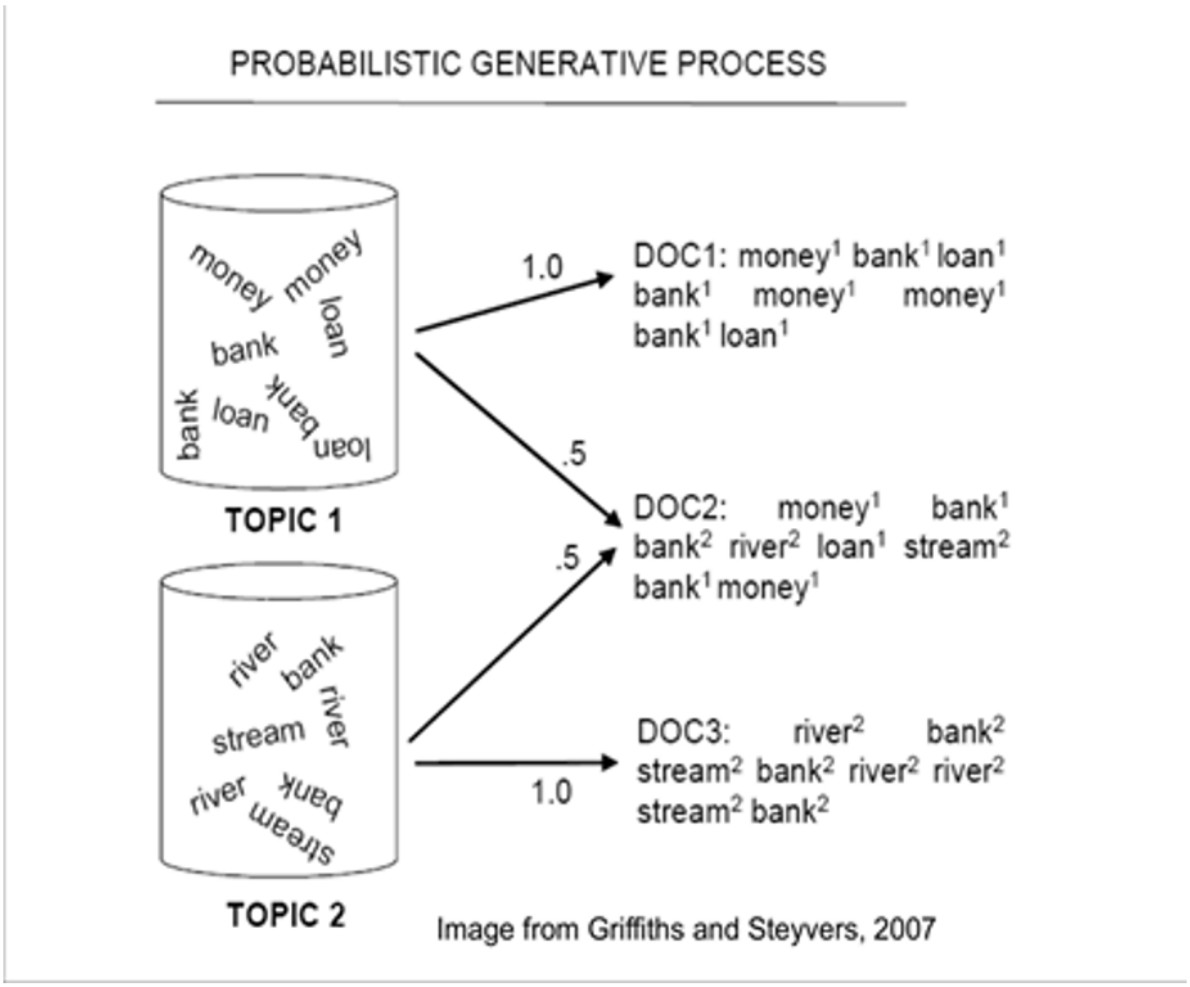
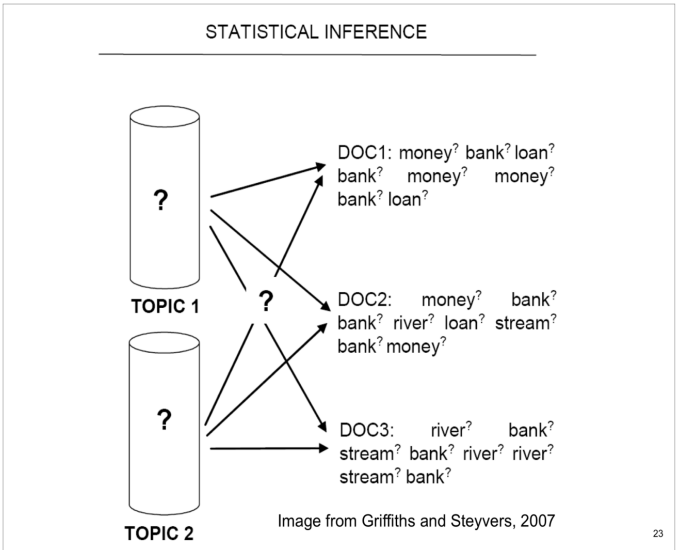
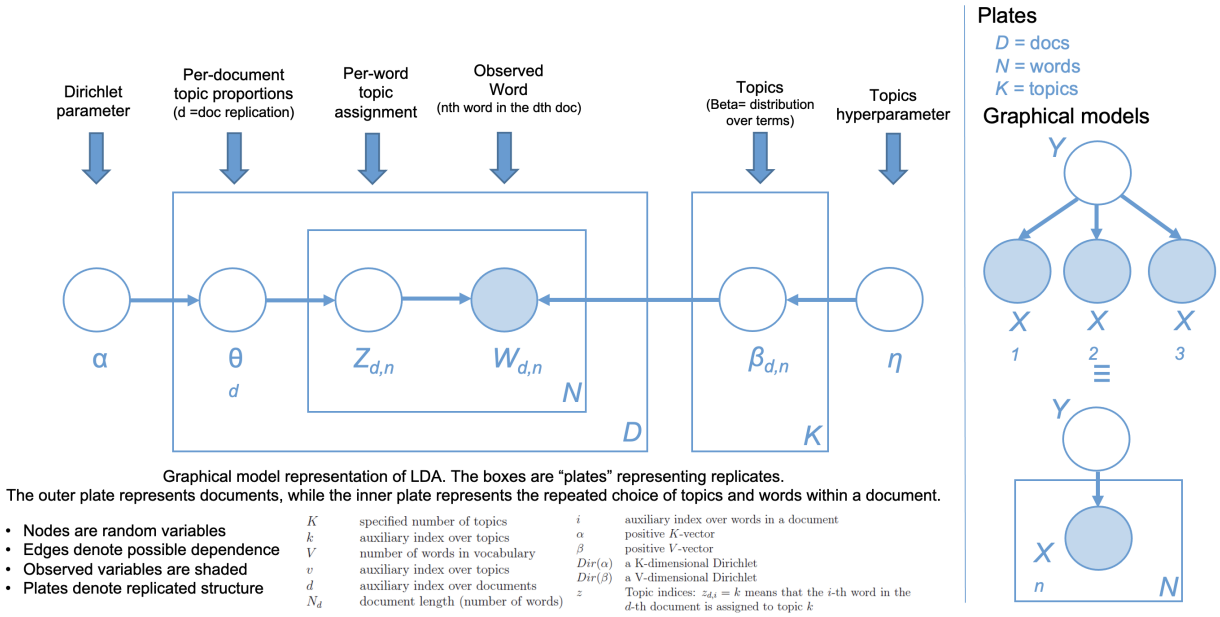


Image from Griffiths and Steyvers, 2007

Topic models | Reality: Documents observed, infer topics



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

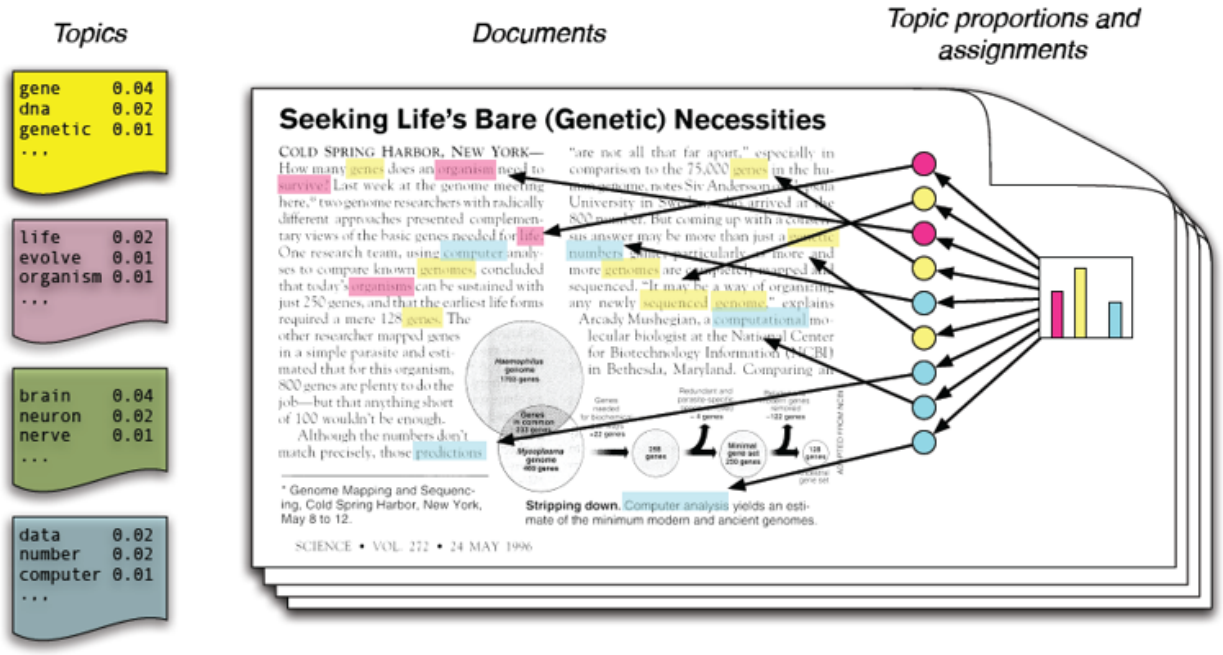
Probabilistic modeling

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?

LDA: Identifying structure in text



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
    - $DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i}(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)})$ ← 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)
- Many ways of influencing the outcome of clustering: number of clusters, similarity measure, representation of documents
- Evaluation

Practical 4