

Text Clustering and Topic Modeling

Applied Text Mining

Dr. Maryam Movahedifar

14-17 July 2025

University of Bremen, Germany

movahedm@uni-bremen.de



Universität
Bremen



DATA SCIENCE
CENTER

Introduction to Clustering

Clustering Methods

Hard vs. Soft Clustering

Partitional Clustering

Topic Modeling

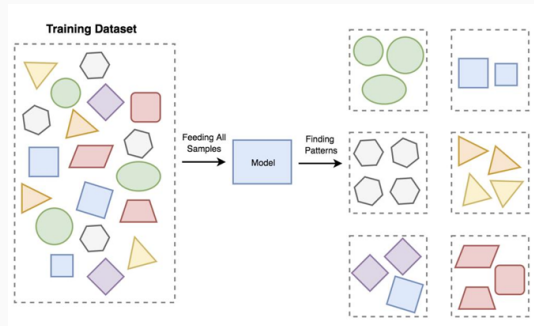
Cluster Validation

Introduction to Clustering

What is Clustering?

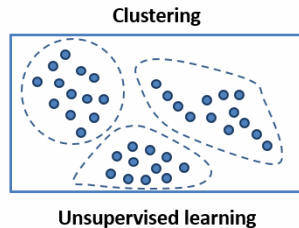
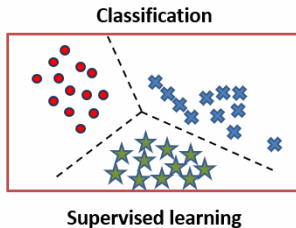
Clustering is the process of grouping similar objects into clusters without prior knowledge of the categories. It helps discover the natural structure in the data.

- **Criterion:** High intra-cluster similarity, low inter-cluster similarity.
- **Applications:** Grouping tweets, customer reviews, scientific articles, etc.



Clustering vs. Classification

- **Clustering:** Unsupervised learning where clusters are inferred from data without labeled input.
- **Classification:** Supervised learning that assigns predefined labels to data.



Clustering Methods

Clustering Algorithms Overview

Clustering algorithms can be categorized by their approach:

Hard vs. Soft Clustering

Hard: Each data point belongs to one cluster.

Soft: A point may belong to multiple clusters with probabilities.

Hierarchical Clustering

Builds a nested cluster structure using a **dendrogram**.

Useful for multilevel analysis.

Partitional Clustering

Divides data into distinct, non-overlapping clusters.

Examples: **K-means**, **K-medoids**.

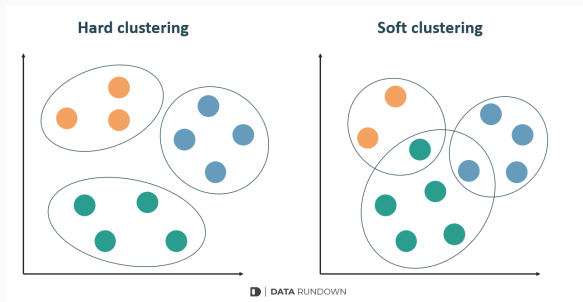
Topic Modeling

Discovers hidden topics in text corpora.
Example: **LDA**. Documents = mixtures of topics.

Hard vs. Soft Clustering

Hard vs. Soft Clustering

- **Hard Clustering:** Each document belongs to exactly one cluster.
- **Soft Clustering:** A document can belong to multiple clusters.



Partitional Clustering

Partitional Clustering: Overview

Partitional clustering divides n documents into K clusters by optimizing a partitioning criterion.

Objective: Minimize **intra-cluster distance**, maximize **inter-cluster distance**.

Challenges: Finding the globally optimal partition is computationally hard for many objective functions.

Heuristic Method: K-Means

- Treats each document as a real-valued vector.
- Assigns each document to the nearest cluster centroid.
- Iteratively updates centroids and reassigns points until convergence.

K-Means Algorithm: Mathematical Explanation

Objective: Group points into K clusters by minimizing the total squared distance between each point and its cluster center.

K-Means Steps:

1. **Initialization:** Select K random centroids μ_1, \dots, μ_K .
2. **Assignment:** Assign each point \mathbf{x}_i to the nearest centroid:

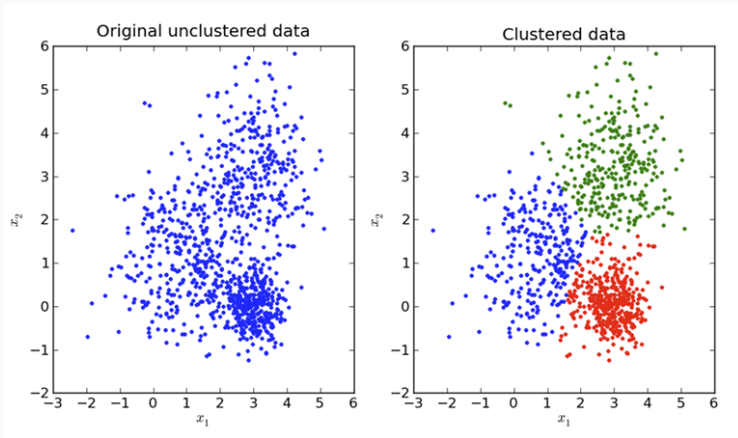
$$C_j = \{\mathbf{x}_i : \|\mathbf{x}_i - \mu_j\|^2 \leq \|\mathbf{x}_i - \mu_k\|^2, \forall k\}$$

3. **Update:** Recompute each centroid using:

$$\mu_j = \frac{1}{|C_j|} \sum_{\mathbf{x}_i \in C_j} \mathbf{x}_i$$

4. **Repeat:** Continue steps 2 and 3 until centroids stop changing.

Example of K-Means



Hierarchical Clustering Overview

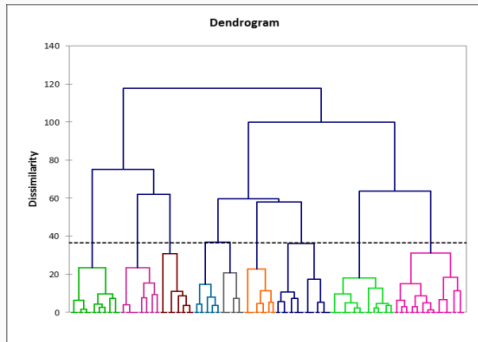
Hierarchical clustering builds a tree-based hierarchical taxonomy (*dendrogram*) to group data into clusters based on their similarity.



Dendrogram: A tree structure representing the nested clustering of documents.

✂ **Cutting the dendrogram** at a chosen level produces clusters — each connected component becomes one.

❓ **No need to predefine** the number of clusters.



Example of a dendrogram.

Types of Hierarchical Clustering

Hierarchical clustering is divided into two main types:

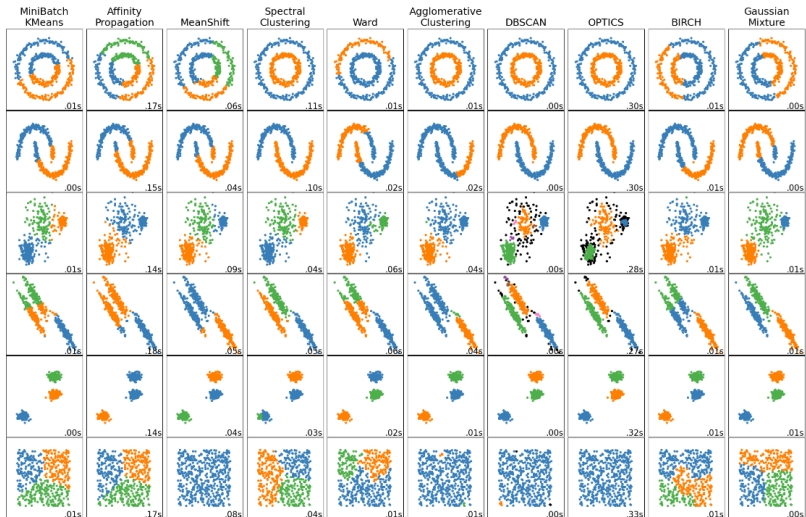
- **Top-down Divisive Clustering**
 - Start with all data in one big cluster.
 - Keep splitting clusters into two smaller ones.
 - Works well when data naturally splits into big groups first.
- **Bottom-up Agglomerative Clustering (HAC)**
 - Start with each item as its own cluster.
 - Merge the most similar pairs step by step.
 - Builds a hierarchy shown as a dendrogram (tree structure).

Linkage Methods in Agglomerative Clustering

How do we decide which clusters to merge? Different linkage methods define how distance between clusters is measured.

- **Single Link** – Merges clusters based on the *shortest distance* between any two points (nearest neighbor).
- **Complete Link** – Uses the *largest distance* between points in different clusters (farthest neighbor).
- **Average Link** – Computes the *average distance* between all point pairs from two clusters.
- **Centroid Link** – Uses the distance between the *centers* (centroids) of clusters.
- **Ward's Method** – Merges the pair of clusters that results in the *smallest increase in total variance*.

Comparison of Clustering Algorithms



A comparison of the clustering algorithms in scikit-learn

Topic Modeling

Topic Modeling in Machine Learning

Topic modeling

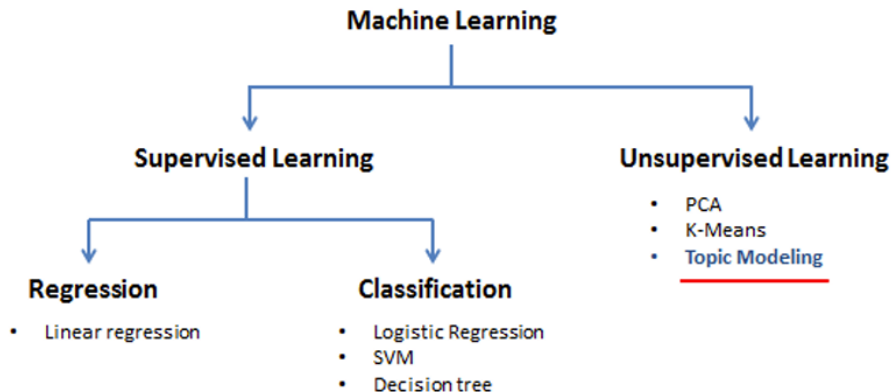


Figure 2: Overview of Machine Learning: Supervised vs. Unsupervised Learning

Topic Modeling and Latent Dirichlet Allocation (LDA)

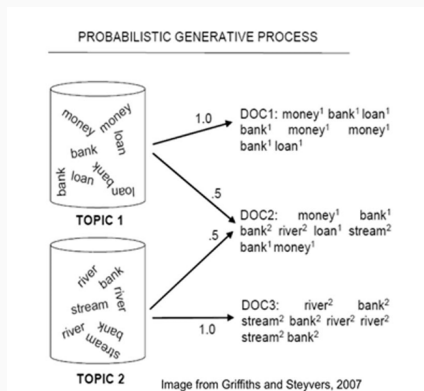
Topic modeling reveals hidden themes in document collections, and **LDA** is a key method.

Each document is represented as a **mixture of topics** with certain probabilities.

► Each topic is a **distribution over words**.

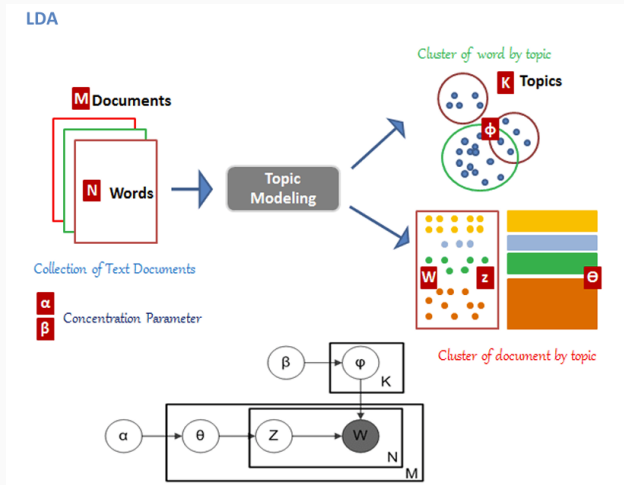
Example interpretation of the diagram:

- **Topic 1** has words like "money," "bank," and "loan."
- **Topic 2** has words like "river," "stream," and "bank."
- Document 2 (DOC2) shows an equal mix of both topics, illustrating topic overlap.



LDA: Process Overview

This slide provides an overview of how LDA clusters words and documents by topics.



Posterior Inference and New Data Integration

Posterior Inference: Estimating topic distributions for existing documents using the learned model.

- 🔍 Identify the topics that describe the current document collection.
- 🕒 Estimate topic proportions for each document using the posterior distribution.

New Data Integration: Incorporating unseen documents into an existing topic model.

- 📄 For a new document, determine how it fits within the existing topics.
- ⚙️ Use the pre-trained topic-word distributions to compute its topic mix.

Example: *"I enjoy eating broccoli while watching football."*

Topic 1 (Food): broccoli, banana

Topic 2 (Sports): football, tennis

LDA assigns: 70% Topic 1, 30% Topic 2

How LDA Learns: Iterative Word Reassignment

LDA improves its understanding of topics through repeated reassignment of word-topic relationships.

1. Initialization: Each word in every document is randomly assigned to a topic.

2. Iterative Update: For each word in each document, the topic assignment is updated based on:

- $p(t|d)$: How often topic t appears in document d .
- $p(w|t)$: How often word w appears in topic t across the whole corpus.

This is repeated for many iterations until topic distributions stabilize.

LDA: Identifying Structure in Text

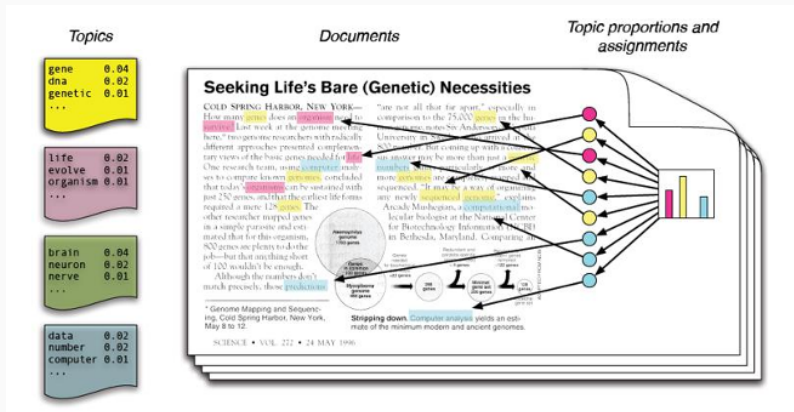





Figure 3: Overview of Identifying Structure in Text


Beyond Basic LDA: Variants and Enhancements

LDA has inspired several advanced variations to handle more complex task data:

 **Hierarchical LDA (hLDA):** Discovers multi-level topic structures — like subtopics within larger themes — forming a tree of topics.

 **Supervised LDA (sLDA):** Learns topics aligned with known outcomes or labels, useful for classification tasks.

 **Hybrid LDA:** Merges topic modeling with additional context, such as metadata or named entities, for richer analysis.

 **LDA + BERT:** Combines LDA with transformer-based models like BERT, bridging statistical modeling and deep learning.

Cluster Validation

Internal Validation:

Evaluates how well data points within the same cluster group together by measuring cluster coherence and separation. Common metrics include the Davies-Bouldin Index.

External Validation:

Assesses clustering quality by comparing the results against known labels or ground truth, using metrics like the Rand Index.

Clustering Performance Evaluation Metrics

Metrics Comparing to True Labels:

Rand Index, Mutual Information, Fowlkes-Mallows Score measure how well clusters match known classes.

Cluster Quality Metrics:

Homogeneity, Completeness, V-measure assess cluster purity and coverage.

Cohesion Separation:

Silhouette Coefficient, Calinski-Harabasz, and Davies-Bouldin Index measure cluster tightness and distinctness.

Pairwise Analysis:

Contingency and Pair Confusion Matrices detail cluster assignment accuracy.

Summary of Text Clustering and Evaluation

What is Text Clustering?

Text clustering is an unsupervised learning technique that groups similar documents based on their content, without relying on labeled data.

What Influences the Results?

Clustering results are shaped by:

- The number of clusters you choose
- The similarity measure used (e.g., cosine, Euclidean)
- How documents are represented (e.g., TF-IDF, embeddings)

Why Do We Evaluate Clustering?

Evaluation helps determine if the clusters are meaningful, consistent, and useful for downstream analysis or interpretation.

Practical 4