PTC | POLYTECHNIC COLLEGE SURINAME

UNIVERSITY OF APPLIED SCIENCES

Clustering and Classifying Text

NLP Techniques for Grouping and Categorizing Text

Tuesday, November 18, 2025

Introduction

Text Mining often involves understanding large amounts of unstructured text.

Two common techniques used are:

- **Clustering** → Unsupervised grouping of similar documents
- **Classification** → Supervised categorization based on labeled data

Goal: Extract meaningful structure, patterns, and insights from text data.

Clustering Text (Overview)

Definition:
Clustering is the process of automatically grouping similar documents or texts without predefined labels.

Purpose:

- Discover hidden patterns in text data
- Organize large document collections
- Identify topics or themes

Examples:

- Grouping news articles by topic
- Organizing customer reviews by sentiment
- Detecting spam vs. non-spam (unsupervised)

Text Representation

Before clustering or classification, text must be numerically represented.

Common Methods:

- Bag-of-Words (BoW): Count frequency of words
- TF-IDF: Weights important words higher
- Word Embeddings: Represent words in continuous vector space (e.g., Word2Vec, GloVe)

K-Means Clustering

What it is:
An unsupervised machine learning algorithm that divides data into k clusters based on similarity.

Algorithm Steps:

1. Choose number of clusters k
2. Randomly select k centroids
3. Assign each document to the nearest centroid
4. Recalculate centroids
5. Repeat until centroids stabilize

K-Means Clustering

Advantages:

- Fast and simple
- Works well with large datasets

Limitations:

- Must choose k in advance
- Sensitive to initial starting points
- Works best with spherical clusters

Example K-Means in Text

Suppose we have 1000 news articles:

- After applying TF-IDF, each article becomes a numeric vector.
- K-Means groups them into **k = 5** clusters:
  - Sports
  - Politics
  - Technology
  - Entertainment
  - Health

Result: Each cluster contains articles that are textually similar.

Hierarchical Clustering

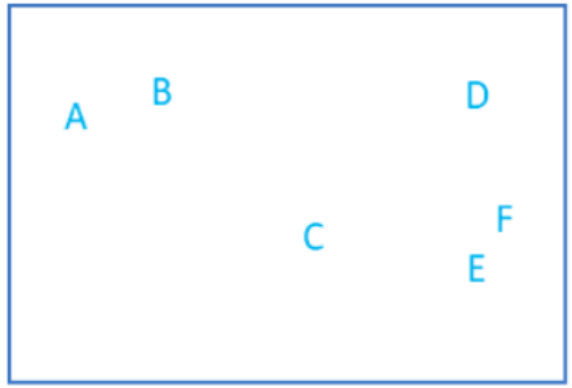
Definition:
A method that builds a hierarchy (tree) of clusters.

Two Types:

1. **Agglomerative (Bottom-Up):**
    1. Start with each document as a single cluster
    2. Merge closest clusters step by step
2. **Divisive (Top-Down):**
    1. Start with all documents in one cluster
    2. Split recursively

Visualization:

- **Dendrogram** → a tree diagram showing cluster merging



Hierarchical Clustering – Pros & Cons

Advantages:

- No need to choose *k*
- Provides a clear tree-like structure
- Good for smaller datasets

Disadvantages:

- Computationally expensive for large data
- Hard to reverse once merged or split

Comparing Clustering Methods

Feature	K-Means	Hierarchical
Type	Flat clustering	Hierarchical (tree structure)
Scalability	High	Low for large data
Need to specify K?	Yes	No (but can cut dendrogram at any level)
Output	Cluster labels	Dendrogram + cluster labels
Speed	Fast	Slower