# DATA MINING AND WAREHOUSING

## Text mining

**DHANYAJA N**
**Assistant Professor**
**STM Kannur**

# Introduction

- Text mining
- Steps in Text mining
- Text mining process
- Techniques in Text mining
- Applications of Text mining

# *Text mining*

- Text mining, also referred to as text data mining, roughly equivalent to text analytics.
- It is the process of deriving high-quality information from text.
- Text mining incorporates and integrates the tools of information retrieval, data mining, machine learning, statistics, and computational linguistics.
- Text mining deals with natural language texts either stored in semi-structured or unstructured formats.

# Steps in Text Mining

1. Gathering unstructured data from multiple data sources.

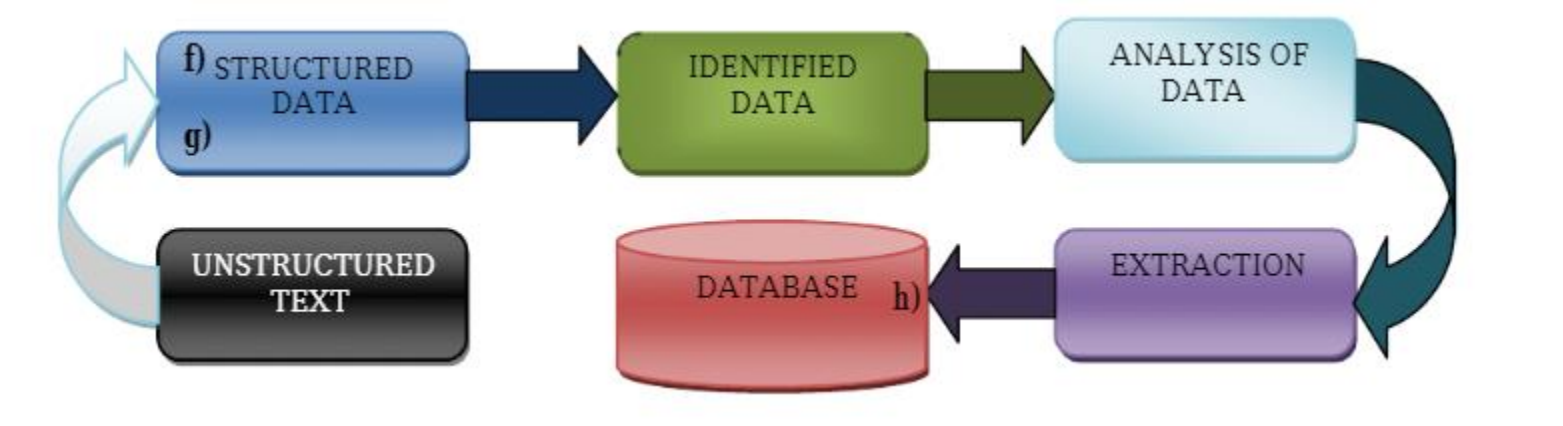
2. Detect and remove anomalies from data by conducting pre-processing and cleansing operations.

3. Convert all the relevant information extracted from unstructured data into structured formats.

4. Analyze the patterns within the data via the Management Information System (MIS).

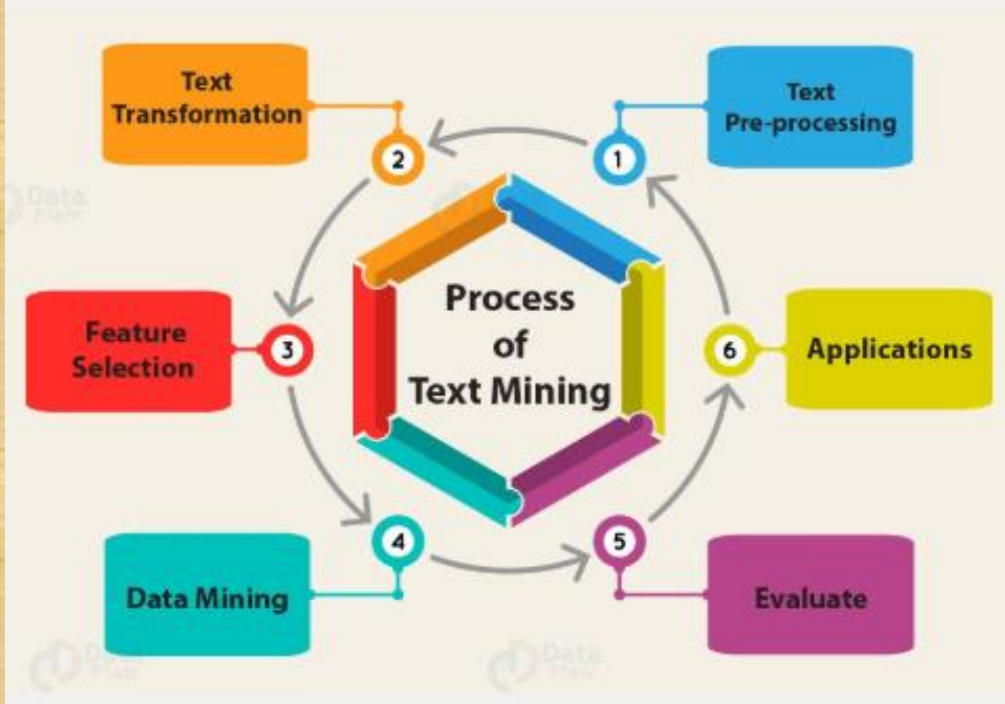
5. Store all the valuable information into a secure database to drive trend analysis and enhance the decision-making process of the organization.

# Steps in Text Mining



# Text Mining Process

- A process of Text mining involves a series of activities to perform to mine the information. These activities are:
-



# Text PreProcessing

- It involves a series of steps as shown in below:
- **Text Cleanup**
  - Text Cleanup means removing any unnecessary or unwanted information. Such as remove ads from web pages, normalize text converted from binary formats.
- **Tokenization**
  - Tokenizing is simply achieved by splitting the text into white spaces.
- **Part of Speech Tagging**
  - Part-of-Speech (POS) tagging means word class assignment to each token. Its input is given by the tokenized text. Taggers have to cope with unknown words (OOV problem) and ambiguous word-tag mappings.

# Transformation

- **Text Transformation** (Attribute Generation)
  - A text document is represented by the words it contains and their occurrences. Two main approaches to document representation are:
    - i. Bag of words
    - ii. Vector Space

# Feature Selection

- **Feature selection**
  - also is known as variable selection.
  - It is the process of selecting a subset of important features for use in model creation.
  - Redundant features are the one which provides no extra information.
  - Irrelevant features provide no useful or relevant information in any context.
- **Data Mining**
  - Classic Data Mining techniques are used in the structured database. Also, it resulted from the previous stages.
- **Evaluate**
  - Evaluate the result, after evaluation, the result discard.

# Applications Of Text Mining

- Risk Management
- Customer Care Service
- Fraud Detection
- Business Intelligence
- Social Media Analysis

# Techniques in text mining

- Techniques used in text mining techniques:
- Information Extraction
- Information Retrieval
- Categorization
- Clustering
- Summarisation

# Techniques in text mining

| Technique | Characteristics | Tools |
|---|---|---|
| Retrieval | Retrievals valuable information from unstructured text | Intelligent Miner, Text Analyst |
| Extraction | Extract information from structured database | Text Finder, Clear Forest Text |
| Summarization | Reduce length by keeping its main points and overall meaning as it is | Tropic Tracking Tool, Sentence Ext Tool |
| Categorization | Document based categorization | Intelligent Miner |
| Cluster | Cluster collection of documents, Clustering, classification and analysis of text document | Carrot, Rapid Miner |

# Thank You !!!

**DHANYAJA N**
**Assistant Professor**
**STM Kannur**