

ADVANCED CONCEPTS

UNIT-V

Web Mining:

- Web mining aims to discover useful information or knowledge from the Web hyperlink structure, page content, and usage data.
- Web mining is a data mining technique to extract knowledge from **web data**.

Web data:

- Web data includes:
 - **web documents**
 - **hyperlinks between documents**
 - **usage logs of web sites**
- The **WWW** is huge, widely distributed, global information service center and, therefore, It is a rich source for data mining.

World Wide Web:

World Wide Web or Web is the biggest and popular source of information available, reachable, and accessible at low cost provides quick response to the users and reduces burden on the users of physical movements.

The data on the Web is noisy. The noise comes from two major sources. First, a Web page contains many pieces of information, e.g., the main content of the page, routing links, advertisements, copyright notices, privacy policies, etc. Second, since the Web does not have quality control of information, i.e., a large amount of information on the Web is of low quality, erroneous, or even misleading.

- ✓ There are about **1.5 billion websites**.
- ✓ But less than **400 million are active**.
- ✓ Grows at about 1 million pages a day
- ✓ By the time you finish this class, **thousands of new sites will spawn**.
- ✓ Lots of duplication (70-80%)

Various types of data

- Text
- Images
- Audio & Video

What is Web Mining?

- Web mining is **the application of data mining techniques to discover useful information from the World Wide Web.**
- It uses automated methods to extract both **structured** and **unstructured** data from web pages, server logs and link structures.

Examples:

- **Web search**, e.g. Google, Yahoo, Bing, Dogpile ,Duck Duck Go Ecosia ,Gigablast ,...
- **Specialized search**: e.g. *Squool Tube* - Search for factual, educational videos., *Elephind* - search the world's historical newspaper archives.
- **eCommerce** : e.g. Amazon,Flipkart,eBay,Fiverr,Upwork.
 - **Advertising**, e.g. Google Adsense
 - Improving Web site design and performance

Process of Web Mining:

Web mining is the process of extracting useful information and knowledge from web data including web content, structure, and usage data. The overall process involves several systematic steps that transform raw web data into meaningful insights.

- **Data Collection**

Web data is collected from various sources, including web pages, databases, and APIs.

Various sources:

- Web pages (HTML, text, images, videos)
- Web logs (user access logs)
- Social media data
- E-commerce sites

Techniques: Web crawlers, APIs, scraping tools.

- **Data Preprocessing**

The collected data is pre-processed to remove irrelevant information, such as advertisements and duplicate content.

Tasks include:

- Removing HTML tags, advertisements, or stop words
- Tokenization (breaking text into words)
- Normalization (lowercasing, stemming)
- Removing duplicates or irrelevant data
- Structuring data into a usable format (tables, JSON, etc.)

- **Data Transformation**

The preprocessed data is transformed into suitable forms for mining.

Common transformations:

- Feature extraction (keywords, hyperlinks, frequency counts)
- Conversion into vector or numerical representation
- Dimensionality reduction (using PCA, SVD, etc.)

- **Web Mining (Core Step)**

This step applies data mining techniques to extract patterns or insights. It is divided into three main categories:

- *Web Content Mining* – Extracting useful information from the actual content of web pages (text, images, videos, etc.).
- *Web Structure Mining* – Analyzing the structure of hyperlinks between web pages to understand relationships and authority.
- *Web Usage Mining* – Studying user behavior through server logs, cookies, and clickstreams to understand navigation patterns.

- **Pattern Evaluation**

The discovered patterns or relationships are evaluated to identify the most meaningful and useful ones.

Techniques:

- Statistical validation
- Correlation analysis
- Interestingness measures (support, confidence, lift)

- **Knowledge Presentation**

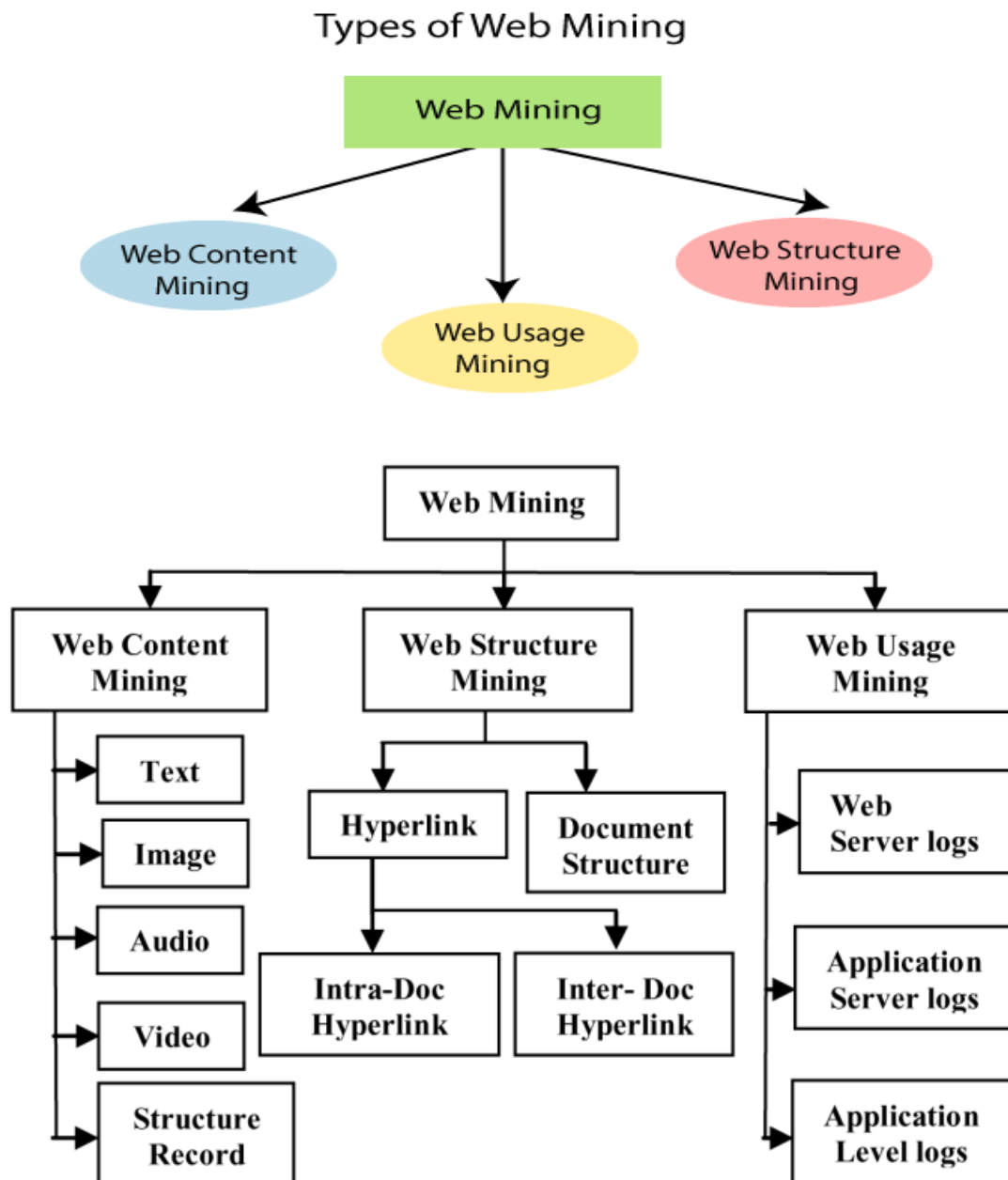
The final patterns or knowledge are presented in a human-understandable form.

Methods:

- Visualization (graphs, charts)
- Reports and summaries
- Dashboards or recommendation systems

Types of Web Mining:

Web mining tasks can be categorized into three types: *Web structure mining*, *Web content mining*, and *Web usage mining*.



Web Content Mining:

- **Web content** mining can be used to extract useful data, information, knowledge from the web page content.
- The **Web page content** includes text, images, audio, video, and structure records.
- In web content mining, **each web page** is considered as an individual document.
- The primary task of **content mining** is data extraction, where **structured** data is extracted from **unstructured** websites.
- Web content mining can be utilized to distinguish topics on the web.
- The technologies that are normally used in web content mining are **NLP** (Natural language processing) and **IR** (Information Retrieval).
- Techniques in Web Content Mining:
 - Classification
 - Clustering

For Example, if any user searches for a **specific book** on the search engine, then the user will get a list of suggestions.

Web Structure Mining:

- Web structure mining is the application of discovering **structure information** from the web. The structure of the web **graph** consists of web pages as **nodes**, and hyperlinks as **edges** connecting related pages.
- The Web structure includes **hyperlinks** and **document structures**.
- Structure mining basically shows the **structured summary** of a particular website. It identifies **relationship** between web pages linked by information or direct link connection.
- Techniques in Web Structure Mining:
 - Association Rules
 - Classification

Example:

- The most important application in this regard is the **Google search** engine, which estimates the ranking of its outcomes primarily with the PageRank algorithm.
- The rank of a page is decided by the number and quality of links pointing to the target node.

Web Usage Mining:

- Web usage mining is the application of identifying or discovering interesting **usage patterns** from large data sets. And these patterns enable you to understand the user behaviors or something like that.
- The **Web usage** includes web server logs, Application server logs, and Application level logs.
- Web usage mining is used for mining the web log records (access information of web pages) and helps to discover the user access patterns of web pages.
- **Web server** registers a web **log entry** for every web page.
- The main source of data here is **Web Server** and **Application Server**.
- Log files are created when a user/customer interacts with a web page.
- Techniques in Web Usage Mining:
 - Association Rules
 - Classification
 - Clustering

Advantage:

- This technology has enabled **e-commerce** to do **personalized marketing**, which eventually results in higher trade volumes.

Disadvantage:

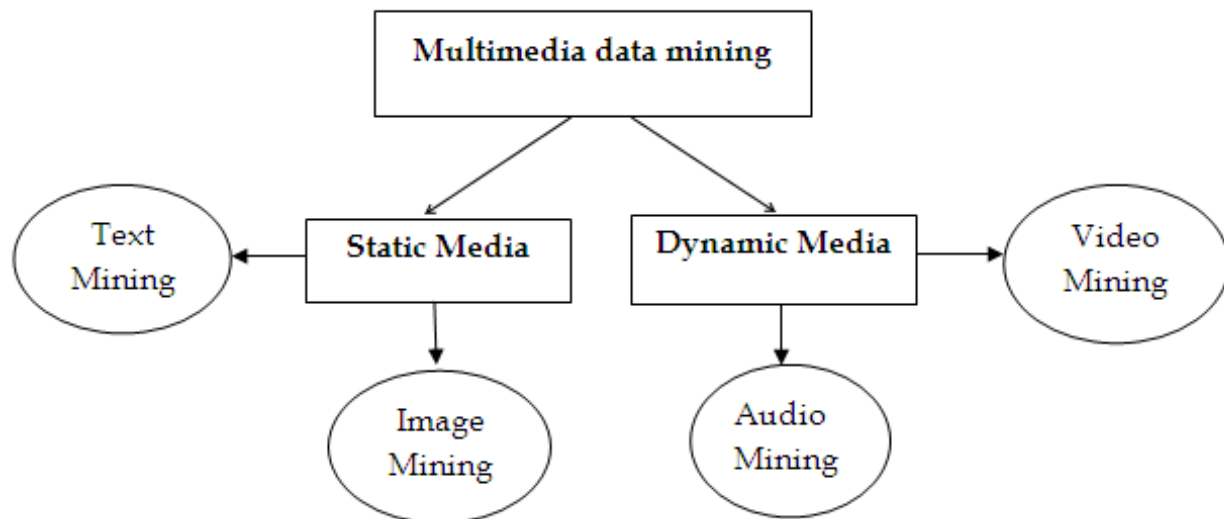
- This technology when used on data of **personal nature** might cause concerns. The most criticized **ethical issue** involving web usage mining is the invasion of **privacy**.

Multimedia Data Mining:

- Multimedia data mining discovers interesting patterns from multimedia databases that **store** and **manage** large collections of multimedia objects.
- The Multimedia data includes the following:
 - image data,
 - video data,
 - audio data,
 - sequence data,
 - hypertext data containing text.
- Multimedia data mining has a number of uses in today's society. An example of this would be the use of **traffic camera footage** to analyze **traffic flow**.
- Multimedia data mining can be defined as a process that finds patterns in various types of data, including images, audio, video, and animation.

Types of Multimedia data mining:

- Multimedia data mining is classified into two broad categories: **static** and **dynamic** media.



Text mining

- Text Mining also referred to as text data mining and it is used to find meaningful information from the unstructured texts that are from various sources.
- Text is the foremost general medium for the proper exchange of information.

- Text Mining is to evaluate huge amount of usual language text and it detects exact patterns to find useful information.

Image mining

- Image mining systems can discover meaningful information or image patterns from a huge collection of images.
- Image mining determines how low-level pixel representation consists of a raw image or image sequence can be handled to recognize high-level spatial objects and relationship.
- It includes digital image processing, image understanding, database, AI and so on.

Video mining

- Video mining has the objective of describing interesting patterns form large amount of video data.
- Video has several type of multimedia data such as image, text, audio, visual etc.
- It is commonly used in various applications like security and surveillance, entertainment, medicine, sports, and education programs.

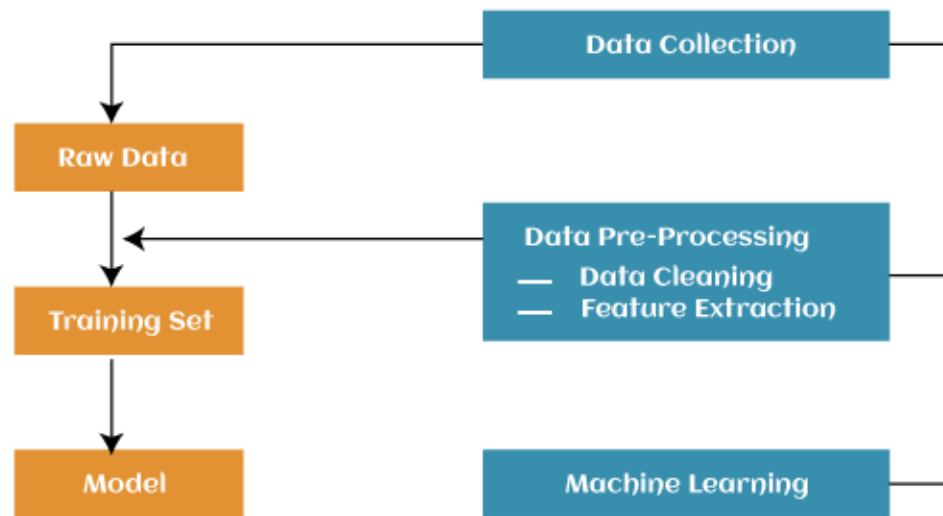
Audio mining

- Audio mining plays an important role in multimedia applications, is a technique by which the content of an audio signal can be automatically searched, analyzed and rotten with wavelet transformation.
- Band energy, frequency centroid, zero crossing rate, pitch period and band-width are often used features for audio processing.
- It is generally used in the field of automatic speech recognition, where the analysis efforts to find any speech within the audio.

Applications of Multimedia Mining:

- Digital Library
- Traffic Video Sequences
- Medical Analysis
- Media Making and Broadcasting
- Surveillance system
- Customer Insight

Process of Multimedia Data Mining:

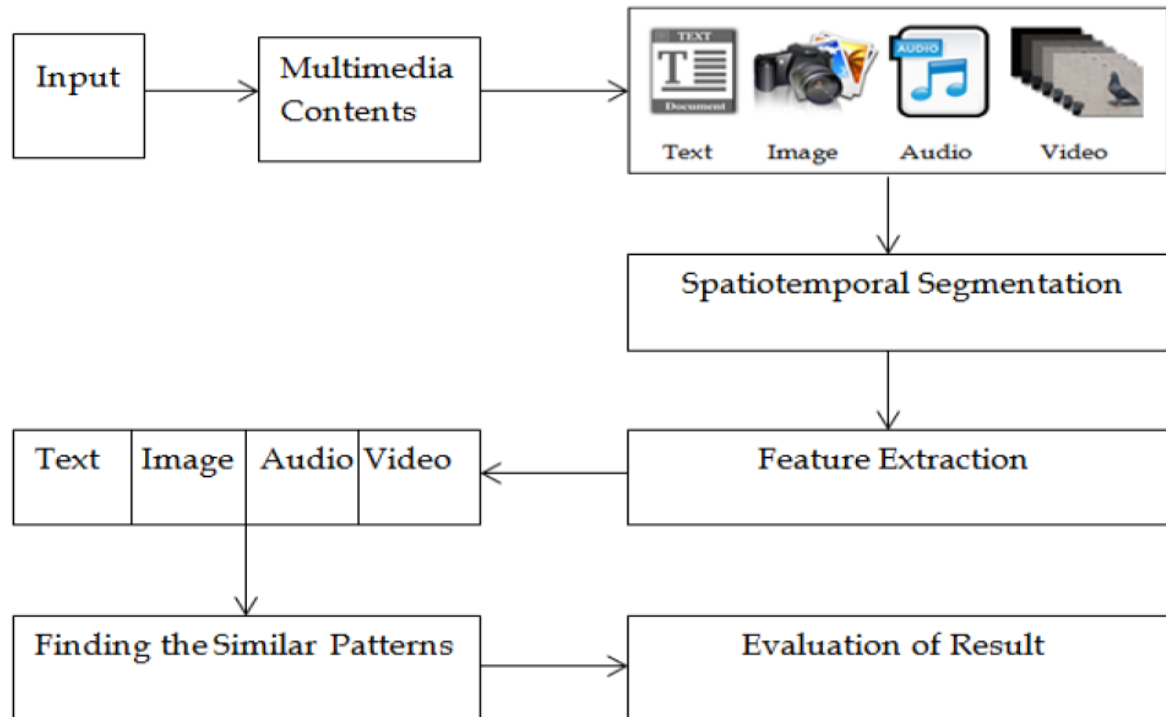


Multimedia Mining Process

The above diagram describes the process of applying multimedia mining in different multimedia types.

- **Data collection** is the starting point of a learning system, as the quality of raw data determines the overall achievable performance. Then, the goal of data pre-processing is to discover important features from raw data.
- **Data pre-processing** includes data cleaning, normalization, transformation, feature selection, etc. Learning can be straightforward, if informative features can be identified at pre-processing stage. Detailed procedure depends highly on the nature of raw data and problem's domain.
- **Training set** is the product of data pre-processing. A learning model must be chosen to learn from training set.
- Create a **model** on training dataset using a **machine learning** algorithm.

Architecture for Multimedia Data Mining:



In the above architecture,

- **Input stage** comprises which multimedia database is used for finding the patterns and to perform data mining process.
- **Multimedia Content** is the data selection stage which requires the user to select the databases, subset of fields or data to be used for data mining.
- **Spatio-temporal segmentation** is nothing but moving objects in image sequences in the videos and it is useful for object segmentation.
- **Feature extraction** is the pre-processing step that involves integrating data from various sources and making choices regarding characterizing or coding certain data fields to serve when inputs to the pattern finding stage. In Multimedia Data mining the preprocessing stage is significant since the unstructured nature of multimedia records.
- **Finding the Similar pattern** stage is the heart of the whole data mining process. The hidden patterns and trends in the data are basically uncovered in this stage. Some approaches of finding similar pattern stage contain association, classification, clustering, regression, time-series analysis, and visualization.

- ***Evaluation of Results*** is a data mining process used to evaluate the results and this is important to determine whether prior stage must be revisited or not. This stage consists of reporting and makes use of the extracted knowledge to produce new actions or products and services or marketing strategies.

We considered two main families of multimedia retrieval systems, i.e. similarity search in multimedia data.

- ***Description-based retrieval system*** creates indices and object retrieval based on image descriptions, such as keywords, captions, size, and creation time.
- ***Content-based retrieval system*** supports image content retrieval, for example, color histogram, texture, shape, objects, and wavelet transform.

Models for Multimedia Mining

The data mining models / techniques that are applied to multimedia data are

- **classification,**
- **clustering,**
- **association rule mining**

Text Mining:

- Text data mining can be described as the process of **extracting** essential data from standard language text.
- All the data that we generate via:
 - **Text messages**
 - **Text documents**
 - **Emails**
 - **Files**
- The primary source of data is *e-commerce websites, social media platforms, published articles, survey*, and many more.
- The larger part of the generated data is **unstructured**, which makes it **challenging** and **expensive** for the organizations to analyze with the help of the people.
- Text mining, also referred to as text data mining, roughly equivalent to **text analytics**, is the process of deriving **high-quality** information from text.
- Text mining uses natural language processing (**NLP**), allowing machines to understand the human language and process it automatically.
- Natural language processing (**NLP**) is the ability of a computer program to understand human language as it is spoken and written. It is a component of artificial intelligence (**AI**).

Why is Text Mining Important?

- Individuals and organizations generate tons of data every day. Stats claim that almost **80% of the existing text data is unstructured**, meaning it's not organized in a predefined way, it's not searchable, and it's almost impossible to manage. In other words, it's just not useful.

How Does Text Mining Work?

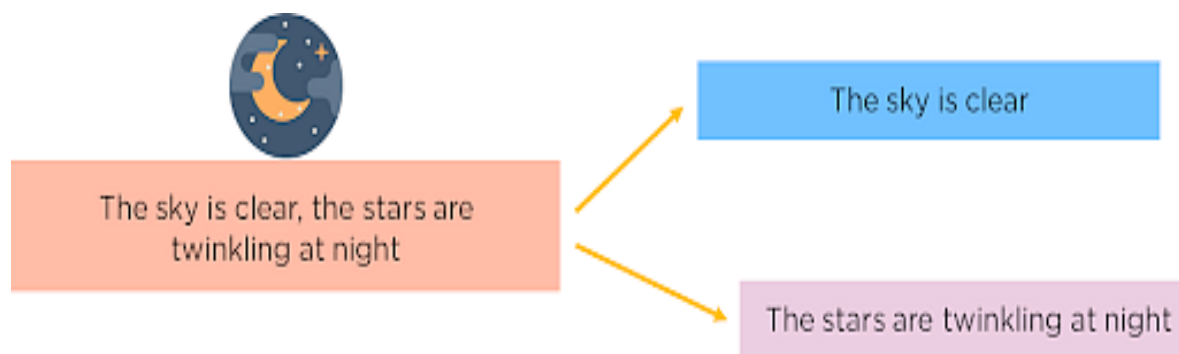
- Text mining helps to analyze large amounts of raw data and find relevant insights. Combined with machine learning, it can create text analysis models that learn to classify or extract specific information based on previous training.
- The **first step** in text mining is **collecting** or **gathering the data**.

- Data can be **internal** (interactions through chats, emails, surveys, spreadsheets, databases, etc) or **external** (information from social media, review sites, news outlets, and any other websites).
- The **second step** is preparing (**preprocessing**) your data. Text mining systems use several **NLP techniques** — like Segmentation, tokenization, lemmatization, stemming and stop removal — to build the inputs of your **machine learning model**.

The steps to perform preprocessing of data:

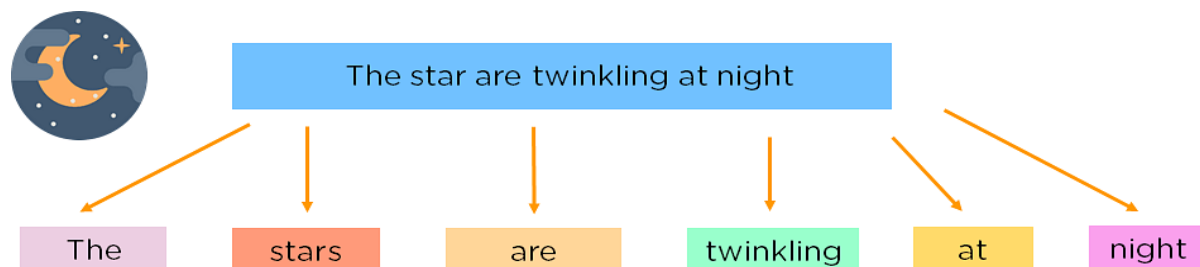
- **Segmentation:**

- Break the entire document/article into its component sentences by its punctuations like full stops and commas.



- **Tokenizing:**

- Tokenization is a process of splitting a text / sentence into smaller units (words) which are also called tokens.



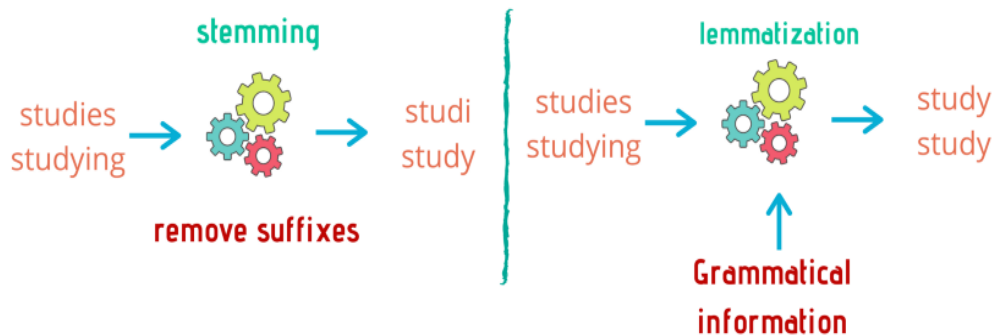
- **Stemming:**

- Stemming is a technique used to extract the base form of the words by removing affixes from them. For example, the stem of the words **eating**, **eats**, **eaten** is **eat**.

- **Lemmatization:**

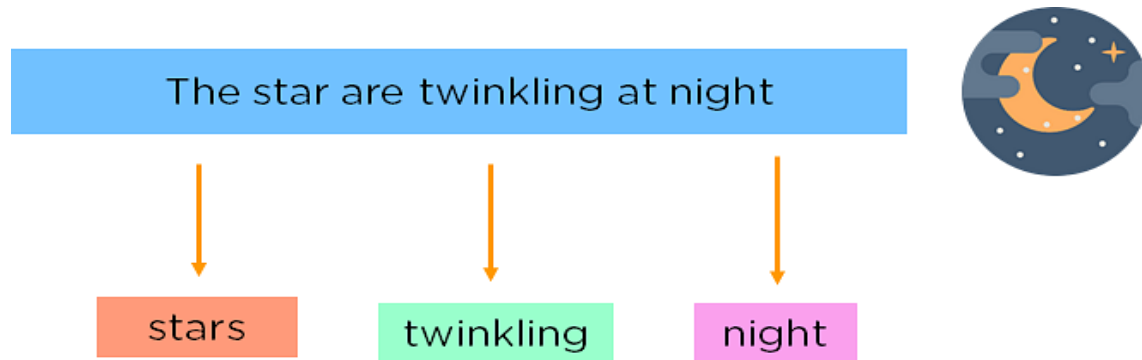
- Lemmatization considers the context and converts the word to its meaningful base form, which is called **Lemma**. For instance, stemming the word '**Caring**' would return '**Car**'. For instance, lemmatizing the word '**Caring**' would return '**Care**'.

STEMMING VS. LEMMATIZATION



- **Filtering (or) Removing Stop Words:**

- It is a process of removing non-essential words, i.e. Words such as **was, in, is, and, the**, are called stop words and can be removed.



- The **third step** is **Feature Extraction** from pre processed data.
- The mapping from textual data to real-valued vectors is called **feature extraction**.
- One of the commonly used technique to extract the features from textual data is **calculating the frequency of words/tokens** in the document/corpus.

Bag of Words (BOW):

- One of the simplest technique is **Bag of Words (BOW)** to represent the text in numerical format(**Vectors**).
- In **BOW**, we make a list of unique words in the text corpus called vocabulary. Then we can represent each **sentence** or **document** as a **vector**, with each word represented as **1** for **presence** and **0** for **absence**.
- The **bag-of-words (BOW)** model is a representation that turns arbitrary text into **fixed-length vectors** by counting how many times each word appears. This process is often referred to as **vectorization**.

- Let's understand this with an example. Suppose we wanted to vectorize the following:

Document 1 : **the cat sat**

Document 2: **the cat sat in the hat**

Document 3: **the cat with the hat**

Step 1: Determine the Vocabulary

- We first define our **vocabulary**, which is the set of all unique words found in our document set.
- The Words are :

the, cat, sat, in, hat, with

Step 2: Count

Document	the	cat	sat	in	hat	with
<i>the cat sat</i>	1	1	1	0	0	0
<i>the cat sat in the hat</i>	2	1	1	1	1	0
<i>the cat with the hat</i>	2	1	0	0	1	1

Step 3: Vector Representation

Now we have length-6 vectors for each document.

- the cat sat*: [1, 1, 1, 0, 0, 0]
 - the cat sat in the hat*: [2, 1, 1, 1, 1, 0]
 - the cat with the hat*: [2, 1, 0, 0, 1, 1]
- Notice that we lose **contextual information**, e.g. where in the document the word appeared, when we use **BOW**. It's like a literal **bag-of-words**: it only tells you *what* words occur in the document, not *where* they occurred.



- Term frequency — Inverse document frequency (**TFIDF**)
- Word2Vec (**W2V**)

Text Mining Process:

Text mining involves a series of activities to be performed to efficiently mine the information. These activities are:

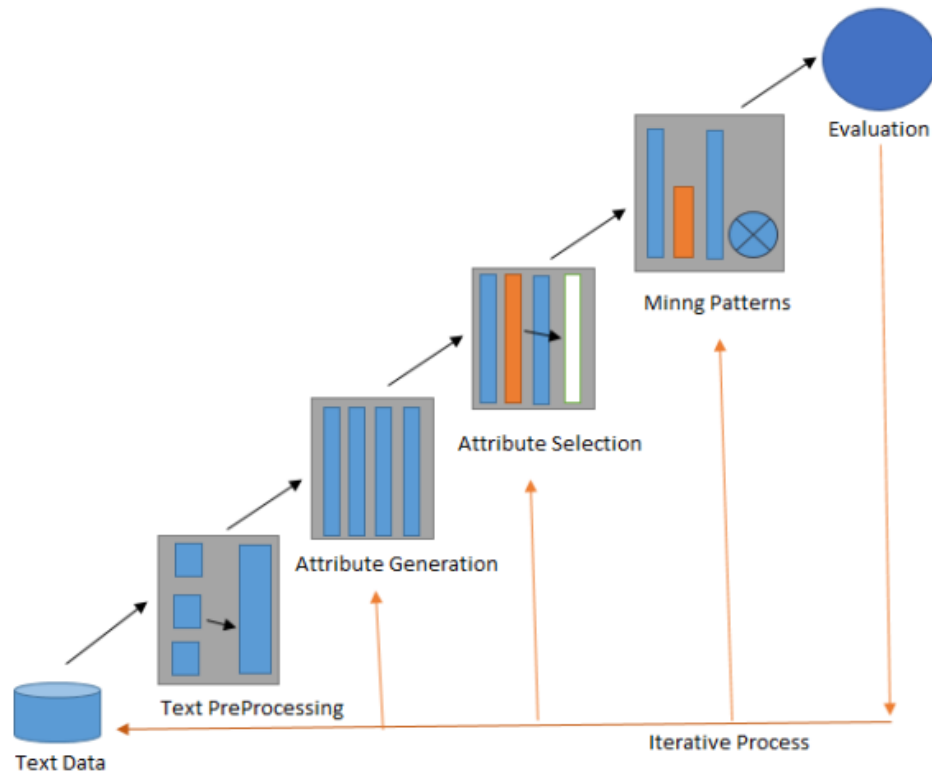


Fig: Text Mining Process

- **Text Pre-processing:**

It involves a series of steps,

- *Text Cleanup:*

Text Cleanup means removing of any unnecessary or unwanted information such as remove ads from web pages, normalize text converted from binary formats, deal with tables, figures, and formulas.

- *Tokenization:*

Tokenizing is simply achieved by splitting the text on white spaces and at punctuation marks that do not belong to abbreviations identified in the preceding step.

- *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. Rule-based approaches like ENGTWOL operate on dictionaries containing word forms together with the associated POS labels and morphological and syntactic features and context sensitive rules to choose the appropriate labels during application.

- **Text Transformation (Attribute Generation):**

A text document is represented by the words (features) it contains and their occurrences. The two main approaches of document representation are

- a) Bag of words
- b) Vector Space.

- **Feature Selection (Attribute Selection):**

Feature selection also known as variable selection, is the process of selecting a subset of important features for use in model creation. The main assumption when using a feature selection technique is that the data contain many redundant or irrelevant features.

Redundant features are the one which provides no extra information. Irrelevant features provide no useful or relevant information in any context. The major feature selection techniques are simple counting and statistics.

- **Data Mining:**

At this point the Text mining process merges with the traditional Data Mining process. Classic Data Mining techniques such as *Classification*, *Clustering*, *Association* are used in the structured database that resulted from the previous stages.

- **Evaluate:**

Evaluate the result, after evaluation the result can be discarded or the generated result can be used as an input for the next set of sequence.

Applications:

Text Mining can be applied in a variety of areas. Some of the most common areas are:

- Web Mining
- Medical
- Emails
- Resume Filtering