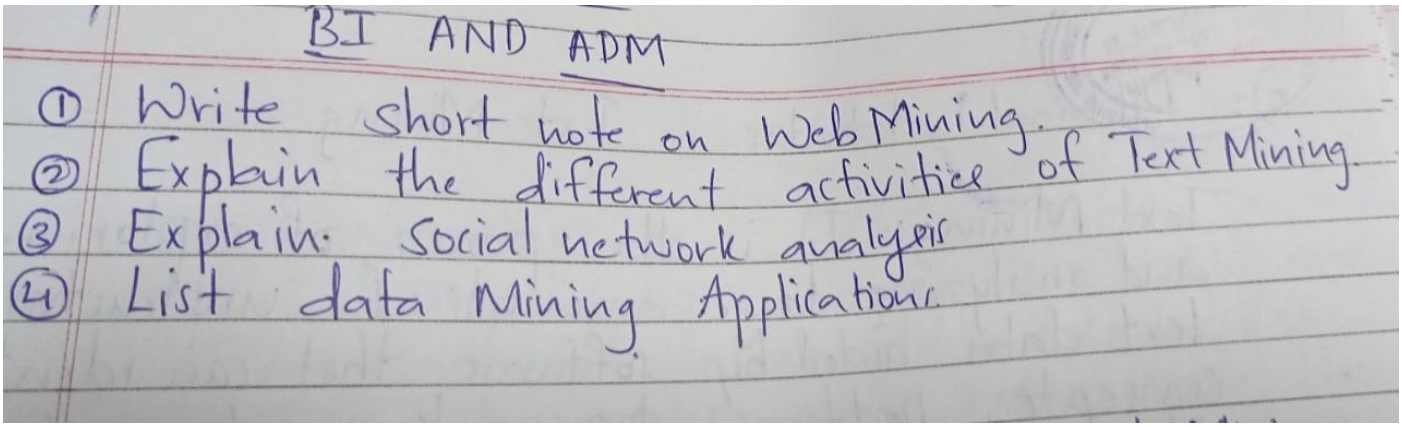


Graph Mining (GM) :

The task of graph mining is to extract patters (sub-graphs) of interest from graphs, that describe the underlying data and could be used further, e.g., for classification or clustering.

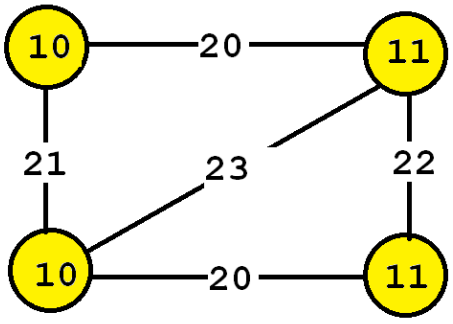
Graph mining has a vast number of applications, e.g. biological networks or web data.

Graph Mining (GM) is essentially the problem of discovering repetitive subgraphs occurring in the input graphs.

Graph mining is a process in which the mining techniques are used in finding a pattern or relationship in the given real-world collection of graphs.

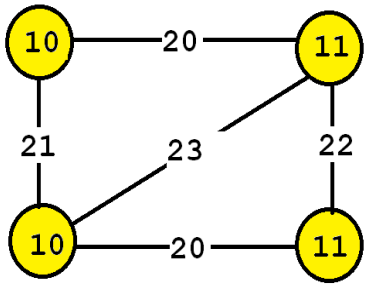
What is a graph?

- A graph is a set of vertices and edges, having some labels.
- This graph contains four vertices (depicted as yellow circles).
- These vertices have labels such as “10” and “11”.
- These labels provide information about the vertices.

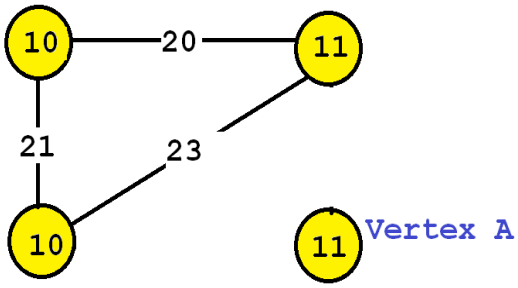


Types of graphs:

- Connected
- Disconnected

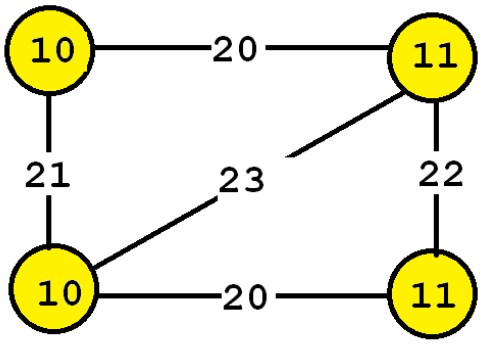


A connected graph

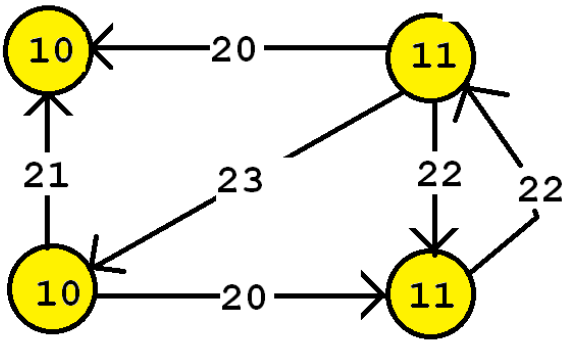


A disconnected graph

- Directed Graph
- Un-Directed Graph



An undirected graph



A directed graph

Differences:

- It is also useful to distinguish between directed and undirected graphs.
- In an undirected graph, edges are bidirectional, while in a directed graph, the edges can be unidirectional or bidirectional.
- Some data mining algorithms are designed to work only with undirected graphs, directed graphs, or support both

By mining the graph, frequent substructures and relationships can be identified.

It helps

- in clustering the graph sets,
- finding a relationship between graph sets,
- discriminating or characterizing graphs.
- Predicting these patterning trends can help in building models for the enhancement of any application that is used in real-time.

To implement the process of graph mining, one must learn to mine **frequent subgraphs**.

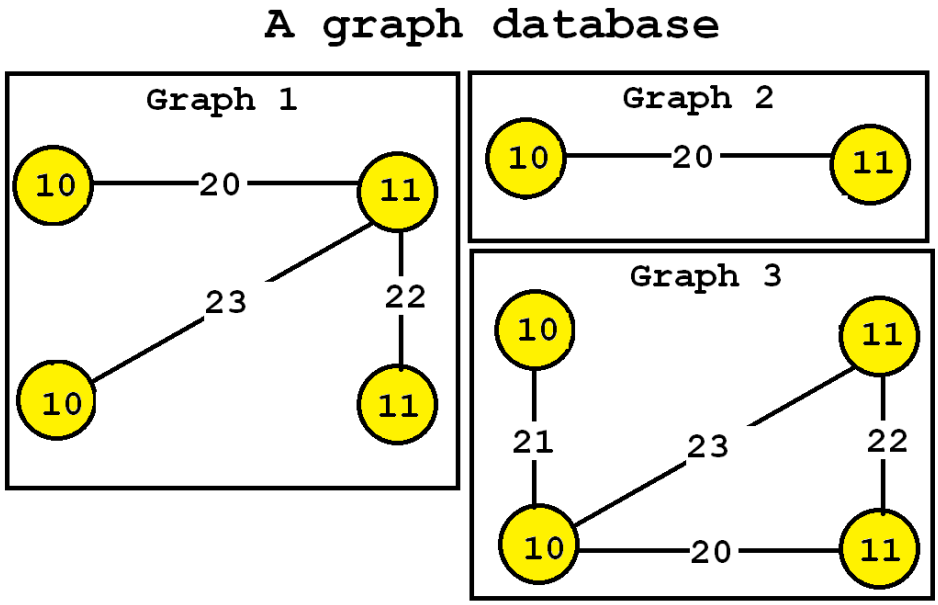
Frequent Subgraph Mining:

The task of finding frequent subgraphs in a set of graphs is called frequent subgraph mining.

As input the user must provide:

- A graph database (a set of graphs)
- A parameter called the minimum support threshold (minsup).

- Then, a frequent subgraph mining algorithm will enumerate as output all frequent subgraphs.
- A frequent subgraph is a subgraph that appears in at least minsup graphs from a graph database.



Steps in finding frequent subgraphs:

There are two steps in finding frequent subgraphs.

- The first step is to create frequent substructure candidates.
- The second step is to find the support of each and every candidate. We must optimize and enhance the first step because the second step is an NP-completed set where the computational complexity is accurate and high.

There are two methods for frequent substructure mining.

- **The Apriori-based approach:**
- **The Pattern- growth approach:**

The Apriori-based approach:

- The approach to find the frequent graphs begin from the graph with a small size.

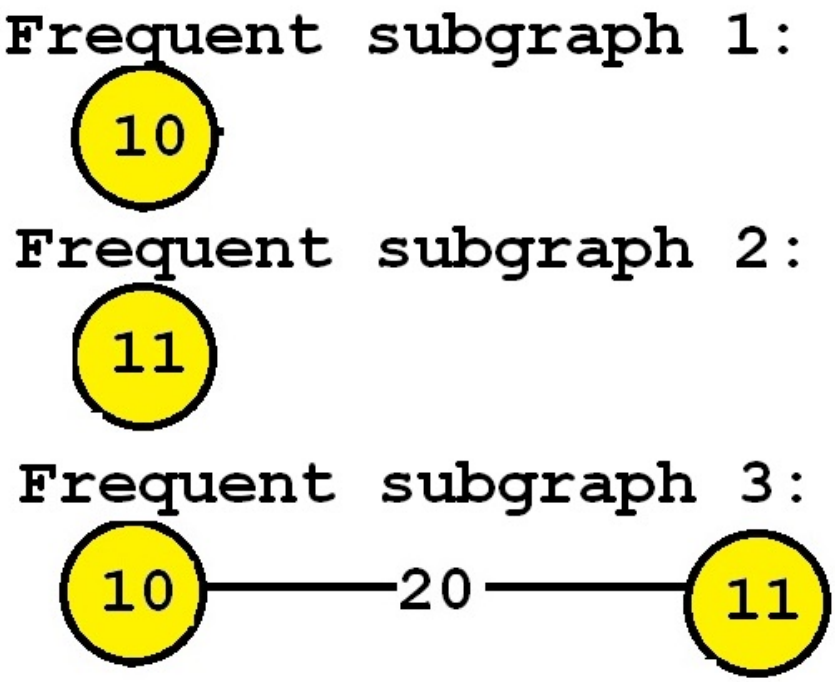
- The approach advances in a bottom-up way by creating candidates with extra vertex or edge.
- This algorithm is called an **Apriori Graph**.

The Pattern- growth approach:

- This pattern-growth approach can use both BFS and DFS(Depth First Search).
- DFS is preferred for this approach due to its less memory consumption nature
- The duplicate graphs generated can be removed but it increases the time and work.
- To avoid the creation of duplicate graphs, the frequent graphs should be introduced very carefully and conservatively which calls the need for other algorithms.

From Database…

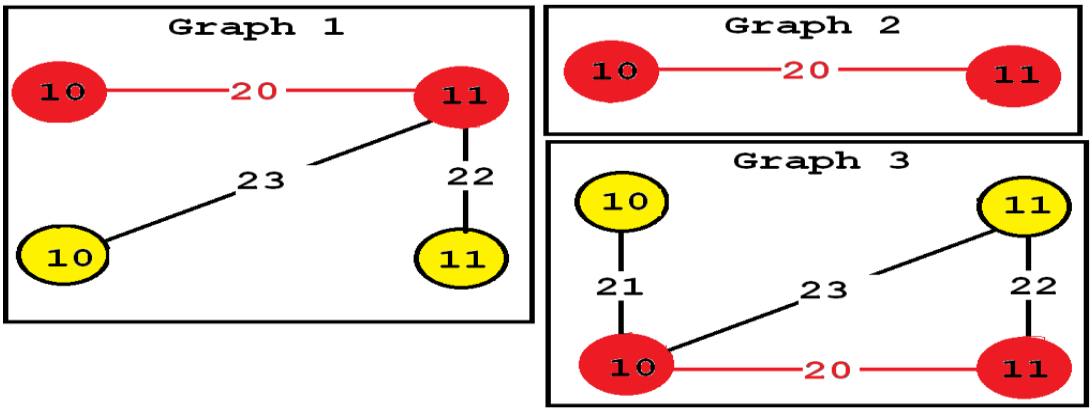
- Now, let's say that we want to discover all subgraphs that appear in at least three graphs.
- By applying a **frequent subgraph mining algorithm**, we will obtain the set of all subgraphs appearing in at least three graphs:



Consider the third subgraph ("Frequent subgraph 3").

- This subgraph is frequent and is said to have a support (a frequency) of 3 since it appears in three of the input graphs.

- These occurrences are highlighted in red, below:



Social Networks Analysis

Social Network Analysis (SNA) is the process of exploring or examining the social structure by using graph theory.

It is used for measuring and analyzing the structural properties of the network.

It helps to measure relationships and flows between groups, organizations, and other connected entities. We need specialized tools to study and analyze social networks.

Basically, there are two types of social networks:

- Ego network Analysis
- Complete network Analysis

1. Ego Network Analysis

- Ego network Analysis is the one **that finds the relationship among people**.
- The analysis is done for a particular sample of people chosen from the whole population.
- This **sampling is done randomly** to analyze the relationship.
- The attributes involved in this ego network analysis are a person's size, diversity, etc.

- This analysis is done by traditional surveys.
- The surveys involve that they people are asked with whom they interact with and their name of the relationship between them.
- It is not focused to find the relationship between everyone in the sample.
- It is an effort to find the density of the network in those samples.
- This hypothesis is tested using some statistical hypothesis testing techniques.

The following functions are served by Ego Networks:

- Propagation of information efficiently.
- Sensemaking from links, For example, Social links, relationships.

- Access to resources, efficient connection path generation.
- Community detection, identification of the formation of groups.
- Analysis of the ties among individuals for social support.

2. Complete Network Analysis

- Complete network analysis is the analysis that is used in all network analyses. It analyses the relationship among the sample of people chosen from the large population.
- Subgroup analysis, centrality measure, and equivalence analysis are based on the complete network analysis. This analysis measure helps the organization or the company to make any decision with the help of their relationship.
- Testing the sample will show the relationship in the whole network since the sample is taken from a single set of domains.

Difference between Ego network analysis and Complete network analysis:

The difference between ego and complete network analysis is that the ego network focus on collecting the relationship of people in the sample with the outside world whereas, in Complete network, it is focused on finding the relationship among the samples.

The majority of the network analysis will be done only for a particular domain or one organization. It is not focused on the relationships between the organization. So many of the social network analysis measure uses only Complete network analysis.

Text Data 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, documents, emails, files are written in common language text. Text mining is primarily used to draw useful insights or patterns from such data.

Text databases consist of huge collection of documents. They collect these information from several sources such as news articles, books, digital libraries, e-mail messages, web pages, etc. Due to increase in the amount of information, the text databases are growing rapidly. In many of the text databases, the data is semi-structured.

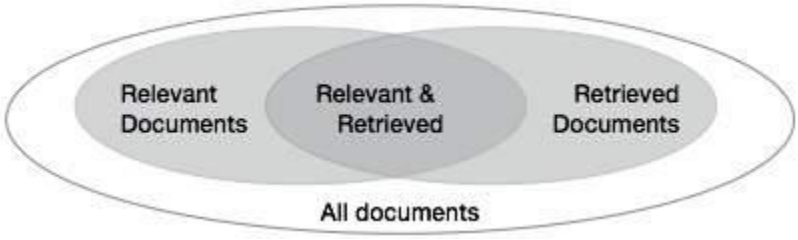
For example, a document may contain a few structured fields, such as title, author, publishing_date, etc. But along with the structure data, the document also contains unstructured text components, such as abstract and contents. Without knowing what could be in the documents, it is difficult to formulate effective queries for analyzing and extracting useful information from the data. Users require tools to compare the documents and rank their importance and relevance. Therefore, text mining has become popular and an essential theme in data mining

The text mining market has experienced exponential growth and adoption over the last few years and also expected to gain significant growth and adoption in the coming future. One of the primary reasons behind the adoption of text mining is higher competition in the business market, many organizations seeking value-

added solutions to compete with other organizations. With increasing completion in business and changing customer perspectives, organizations are making huge investments to find a solution that is capable of analyzing customer and competitor data to improve competitiveness. The primary source of data is e-commerce websites, social media platforms, published articles, survey, and many more.

Basic Measures for Text Retrieval

We need to check the accuracy of a system when it retrieves a number of documents on the basis of user's input. Let the set of documents relevant to a query be denoted as {Relevant} and the set of retrieved document as {Retrieved}. The set of documents that are relevant and retrieved can be denoted as {Relevant} ∩ {Retrieved}. This can be shown in the form of a Venn diagram as follows −

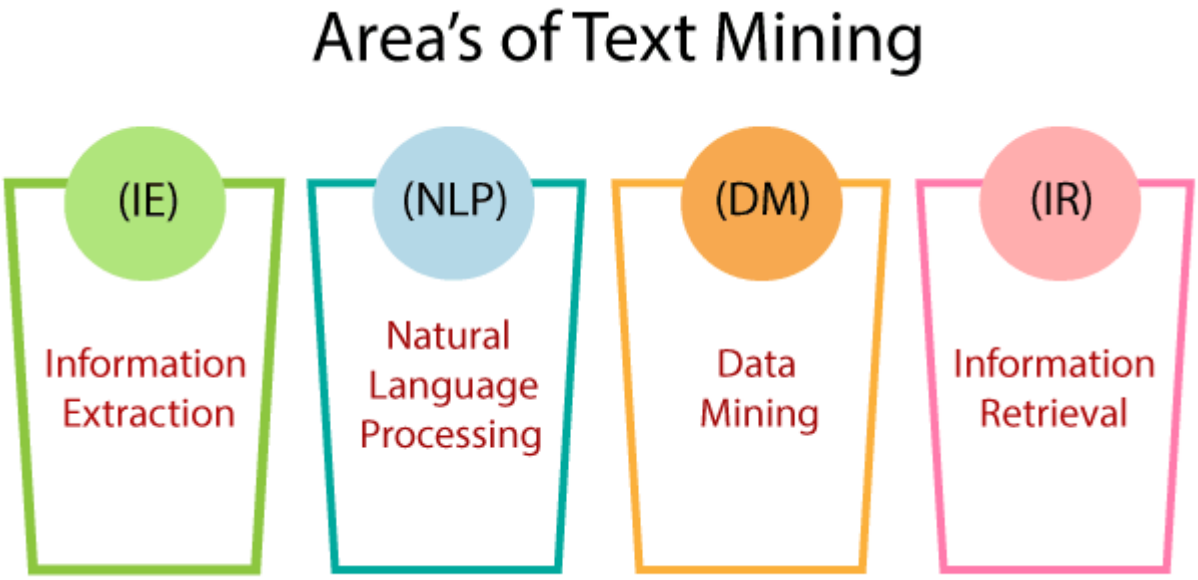


There are three fundamental measures for assessing the quality of text retrieval −

- Precision
- Recall
- F-score

Areas of text mining in data mining:

These are the following area of text mining :



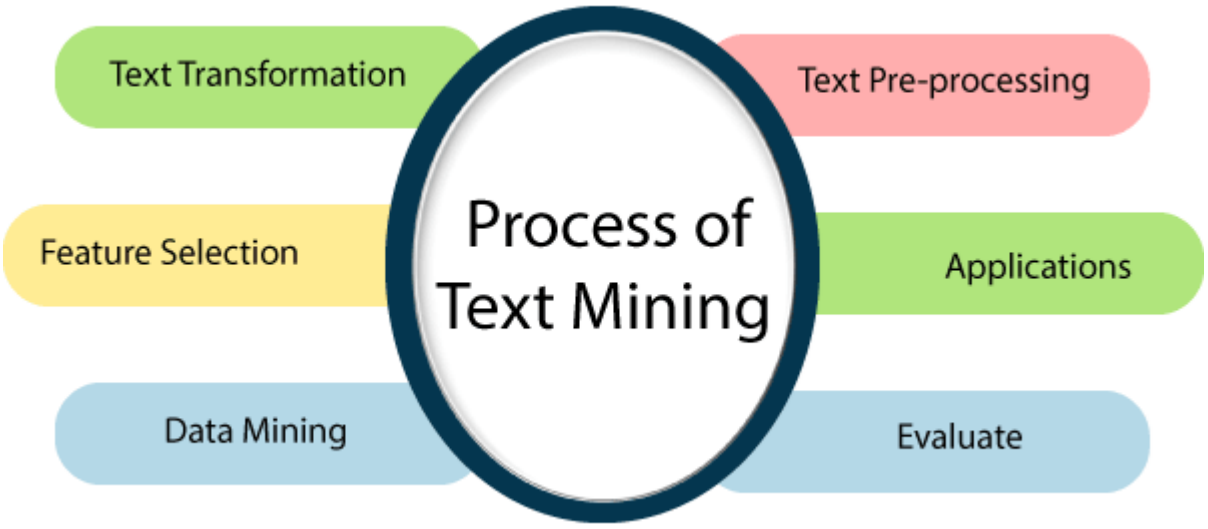
- **Information Extraction:**
  The automatic extraction of structured data such as entities, entities relationships, and

attributes describing entities from an unstructured source is called information extraction.

- **Natural Language Processing:**
  NLP stands for Natural language processing. Computer software can understand human language as same as it is spoken. NLP is primarily a component of artificial intelligence(AI). The development of the NLP application is difficult because computers generally expect humans to "Speak" to them in a programming language that is accurate, clear, and exceptionally structured. Human speech is usually not authentic so that it can depend on many complex variables, including slang, social context, and regional dialects.
- **Data Mining:**
  Data mining refers to the extraction of useful data, hidden patterns from large data sets. Data mining tools can predict behaviors and future trends that allow businesses to make a better data-driven decision. Data mining tools can be used to resolve many business problems that have traditionally been too time-consuming.
- **Information Retrieval:**
  Information retrieval deals with retrieving useful data from data that is stored in our systems. Alternately, as an analogy, we can view search engines that happen on websites such as e-commerce sites or any other sites as part of information retrieval.

Text Mining Process:

The text mining process incorporates the following steps to extract the data from the document.



- **Text transformation**
  A text transformation is a technique that is used to control the capitalization of the text.
  Here the two major way of document representation is given.
  1. Bag of words
  2. Vector Space
- **Text Pre-processing**
  Pre-processing is a significant task and a critical step in Text Mining, Natural

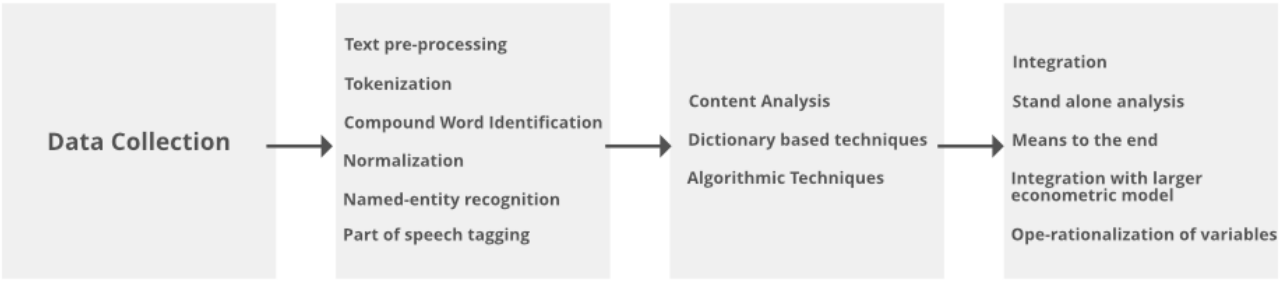
Language Processing (NLP), and information retrieval(IR). In the field of text mining, data pre-processing is used for extracting useful information and knowledge from unstructured text data. Information Retrieval (IR) is a matter of choosing which documents in a collection should be retrieved to fulfill the user's need.

- **Feature selection:**
  Feature selection is a significant part of data mining. Feature selection can be defined as the process of reducing the input of processing or finding the essential information sources. The feature selection is also called variable selection.
- **Data Mining:**
  Now, in this step, the text mining procedure merges with the conventional process. Classic Data Mining procedures are used in the structural database.
- **Evaluate:**
  Afterward, it evaluates the results. Once the result is evaluated, the result abandon.
- **Applications: These are the following text mining applications:**

  - **Risk Management:**
    Risk Management is a systematic and logical procedure of analyzing, identifying, treating, and monitoring the risks involved in any action or process in organizations. Insufficient risk analysis is usually a leading cause of disappointment. It is particularly true in the financial organizations where adoption of Risk Management Software based on text mining technology can effectively enhance the ability to diminish risk. It enables the administration of millions of sources and petabytes of text documents, and giving the ability to connect the data. It helps to access the appropriate data at the right time.
  - **Customer Care Service:**
    Text mining methods, particularly NLP, are finding increasing significance in the field of customer care. Organizations are spending in text analytics programming to improve their overall experience by accessing the textual data from different sources such as customer feedback, surveys, customer calls, etc. The primary objective of text analysis is to reduce the response time of the organizations and help to address the complaints of the customer rapidly and productively.
  - **Business Intelligence:**
    Companies and business firms have started to use text mining strategies as a major aspect of their business intelligence. Besides providing significant insights into customer behavior and trends, text mining strategies also support organizations to analyze the qualities and weaknesses of their opponent's so, giving them a competitive advantage in the market.
  - **Social Media Analysis:**
    Social media analysis helps to track the online data, and there are numerous text mining tools designed particularly for performance analysis of social media sites. These tools help to monitor and interpret the text generated via the internet from the news, emails, blogs, etc. Text mining tools can precisely analyze the total no of posts, followers, and total no of likes of your brand on a social media platform that enables you to understand the response of the individuals who are interacting with your brand and content.

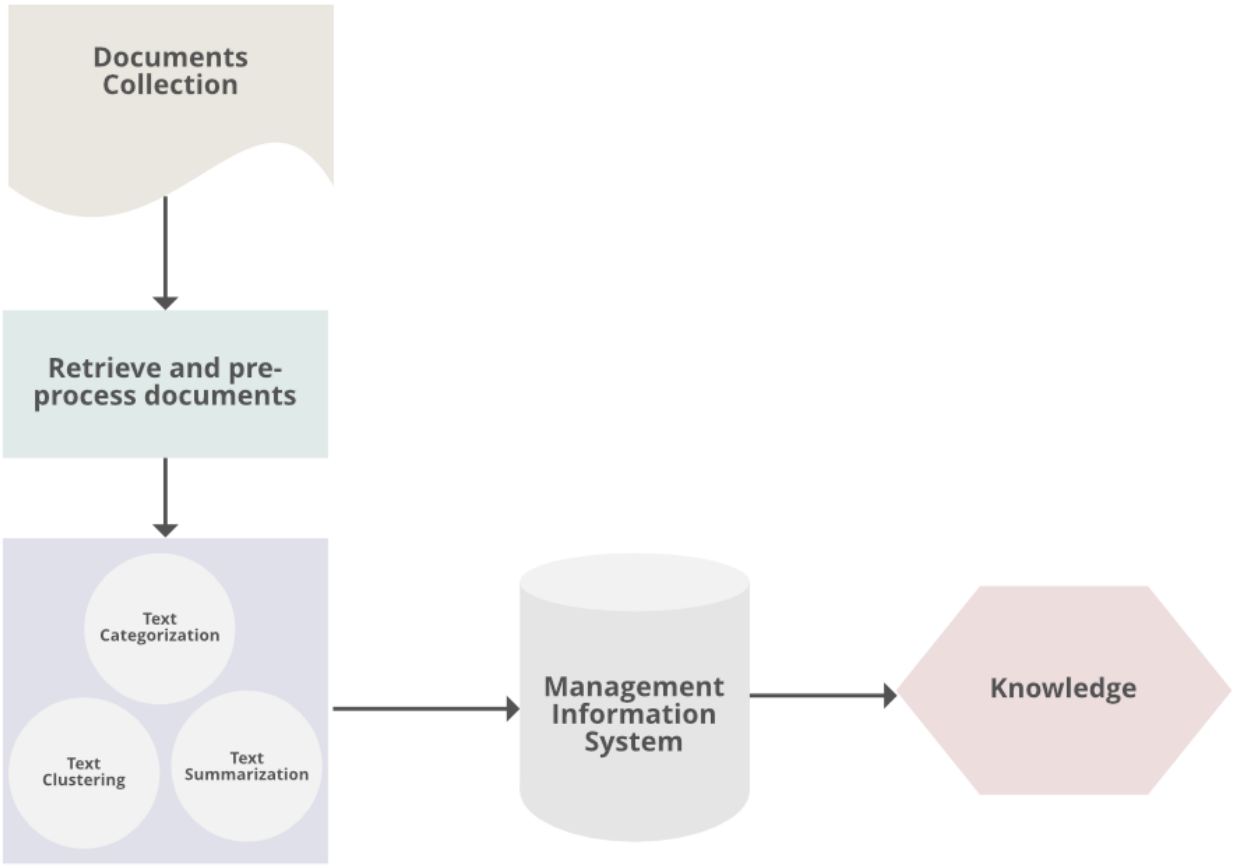
The conventional process of text mining as follows:

- Gathering unstructured information from various sources accessible in various document organizations, for example, plain text, web pages, PDF records, etc.
- Pre-processing and data cleansing tasks are performed to distinguish and eliminate inconsistency from the data. The data cleansing process makes sure to capture the genuine text, and it is performed to eliminate stop words stemming (the process of identifying the root of a certain word and indexing the data.
- Processing and controlling tasks are applied to review and further clean the data set.
- Pattern analysis is implemented in Management Information System.
- Information processed in the above steps is utilized to extract important and applicable data for a powerful and convenient decision-making process and trend analysis.



Procedures of analyzing Text Mining:

- **Text Summarization:** To extract its partial content reflection its whole content automatically.
- **Text Categorization:** To assign a category to the text among categories predefined by users.
- **Text Clustering:** To segment texts into several clusters, depending on the substantial relevance.

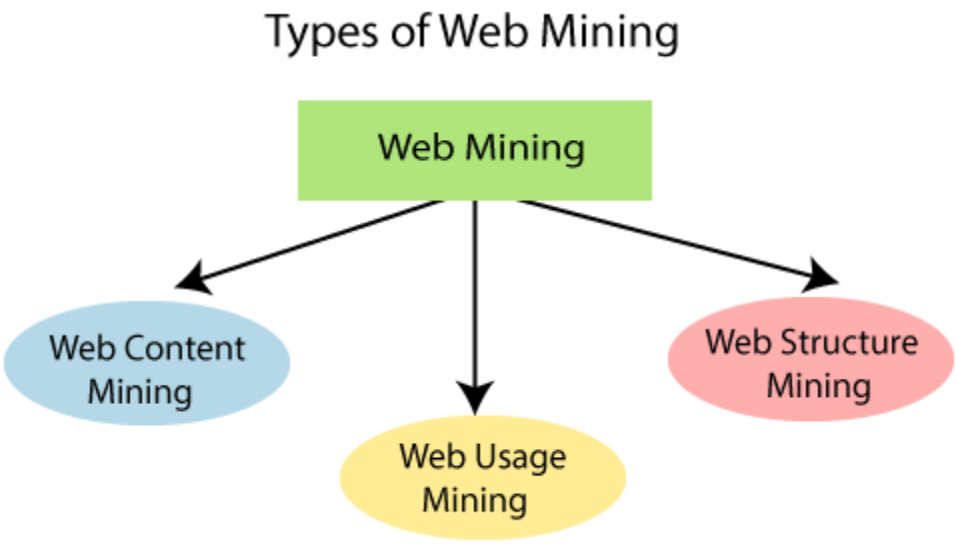


TextMining Techniques:

- **Information Extraction**: It is a process of extract meaningful words from documents.
- **Information Retrieval**: It is a process of extracting relevant and associated patterns according to a given set of words or text documents.
- **Natural Language Processing:** It concerns the automatic processing and analysis of unstructured text information.
- **Clustering:** It is an unsupervised learning process that grouping of text according to their similar characteristics.
- **Text Summarization:** To extract its partial content reflection it's whole content automatically.

WEB MING:

Web mining can widely be seen as the application of adapted data mining techniques to the web, whereas data mining is defined as the application of the algorithm to discover patterns on mostly structured data embedded into a **knowledge discovery process**. Web mining has a distinctive property to provide a set of various data types. The web has multiple aspects that yield different approaches for the mining process, such as web pages consist of text, web pages are linked via hyperlinks, and user activity can be monitored via web server logs. These three features lead to the differentiation between the three areas are web content mining, web structure mining, web usage mining.



Web Content Mining:

Web content mining can be used to extract useful data, information, knowledge from the web page content. In web content mining, each web page is considered as an individual document. The individual can take advantage of the semi-structured nature of web pages, as HTML provides information that concerns not only the layout but also logical structure. The primary task of content mining is data extraction, where structured data is extracted from unstructured websites. The objective is to facilitate data aggregation over various web sites by using the extracted structured data. Web content mining can be utilized to distinguish topics on the web. For Example, if any user searches for a specific task on the search engine, then the user will get a list of suggestions.

2. Web Structured Mining:

The web structure mining can be used to find the link structure of hyperlink. It is used to identify that data either link the web pages or direct link network. In Web Structure Mining, an individual considers the web as a directed graph, with the web pages being the vertices that are associated with hyperlinks. The most important application in this regard is the Google search engine, which estimates the ranking of its outcomes primarily with the PageRank algorithm. It characterizes a page to be exceptionally relevant when frequently connected by other highly related pages. Structure and content mining methodologies are usually combined. For example, web structured mining can be beneficial to organizations to regulate the network between two commercial sites.

3. Web Usage Mining:

Web usage mining is used to extract useful data, information, knowledge from the weblog records, and assists in recognizing the user access patterns for web pages. In Mining, the usage of web resources, the individual is thinking about records of requests of visitors of a website, that are often collected as web server logs. While the content and structure of the collection of web pages follow the intentions of the authors of the pages, the individual requests demonstrate how the consumers see these pages. Web usage mining may disclose relationships that were not proposed by the creator of the pages.

Data Visualization

Data visualization is a graphical representation of quantitative information and data by using visual elements like graphs, charts, and maps.

Data visualization convert large and small data sets into visuals, which is easy to understand and process for humans.

Data visualization tools provide accessible ways to understand outliers, patterns, and trends in the data.

In the world of Big Data, the data visualization tools and technologies are required to analyze vast amounts of information.

Data visualizations are common in your everyday life, but they always appear in the form of graphs and charts. The combination of multiple visualizations and bits of information are still referred to as Infographics.

Data visualizations are used to discover unknown facts and trends. You can see visualizations in the form of line charts to display change over time. Bar and column charts are useful for observing relationships and making comparisons. A pie chart is a great way to show parts-of-a-whole. And maps are the best way to share geographical data visually.

Characteristics of Effective Graphical Visual :

- It shows or visualizes data very clearly in an understandable manner.
- It encourages viewers to compare different pieces of data.
- It closely integrates statistical and verbal descriptions of data set.
- It grabs our interest, focuses our mind, and keeps our eyes on message as human brain tends to focus on visual data more than written data.
- It also helps in identifying area that needs more attention and improvement.
- Using graphical representation, a story can be told more efficiently. Also, it requires less time to understand picture than it takes to understand textual data.

Data visualization have some more specialties such as:

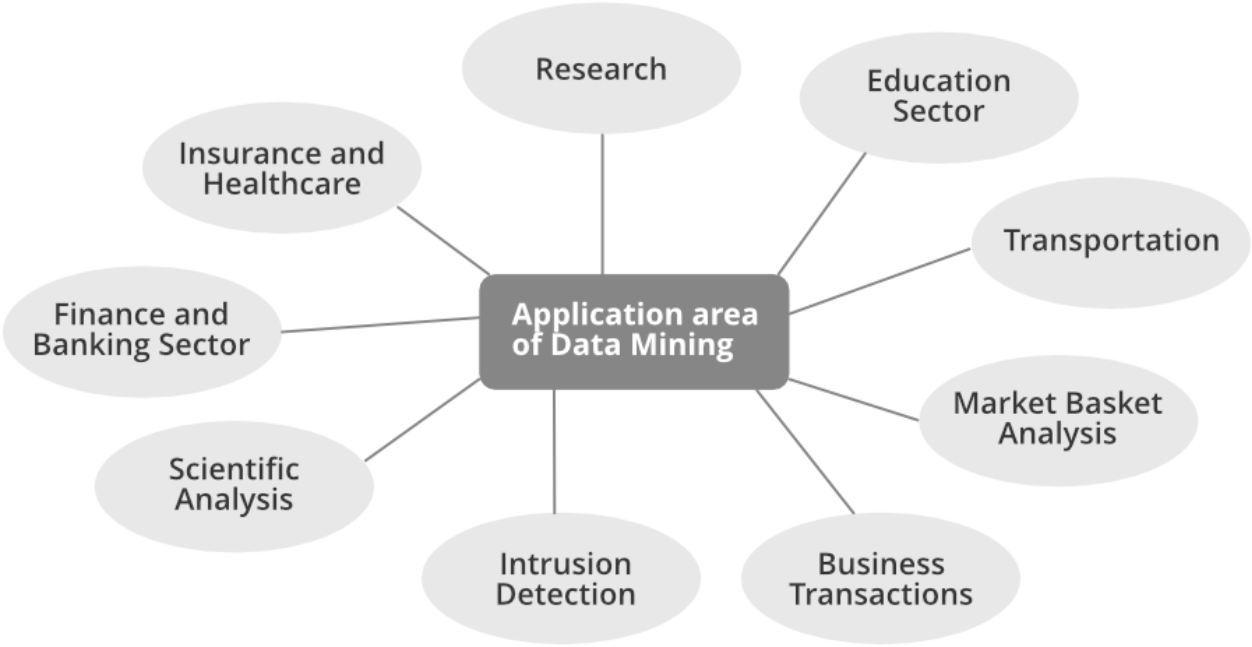
- Data visualization can identify areas that need improvement or modifications.
- Data visualization can clarify which factor influence customer behavior.
- Data visualization helps you to understand which products to place where.
- Data visualization can predict sales volumes.

Data visualization tools have been necessary for democratizing data, analytics, and making data-driven perception available to workers throughout an organization. They are easy to operate in comparison to earlier versions of BI software or traditional statistical analysis software.

Use Data Visualization

1. To make easier in understand and remember.
2. To discover unknown facts, outliers, and trends.
3. To visualize relationships and patterns quickly.
4. To ask a better question and make better decisions.
5. To competitive analyse.
6. To improve insights.

APPLICATION OF DATA MINIG :



https://www.geeksforgeeks.org/applications-of-data-mining/

https://www.tutorialspoint.com/data_mining/dm_applications_trends.htm