# ACTIVIDAD #1

## Tipo actividad: Reading comprehension: "What is Text Mining?" and related activities

### 3) Reading comprehension: "What is Text Mining?"

### What is text mining?

Text mining, also known as text data mining, is the process of transforming unstructured text into a structured format to identify meaningful patterns and new insights. By applying advanced analytical techniques, such as Naïve Bayes, Support Vector Machines (SVM), and other deep learning algorithms, companies are able to explore and discover hidden relationships within their unstructured data.

Text is one of the most common data types within databases. Depending on the database, this data can be organized as:

**Structured data:** This data is standardized into a tabular format with numerous rows and columns, making it easier to store and process for analysis and machine learning algorithms. Structured data can include inputs such as names, addresses, and phone numbers.

**Unstructured data:** This data does not have a predefined data format. It can include text from sources, like social media or product reviews, or rich media formats like, video and audio files.

**Semi-structured data**: As the name suggests, this data is a blend between structured and unstructured data formats. While it has some organization, it doesn't have enough structure to meet the requirements of a relational database. Examples of semi-structured data include XML, JSON and HTML files.

Since roughly 80% of data in the world resides in an unstructured format (link resides outside ibm.com), text mining is an extremely valuable practice within organizations. Text mining tools and natural language processing (NLP) techniques, like information extraction (link reside outside of IBM), allow us to transform unstructured documents into a structured format to enable analysis and the generation of high-quality insights. This, in turn, improves the decision-making of organizations, leading to better business outcomes.

**Begin your journey to AI**

## Text mining vs. text analytics

The terms, text mining and text analytics, are largely synonymous in meaning in conversation, but they can have a more nuanced meaning. Text mining and text analysis identifies textual patterns and trends within unstructured data through the use of machine learning, statistics, and linguistics. By transforming the data into a more structured format through text mining and text analysis, more quantitative insights can be found through text analytics. Data visualization techniques can then be harnessed to communicate findings to wider audiences.

## Text mining techniques

The process of text mining comprises several activities that enable you to deduce information from unstructured text data. Before you can apply different text mining techniques, you must start with text preprocessing, which is the practice of cleaning and transforming text data into a usable format. This practice is a core aspect of natural language processing (NLP) and it usually involves the use of techniques such as language identification, tokenization, part-of-speech tagging, chunking, and syntax parsing to format data appropriately for analysis. When text preprocessing is complete, you can apply text mining algorithms to derive insights from the data. Some of these common text mining techniques include:

### Information retrieval

Information retrieval (IR) returns relevant information or documents based on a pre-defined set of queries or phrases. IR systems utilize algorithms to track user behaviors and identify relevant data. Information retrieval is commonly used in library catalogue systems and popular search engines, like Google. Some common IR sub-tasks include:

**Tokenization:** This is the process of breaking out long-form text into sentences and words called "tokens". These are, then, used in the models, like bag-of-words, for text clustering and document matching tasks.

**Stemming:** This refers to the process of separating the prefixes and suffixes from words to derive the root word form and meaning. This technique improves information retrieval by reducing the size of indexing files.

### Natural language processing (NLP)

Natural language processing, which evolved from computational linguistics, uses methods from various disciplines, such as computer science, artificial intelligence, linguistics, and data science, to enable computers to understand human language in both written and verbal forms. By analyzing sentence structure and grammar, NLP sub-tasks allow computers to "read". Common sub-tasks include:

**Summarization:** This technique provides a synopsis of long pieces of text to create a concise, coherent summary of a document's main points.

**Part-of-Speech (PoS) tagging:** This technique assigns a tag to every token in a document based on its part of speech—i.e. denoting nouns, verbs, adjectives, etc. This step enables semantic analysis on unstructured text.

**Text categorization:** This task, which is also known as text classification, is responsible for analyzing text documents and classifying them based on predefined topics or categories. This sub-task is particularly helpful when categorizing synonyms and abbreviations.

**Sentiment analysis:** This task detects positive or negative sentiment from internal or external data sources, allowing you to track changes in customer attitudes over time. It is commonly used to provide information about perceptions of brands, products, and services. These insights can propel businesses to connect with customers and improve processes and user experiences.

### Information extraction

Information extraction (IE) surfaces the relevant pieces of data when searching various documents. It also focuses on extracting structured information from free text and storing these entities, attributes, and relationship information in a database. Common information extraction sub-tasks include:

**Feature selection**, or attribute selection, is the process of selecting the important features (dimensions) to contribute the most to output of a predictive analytics model.

**Feature extraction** is the process of selecting a subset of features to improve the accuracy of a classification task. This is particularly important for dimensionality reduction.

**Named-entity recognition** (NER) also known as entity identification or entity extraction, aims to find and categorize specific entities in text, such as names or locations. For example, NER identifies "California" as a location and "Mary" as a woman's name.

### Data mining

Data mining is the process of identifying patterns and extracting useful insights from big data sets. This practice evaluates both structured and unstructured data to identify new information, and it is commonly utilized to analyze consumer behaviors within marketing and sales. Text mining is essentially a sub-field of data mining as it focuses on bringing structure to unstructured data and analyzing it to generate novel insights. The techniques mentioned above are forms of data mining but fall under the scope of textual data analysis.

### Text mining applications

Text analytics software has impacted the way that many industries work, allowing them to improve product user experiences as well as make faster and better business decisions. Some use cases include:

**Customer service:** There are various ways in which we solicit customer feedback from our users. When combined with text analytics tools, feedback systems, such as chatbots, customer surveys, NPS (net-promoter scores), online reviews, support tickets, and social media profiles, enable companies to improve their customer experience with speed. Text mining and sentiment analysis can provide a mechanism for companies to prioritize key pain points for their customers, allowing businesses to respond to urgent issues in real-time and increase customer satisfaction. Learn how Verizon is using text analytics in customer service.

**Risk management:** Text mining also has applications in risk management, where it can provide insights around industry trends and financial markets by monitoring shifts in sentiment and by extracting information from analyst reports and whitepapers. This is particularly valuable to banking institutions as this data provides more confidence when considering business investments across various sectors. Learn how CIBC and EquBot are using text analytics for risk mitigation.

**Maintenance:** Text mining provides a rich and complete picture of the operation and functionality of products and machinery. Over time, text mining automates decision making by revealing patterns that correlate with problems and preventive and reactive maintenance procedures. Text analytics helps maintenance professionals unearth the root cause of challenges and failures faster.

**Healthcare:** Text mining techniques have been increasingly valuable to researchers in the biomedical field, particularly for clustering information. Manual investigation of medical research can be costly and time-consuming; text mining provides an automation method for extracting valuable information from medical literature.

**Spam filtering:** Spam frequently serves as an entry point for hackers to infect computer systems with malware. Text mining can provide a method to filter and exclude these emails from inboxes, improving the overall user experience and minimizing the risk of cyber-attacks to end users.

**4) True/False activity.**

**1.** Text mining is the process of converting structured data into an unstructured format to explore hidden relationships.

- **True / False**

**2.** Named-Entity Recognition (NER) is a text mining technique that categorizes specific entities in text, such as names or locations.

- **True / False**

**3.** Information retrieval is a text mining technique that focuses on breaking long-form text into sentences and words called "tokens."

- **True / False**

**4.** Text analytics is a distinct process from text mining and involves transforming data into a more structured format using machine learning, statistics, and linguistics.

- **True / False**

**5.** Sentiment analysis is a text mining task that detects positive or negative sentiment from internal or external data sources.

- **True / False**

**5) Fill in the blank vocabulary activity.**

**1.** Text mining is also known as text data mining, and it involves transforming unstructured text into a structured format to identify meaningful patterns and new insights by applying advanced analytical techniques such as Naïve Bayes, Support Vector Machines (SVM), and other deep learning algorithms. This process helps companies explore and discover hidden relationships within their ____________ data.

**2.** Structured data is standardized into a tabular format with numerous rows and columns, making it easier to store and process for analysis and machine learning algorithms. Examples of structured data include names, addresses, and phone numbers, while ____________ data has some organization but doesn't have enough structure to meet the requirements of a relational database.

**3.** Text preprocessing is a core aspect of natural language processing (NLP) and involves cleaning and transforming text data into a usable format. It includes techniques such as language identification, tokenization, part-of-speech tagging, chunking, and syntax parsing to format data appropriately for ____________.

**4.** Information retrieval (IR) returns relevant information or documents based on pre-defined queries or phrases. Tokenization, the process of breaking out long-form text into sentences and words called "tokens," is one of the common sub-tasks in ____________ that involves using models like bag-of-words for text clustering and document matching tasks.

**5.** Data mining is the process of identifying patterns and extracting useful insights from big data sets. While text mining falls under the scope of textual data analysis, data mining evaluates both structured and unstructured data to identify new information. Text mining is essentially a ____________ of data mining.