UNIT - II
Text Analytics and Text Mining:

1. Machine Versus Men on Jeopardy!: The Story of Watson
2. Text Analytics and Text Mining Concepts and Definitions
3. Natural Language Processing
4. Text Mining Applications
5. Text Mining Tools.
6. Text Mining Process

1) MACHINE VERSUS MEN ON JEOPARDY!: THE STORY OF WATSON

- **Watson Overview**
- **Background**
- **Competing Against The Best**
- **How Does Watson Do It?**
- **High Level Depiction Of DEEPQA Architecture**
- **Conclusion**

Watson Overview

- Watson is an advanced computer system designed to answer natural human language questions.
- Developed in 2010 by IBM Research for the DeepQA project, named after IBM's first president, Thomas J. Watson.

Background

- IBM Research pursued a challenge to rival Deep Blue and advance computer science, benefiting science, business, and society.
- The challenge: build a real-time Jeopardy! contestant capable of listening, understanding, and responding.

Competing Against the Best

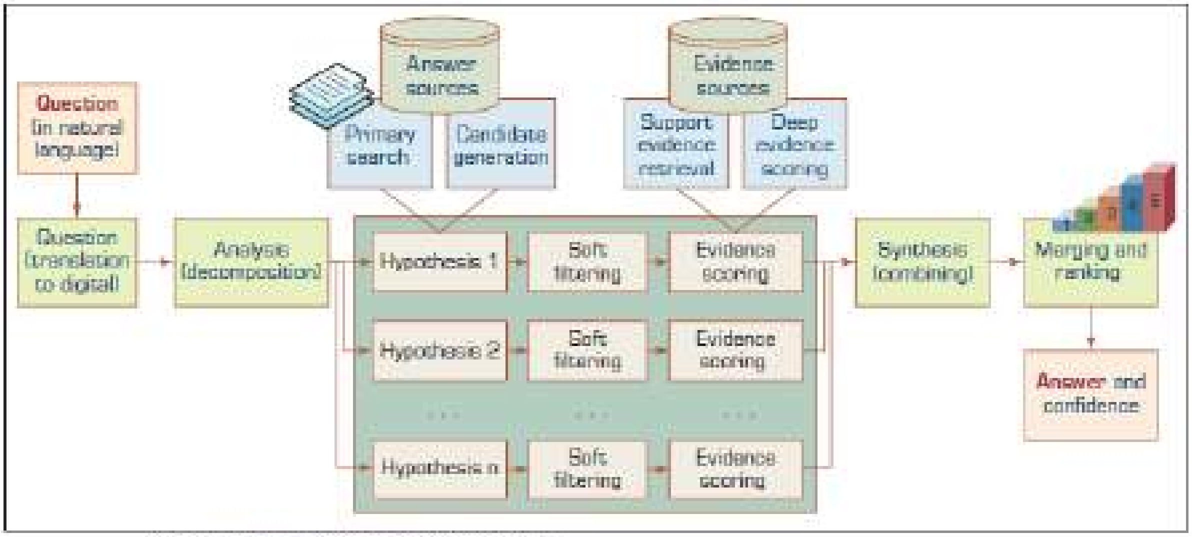
- In 2011, Watson competed in Jeopardy!'s first human-versus-machine match.
- Watson won a two-game match, defeating top players Brad Rutter and Ken Jennings.

- Watson excelled in signaling but had trouble with short, few-word clues.(In the *Jeopardy!* match, Watson was very fast at pressing the buzzer, which gave it an advantage over the human players. However, Watson had trouble understanding clues that were very short and only had a few words. These types of clues were harder for Watson to figure out, so it didn't do as well in those parts of the game.)
- It processed 200 million pages of data using four terabytes of storage without internet access.
- Watson employed multiple QA technologies like text mining, NLP, and question-answering techniques.
- It computed confidence levels to decide whether to "buzz in" within 1-6 seconds, averaging 3 seconds.

HOW DOES WATSON DO IT?

- Watson's system, called DeepQA, is a powerful, parallel-processing architecture that uses over 100 techniques to analyze language, generate and evaluate answers, and rank them. The key to its success lies in how these techniques work together to improve accuracy, speed, and confidence.
- **Key principles of DeepQA**:
- **Massive parallelism**: It examines multiple interpretations and answers at once.
- **Many experts**: Uses a range of probabilistic methods for analyzing questions and content.
- **Pervasive confidence estimation**: Each part of the system scores potential answers with confidence levels, which are combined to find the best answer.
- **Shallow and deep knowledge**: It balances between detailed and broader understandings of language and knowledge structures.

A HIGH LEVEL DEPICTION OF DEEPQA ARCHITECTURE



A High-Level Depiction of DeepQA Architecture.

Working

- **Question (Natural Language)**:)(The process starts with a regular question.A regular question, like "Who wrote 'Pride and Prejudice'?", is asked.
- **Question (Translation to Digital)**: The question is converted into a computer-friendly format by breaking it into tokens and analyzing meaning.
- **Decomposition**: The system breaks the question down into smaller parts (e.g., subject, verb, noun phrases).
- **Hypotheses Generation**: The system generates multiple possible answers.
- **Answer Sources**: It searches documents, websites, or databases for relevant information.
- **Primary search:** This involves searching through a vast corpus of text documents or knowledge bases to identify relevant information that might be useful for answering the given question.
- **Candidate generation:** Once relevant documents are identified, the system generates potential candidates or answers based on the information contained within those documents.
- **Soft filtering** is a way to remove unlikely or irrelevant answers from a list of possible answers in a DeepQA system. It uses initial analysis to decide which answers are more likely to be correct.
- **Soft Filtering**: Unlikely answers are removed.

- **Evidence Scoring**: The remaining answers are ranked based on evidence.(**Evidence Scoring:** The evidence supporting each remaining answer is evaluated and scored to determine its strength.
- **Synthesis and Merging**: The strongest answers are combined.
- **Answer and Confidence**: watson selects The best answer , and gives with confidence level.
- So in short, Watson takes a question, breaks it down, finds potential answers, backs them up with evidence, and then picks the most reliable one!"

SUMMARY

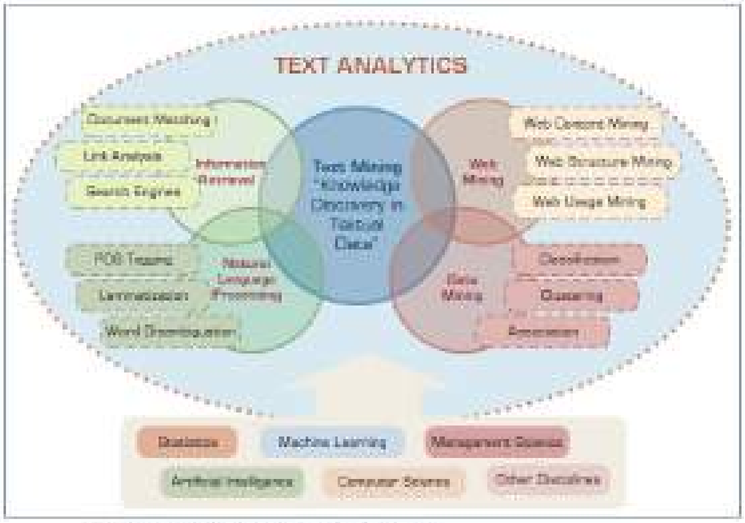
- The *Jeopardy!* challenge helped IBM define the requirements for creating Watson and the DeepQA architecture.
- A team of about 20 researchers worked for 3 years to develop Watson, making it capable of performing at expert human levels in terms of accuracy, confidence, and speed.
- IBM created various computational and language-processing algorithms to solve different challenges in question answering.
- Although the specific details of these algorithms are not publicly known, they heavily relied on text analytics and text mining techniques.
- IBM is now working on adapting Watson to address important challenges in healthcare and medicine.

2) Text Analytics and Text Mining Concepts and Definitions

Index

Introduction

- **Data Growth**: The information age has led to rapid data growth, with 85% of corporate data being unstructured (mostly text). This unstructured data is doubling every 18 months.
- **Importance of Text Data**: Businesses that effectively analyze their unstructured text data will gain valuable knowledge, allowing them to make better decisions and achieve a competitive edge.
- **Text Analytics and Text Mining**:
  - Both aim to turn unstructured text data into actionable insights using natural language processing (NLP) and analytics.
  - **Text Analytics**: A broader concept that includes information retrieval, extraction, data mining, and web mining.
  - **Text Mining**: Focuses on discovering new, useful knowledge from text data.
- **Relationship**:
  - **Text Analytics** = Information Retrieval + Information Extraction + Data Mining + Web Mining.
  - Or, simplified: **Text Analytics** = Information Retrieval + Text Mining.
- In essence, text analytics encompasses multiple processes, while text mining specifically focuses on knowledge discovery from text sources



Text Analytics, Related Application Areas, and Enabling Disciplines.

Terminology:

- **Text Analytics**: A newer term, often used in business contexts, focusing on analyzing text data to gain insights.
- **Text Mining**: An older term, commonly used in academic research, involving the extraction of patterns and knowledge from text.
- Purpose:
- Both terms aim to convert unstructured text data into actionable insights using methods like natural language processing (NLP) and analytics.

Text Mining Definition:

- **Text Mining (also known as text data mining or knowledge discovery in textual databases):** A semi-automated process to extract useful patterns and knowledge from large amounts of unstructured data.
- **ComparisonText Mining with Data Mining**:
- **Data Mining**: Identifies patterns in structured data (e.g., databases with records and variables).
- **Text Mining**: Uses similar techniques but focuses on unstructured data (e.g., Word documents, PDFs, text excerpts).
- **Process of Text Mining**:
- **Step 1**: Impose structure on unstructured text data.
- **Step 2**: Use data mining techniques to extract and analyze information from this structured data.

Most popular application areas of text mining:

1. Information extraction.

2. Topic tracking

3. Summarization.

4. Categorization.

5. Clustering.

6. Concept linking.

7. Question answering.

1. **Information extraction.**

Identification of key phrases and relationships within text by looking for predefined objects and sequences in text by way of pattern matching. Perhaps the most commonly used form of information extraction is named entity extraction.

Named entity extraction includes named entity recognition (recognition of known entity names-for people and organizations, place names, temporal expressions, and certain types of numerical expressions, using existing knowledge of the domain), co-reference resolution (detection of co-reference and anaphoric links between text entities), and relationship extraction (identification of relations between entities).

- "**Co-reference Resolution:**It identifies when different words refer to the **same entity**.
- **Example**: In "John went to the store. He bought milk," "John" and "He" refer to the same person.

2. **Topic tracking.**

- Based on a user profile and documents that a user views, text mining can predict other documents of interest to the user.
- It is used in search engines to recommend content based on user searches.
- For example, if a user frequently searches for deep learning and AI, the engine will provide more related suggestions to enhance their experience.

3. **Summarization.** Summarizing a document to save time on the part of the reader. For example, you can summarize a one-page paragraph into a two-line summary.

4.Categorization means Identifying the main themes of a document and then placing the document into a predefined set of categories based on those themes.

- Example: A blog post discussing healthy eating habits is categorized as "Health" and "Nutrition" based on its main themes

5.Clustering means Grouping similar documents

- Example: A digital collection of books contains various books, and through clustering, novels are grouped together, while non-fiction and reference materials are placed in separate clusters based on their content similarities.

6. Concept linking connectsrelated documents by identifying their shared concepts and, by doing so, helps users find information that they would not have found using traditional search methods.

- **Example**: If a user reads an article about climate change, the system shows a link to another article about renewable energy. This helps the user find more related information they might not have seen before.

7. Question answering :It Finds the best answer to a given question through knowledge-driven pattern matching.

Example: A user asks, "What causes rain?" The system answers, "Rain is caused by condensation of water vapor in the atmosphere.

Text Mining Terminology(Text Mining Lingo)

Text mining lingo refers to the specialized vocabulary and terminology used in the field of text mining

1.Unstructured data (versus structured data).

2.Corpus

3.Terms.

4.Concepts.

5.Stemming.

6.Stop words.

7.Synonyms and polysemes.

8.Tokenizing

9.Term dictionary

10. Word frequency

11. Part-of-speech tagging.

12.Term-by-document matrix (occurrence matrix).

13.Singular-value decomposition

The following list describes some commonly used text mining terms:

1.**Unstructured data (versus structured data).**

- **Structured data** has **a predetermined** format. It is usually organized into records with simple data values (categorical, ordinal, and continuous variables) and stored in databases. In contrast, **unstructured data does** not have a predetermined format and is stored in the form of textual documents.
- **Examples of Structured Data:** Database Records (data in rows and columns), Spreadsheets.
- **Examples of Unstructured data**: Emails, Social Media Posts, Tweets, Facebook posts, Documents: Word files, PDFs.

2. Corpus.

- In linguistics, a corpus is a large and structured set of texts (now usually stored and processed electronically) prepared for the purpose of conducting knowledge discovery. **Example:**
- If you collect all of the news articles from a specific news website for one year, that **collection of articles** would be a corpus.
- A corpus could be a **collection of all the research** papers published in a specific academic journal over the last five years.
- A corpus might **consist of all the tweets** related to a particular event, such as a sports championship, collected over the event's duration.

3.Terms. A term is a single word or multiword phrase extracted directly from the corpus of a specific domain by means of natural language processing (NLP) methods.

- **Example**: In the sentence "The cat sat on the mat," the words "cat," "sat," and "mat" are considered **terms**.

4.Concepts. Concepts are features generated from a collection of documents by means of manual, statistical, rule-based, or hybrid categorization methodology. Compared to terms, concepts are the result of higher level abstraction.

- **Example**: In a collection of health-related documents, the terms "fever," "cough," and "fatigue" might point to the **concept** of "illness.“

3.Terms. A term is a single word or multiword phrase extracted directly from the corpus of a specific domain by means of natural language processing (NLP) methods.

- **Example**: In the sentence "The cat sat on the mat," the words "cat," "sat," and "mat" are considered **terms**.

4.Concepts. Concepts are features generated from a collection of documents by means of manual, statistical, rule-based, or hybrid categorization methodology. Compared to terms, concepts are the result of higher level abstraction.

- **Example**: In a collection of health-related documents, the terms "fever," "cough," and "fatigue" might point to the **concept** of "illness.“

5.Stemming. Stemming is the process of reducing words to their stem (or base or root) form. For instance, stemmer, stemming, and stemmed are all based on the root stem. **Example of Stemming:**

- **Words**: "running," "runner," "ran"
- **Stemmed**: "run"

6.Stop words.

- Stop words (or noise words) are words that are filtered out prior to or after processing of natural language data (i.e., text).
- Even though there is no universally accepted list of stop words, most natural language processing tools use a list that includes articles (a, am, the, of, etc.), auxiliary verbs (is, are, was, were, etc.)..

7. Synonyms and polysemes. Synonyms are different words that have similar meanings, such as "movie," "film," and "motion picture."(e.g., movie, film, and motion picture).(These are different words that have similar meanings.

- **Example**: "happy" and "joyful" are synonyms, as they both convey a sense of happiness.)
- **In contrast, polysemes,** which are also called homonyms, are identical words (i.e., spelled exactly the same) with different meanings.
- **Example**: The word "bank" can refer to:
- A financial institution (e.g., "I deposited money in the bank.") The side of a river (e.g., "We sat by the river bank."))

8. Tokenizing.

- **Tokenizing** is the process of dividing text into categorized blocks called tokens, each assigned a specific meaning based on its function within the sentence. Tokens can be words, phrases, or symbols that provide useful structure to the text.
- **Word Tokenization Example**
- Text: "ChatGPT is very helpful for answering questions."
- Tokens: ["ChatGPT", "is", "very", "helpful", "for", "answering", "questions“])

9.Term dictionary.

- A **term dictionary** in NLP refers to a list or set of predefined words and their associated meanings or categories, which can help with text processing tasks like classification, information retrieval, or sentiment analysis.
- **Example:**
- A company builds a **term dictionary** to classify product reviews as positive or negative.
  - Positive words in the dictionary: "excellent," "amazing," "great," "love."
  - Negative words in the dictionary: "bad," "terrible," "disappointing," "poor."

10. Word frequency. The number of times a word is found in a specific document.

11.Part-of-speech tagging.

- The process of marking up the words in a text as corresponding to a particular part of speech (such as nouns, verbs, adjectives, adverbs, etc.) based on a word's definition and the context in which it is used.

12.Term-by-document matrix (occurrence matrix).

- It shows the relationship between terms(words) and documents in a table.
- Terms are in rows, documents in columns, and the cell values indicate how often each term appears in each document as integers.
- **Documents:**
- **Document 1**: "Apple and banana are fruits. Apple is sweet."
- **Document 2**: "I like to eat a banana and an apple."
- **Document 3**: "An apple a day keeps the doctor away. The apple is healthy."
- **Terms are apple,banana,fruits,eat,doctor,sweet,healthy**

Term	Document 1	Document 2	Document 3
apple	2	1	2
banana	1	1	0
fruits	1	0	0
eat	0	1	0
doctor	0	0	1
sweet	1	0	0
healthy	0	0	1

14.Singular-value decomposition

- SVD is a technique for reducing the size of a term-by-document matrix, making it easier to analyze.

It captures the essential relationships between terms and documents while simplifying the data

3. NATURAL LANGUAGE PROCESSING (NLP):

- **Definition**: A branch of artificial intelligence and computational linguistics that focuses on the interaction between computers and human language.
- **Goal**: Convert human language into structured data that computers can process, aiming for a deeper understanding.

NLP Objectives:

- Moves past simple word counting to include understanding of grammar, semantics, and context.
- True understanding of natural language is difficult because it requires extensive knowledge that goes beyond the text itself.
- NLP has evolved from basic text processing methods to more advanced techniques that attempt to understand language more deeply.

The following are some of the challenges commonly associated with the implementation of NLP:

*1.*Part-of-speech tagging.

2.Text segmentation.

3.Word sense disambiguation.

4. Syntactic ambiguity.

5. Imperfect or irregular input.

6.Speech acts.

1. Part-of-speech tagging.

It is difficult to mark up terms in a text to a particular part of speech (such as nouns, verbs, adjectives, adverbs, etc.) because the part of speech depends not only on the definition of the term but also on the context within which it is used.

- **As a verb**:
- "Please **light** the candle."
- In this case, "light" is an action.
- **As an adjective**:
- "She carried a **light** backpack."
- Here, "light" describes the weight of the backpack

2. Text segmentation.

- *Some written languages, such as Chinese, Japanese, and* Thai, do not have **single-word boundaries**. (**Single-word boundaries** refer to clear separations between individual words in a sentence)
- **Text:** "我喜欢吃苹果" (I like to eat apples in Chinese)
- In these instances, the text-parsing task requires the identification of word boundaries, which is often a difficult task.
- Similar challenges **in speech segmentation** emerge when analyzing spoken language, because sounds representing successive letters and words blend into each other.
- Example: When someone says, "I scream," it can sound like "ice cream" when spoken quickly. The blending of sounds makes it challenging to tell where one word ends and the other begins, causing confusion in understanding.

3. Word sense disambiguation.

- *Many words have more than one meaning.*
- Selecting the meaning that makes the most sense can only be accomplished by taking into account the context within which the word is used.
- Financial Institution: "I need to deposit money at the bank.
- "Riverbank: "We had a picnic by the river bank."

4. Syntactic ambiguity. *The grammar for natural languages is ambiguous; that is,* multiple possible sentence structures often need to be considered. Choosing the most appropriate structure usually requires a fusion of semantic (Semantic focuses on understanding what the words convey, rather than how they are arranged) and contextual information .

- Sentence: "I looked at the man using the telescope."
- Possible Structures:
- "I looked at [the man using the telescope]" – Meaning: I observed a man who was using the telescope.
- "I looked at [the man] [using the telescope]" – Meaning: I observed the man through the telescope.

5. ***Imperfect or irregular input. Foreign or regional accents and vocal impediments*** in speech and typographical or grammatical errors in texts make the processing of the language an even more difficult task.

- **Speech with Accents:** A sentence like "I need to book a flight" might be pronounced differently by someone with a strong regional accent, making it harder for speech recognition systems to accurately transcribe it.
- **Typographical Errors:** In a text message, a typo like "I need a reserch paper" instead of "I need a research paper" can confuse text processing systems and lead to incorrect or incomplete interpretations.

6. Speech acts.

- **Speech act** is something expressed by an individual that not only presents information but performs an **action** as well.
- *The* sentence structure alone may not contain enough information to define this action.
- **Speech acts** might be requests, warnings, promises, apologies, physical actions, greetings.

For example, "Can you pass the class?" requests a simple yes/no answer, whereas "Can you pass the salt?" is a request for a physical action to be performed

Speech acts are tough for computers because understanding language involves more than just recognizing words—it's about grasping deeper meaning, whether it's a yes/no answer or a request for action, which remains challenging for technology.

Applications of NLP Across Various Domains(NLP has successfully been applied to a variety of domains for a variety of tasks)

1. Question answering

2.Automatic summarization

3.Natural language generation.

4. Natural language understanding.

5. Machine translation.

6. Foreign language reading.

7. Foreign language writing

8. Speech recognition.

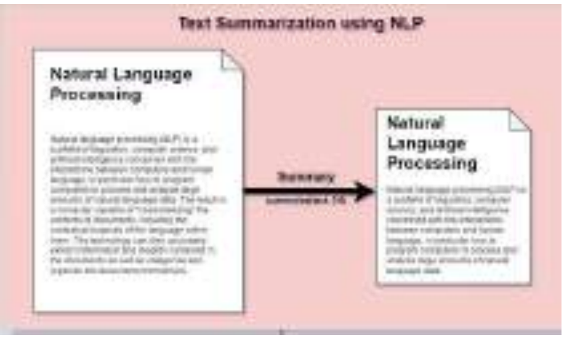
9.Text-to-speech.

10. Text proofing.

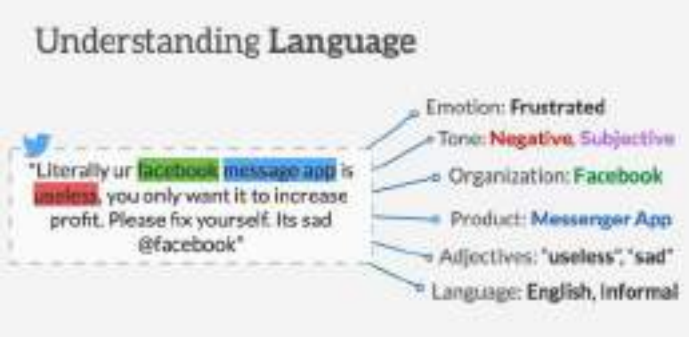
11. Optical character recognition

1. Question answering.

- *It is task of automatically answering a question posed* in natural language; that is, **producing a human-language** answer **when given a human-language question**.
- To find the answer to a question, the computer program may use either a prestructured database or a collection of natural language documents .
- ***2.Automatic summarization.*** *The creation of a shortened version of a textual* document by a computer program that contains the most important points of the original document.


Text Summarization using NLP
Natural Language Processing
Summary
Natural Language Processing

3. Natural language understanding*. Systems convert samples of human language* into more formal representations that are easier for computer programs to manipulate.


Understanding Language
"Literally ur facebook message app is useless, you only want it to increase profit. Please fix yourself. Its sad @facebook"
Emotion: Frustrated
Tone: Negative, Subjective
Organization: Facebook
Product: Messenger App
Adjectives: "useless", "sad"
Language: English, Informal

4.Natural language generation

- NLG, another subset of NLP, involves **the generation of human-like text by computers.** It converts structured data into natural language.
- NLG is widely used in applications such as chatbots for generating responses, automated report writing, and creating written content from data sets.

5. ***Machine translation.*** *The automatic translation of one human language to* another.

6.Foreign language reading. *A computer program that assists a nonnative language* speaker to read a foreign language with correct pronunciation and accents on different parts of the words.

7.Foreign language writing. *A computer program that assists a nonnative language* user in writing in a foreign language.

8. ***Speech recognition.*** *Converts spoken words to machine-readable input. Given a* sound clip of a person speaking, the system produces a text dictation.

9.Text-to-speech. *Also called speech synthesis, a computer program automatically* converts normal language text into human speech .

10 ***Text proofing.***

- This refers to a computer program that reviews a proof copy of text to identify and fix errors.
- It checks for spelling, grammar, punctuation, and formatting mistakes, ensuring the text is polished and error-free before final publication.
- 11. ***Optical character recognition.*** *The automatic translation of images of handwritten,* typewritten , or printed text (usually captured by a scanner) into machine editable textual documents.

4. TEXT MINING APPLICATIONS

A. Marketing Applications
B. Security Applications
C. Biomedical Applications

D. Academic Applications

A. Marketing Applications

1. Cross-Selling and Up-Selling:

- Text mining examines unstructured data from call center notes and voice transcripts to understand customer perceptions.
- This analysis helps identify opportunities for cross-selling (selling additional products) and up-selling (selling higher-end products).

2. Customer Sentiment Analysis:

- Text mining processes blogs, product reviews, and discussion board posts to capture customer sentiments.
- Understanding this rich data helps enhance customer satisfaction and increases their overall lifetime value with the company.

3. Customer Relationship Management (CRM):

- Companies combine unstructured text data with structured data from their databases to gain insights into customer behavior.
- Text mining improves the ability to predict customer churn (attrition), enabling companies to identify at-risk customers for targeted retention efforts.

4. Product Attribute Analysis:

- Text mining systems can identify both explicit and implicit product attributes, allowing retailers to analyze product databases more effectively.
- Explicit attributes are clearly defined features, such as color, size, or brand, while implicit attributes are inferred from data, like customer sentiment or usage context.
- Treating products as sets of attribute-value pairs enhances effectiveness in demand forecasting, product recommendations, and supplier selection.

Product: Smartphone

Explicit Attributes:

- These are clearly defined features that can be directly observed or stated about the product.
- **Brand**: Samsung
- **Color**: Black
- **Storage**: 128GB
- **Price**: $699
- **Implicit Attributes**:
- They require interpretation, often using methods like text mining, sentiment analysis,
- **Customer sentiment**: Positive reviews about the smartphone's battery life or design (e.g., inferred from customer feedback).
- **Usage context**: The smartphone is often used for gaming or photography (inferred from reviews or user discussions).

B. Security Applications

1. ECHELON Surveillance System:

- **E**CHELON is a major classified text mining application in the security field
- It is a classified system rumored to analyze content from phone calls, faxes, emails, and other communications.
- It can intercept information via various channels, including satellites and public networks.

2. EUROPOL's OASIS System:

- It was developed in 2007 to track transnational organized crime by analyzing large volumes of structured and unstructured data.

- **It** Integrates advanced data and text mining technologies, enhancing law enforcement efforts internationally.

3. FBI and CIA Supercomputer System:

- The FBI and CIA are working together to create a comprehensive data warehouse for law enforcement.

4. Supercomputer System Goal is to improve knowledge discovery by connecting previously separate databases, enhancing data accessibility for federal, state, and local agencies.

5. Deception Detection:

- Text mining is used to analyze statements from persons of interest in criminal investigations.
- It Developed models that achieved 70% accuracy in distinguishing between deceptive and truthful statements, relying solely on textual cues.
- This method can be applied to both text and transcriptions of voice recordings, offering a alternative to traditional techniques like polygraphs.

C .Biomedical Applications

- **Growing Medical Literature**:

The field of biomedical literature is expanding rapidly, particularly with the rise of open-source journals.Medical literature is well-organized and standardized, making it easier to mine for valuable insights.

- **Data from Experimental Techniques**:
  - Techniques like DNA microarray analysis, SAGE, and mass spectrometry produce vast amounts of data.
  - Text mining tools are essential for analyzing and interpreting this data by cross-referencing it with existing literature.
- Text mining techniques are enhancing understanding of biological processes by predicting protein locations, extracting disease-gene relationships, and discovering gene-protein interactions.
- **Protein Location Prediction**:
  - Knowing a protein’s location in a cell is crucial for understanding its biological role and potential as a drug target.
  - Shatkay et al. (2007) developed a system that integrates sequence-based and text-based features to predict protein locations. Their system outperformed previous models.

- **Disease-Gene Relationship Extraction**:
  - Chun et al. (2006) created a system to extract relationships between diseases and genes from biomedical literature using a dictionary for disease/gene names.
  - To reduce false positives, they used machine learning-based Named Entity Recognition (NER) filtering, which improved precision by 26.7%.
- **Gene-Protein Relationship Discovery**:
  - Nakov et al. (2005) demonstrated a multilevel text analysis process to discover gene-protein and protein-protein relationships from biomedical texts.
  - This process involves tokenizing text, part-of-speech tagging, and shallow parsing, then matching terms against domain ontologies to interpret relationships.

D.Academic Applications

- **Publishers:** Large databases need indexing for better information retrieval, especially in scientific fields. Initiatives like Nature's Open Text Mining Interface (OTMI) and NIH's Journal Publishing DTD aim to enable machines to answer queries without removing publisher barriers.
- Academic institutions have also launched text mining initiatives. For example, the National Centre for Text Mining, a collaborative effort between the Universities of Manchester and Liverpool, provides customized tools, research facilities, and advice on text mining to the academic community. With an initial focus on text mining in the biological and biomedical sciences, research has since expanded into the social sciences.
- In the United States, the School of Information at the University of California, Berkeley, is developing a program called BioText to assist bioscience researchers in text mining and analysis.

6. TEXT MINING TOOLS

A. Commercial Software Tools

The following are some of the most popular software tools used for text mining. Note that many companies offer demonstration versions of their products on their Web sites.

1. **ClearForest** offers text analysis and visualization tools.

2. IBM offers **SPSS Modeler and data and text analytics toolkits**.

3. **Megaputer Text Analyst** offers semantic analysis of free-form text, summarization, clustering, navigation, and natural language retrieval with search dynamic refocusing.

4. **SAS Text Miner** provides a rich suite of text processing and analysis tools.

5. **KXEN Text Coder (KTC)** offers a text analytics solution for automatically preparing and transforming unstructured text attributes into a structured representation for use in KXEN Analytic Framework.

6. The **Statistica Text Mining** engine provides easy-to-use text mining functionality with exceptional visualization capabilities.

7. **VantagePoint** provides a variety of interactive graphical views and analysis tools with powerful capabilities to discover knowledge from text databases.

8. The **WordStat analysis** module from Provalis Research analyzes textual information such as responses to open-ended questions, interviews, etc.

B. Free Software Tools

Free software tools, some of which are open source, are available from a number of nonprofit organizations:

1. **RapidMiner,** one of the most popular free, open source software tools for data mining and text mining, is tailored with a graphically appealing, drag-and-drop user interface.

2. **Open Calais** is an open source toolkit for including semantic functionality within your blog, content management system, Web site, or application.

3. **GATE** is a leading open source toolkit for text mining. It has a free open source framework (or SDK) and graphical development environment.

4. **LingPipe** is a suite of Java libraries for the linguistic analysis of human language.

5. **S-EM (Spy-EM)** is a text classification system that learns from positive and unlabeled examples.

6. **Vivisimo/ Clusty** is a Web search and text-clustering engine .

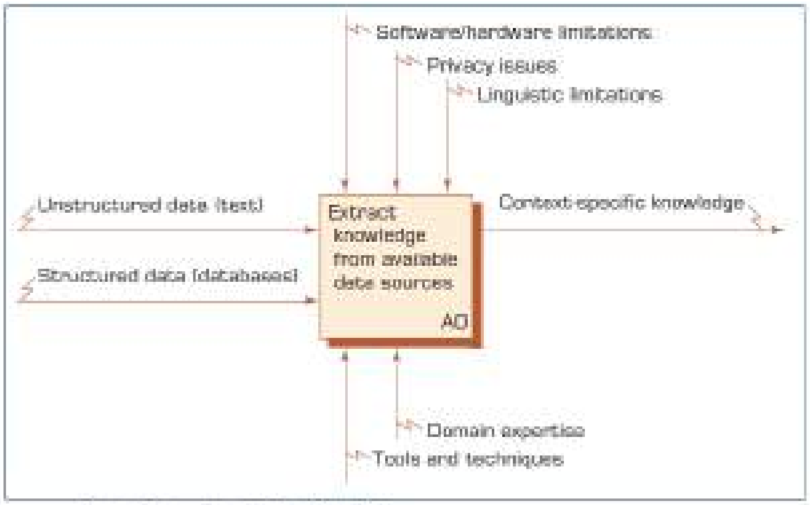
7.Text Mining Process

- **Context Diagram for the Text Mining Process**
- **The Three-Step Text Mining Process**

Context Diagram for the Text Mining Process

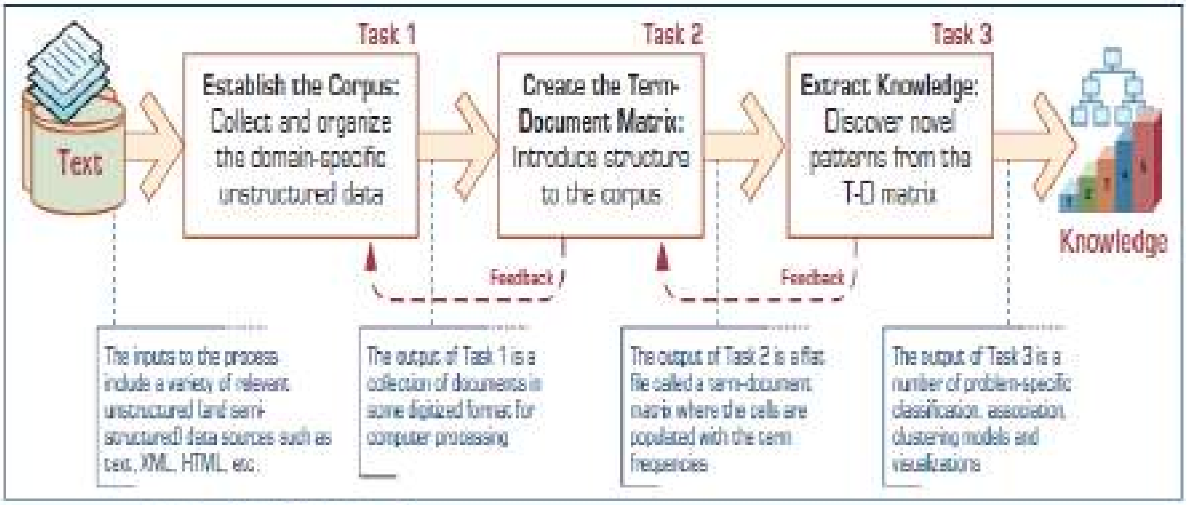
- As the context diagram indicates, the input (inward connection to the left edge of the box) into the text-based knowledge-discovery process is the unstructured as well as structured data collected , stored, and made available to the process.
- The output (outward extension from the right edge of the box) of the process is the context-specific knowledge that can be used for decision making.
- The controls, also called the *constraints (inward connection* to the top edge of the box), of the process include software and hardware limitations, privacy issues, and the difficulties related to processing the text that is presented in the form of natural language.
- The mechanisms (inward connection to the bottom edge of the box) of the process include proper techniques, software tools, and domain expertise.

- The text-based knowledge discovery process involves analyzing both unstructured and structured data to generate context-specific knowledge for decision-making.
- Inputs include collected data, while outputs are actionable insights.
- The process is governed by constraints like software limitations and privacy issues, and relies on techniques and expertise.
- It consists of three tasks, with feedback loops for adjusting outputs if necessary.(Three-Step Text Mining Process)


Software/hardware limitations
Privacy issues
Linguistic limitations
Unstructured data (text)
Structured data (databases)
Extract knowledge from available data sources
A0
Context-specific knowledge
Domain expertise
Tools and techniques

Context Diagram for the Text Mining Process.

The Three-Step Text Mining Process.


Text
Task 1
Establish the Corpus: Collect and organize the domain-specific unstructured data
Task 2
Create the Term-Document Matrix: Introduce structure to the corpus
Task 3
Extract Knowledge: Discover novel patterns from the T-D matrix
Knowledge
Feedback
Feedback
The inputs to the process include a variety of relevant unstructured (and semi-structured) data sources such as text, XML, HTML, etc.
The output of Task 1 is a collection of documents in some digitized format for computer processing
The output of Task 2 is a flat file called a term-document matrix where the cells are populated with the term frequencies
The output of Task 3 is a number of problem-specific classification, association, clustering models and visualizations

The Three-Step/Task Text Mining Process.

- **The primary purpose of text mining** (within the context of knowledge discovery) is to process unstructured (textual) data (along with structured data, if relevant to the problem being addressed and available) to extract meaningful and actionable patterns for better decision making.
- At a very high level, **the text mining process can be broken down into three consecutive tasks**, each of which has specific inputs to generate certain outputs (see Figure 7.6).
- If, for some reason ,**the output of a task is not what is expected, a backward redirection** to the previous task execution is necessary.

- The diagram shows the **three-step text mining process**:

1. Establish the Corpus: Collect and organize domain-specific unstructured data (like text, XML, HTML).

2. Create the Term-Document Matrix: Structure the data into a matrix where each cell represents term frequency in documents.

3. Extract Knowledge: Use the matrix to find patterns using classification, clustering, or association techniques

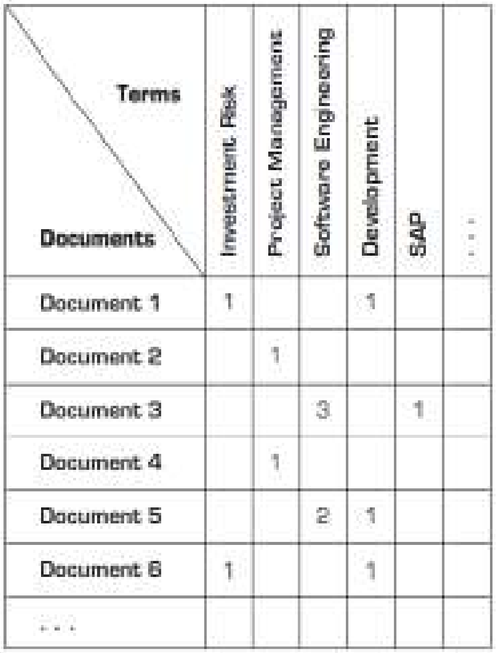
Explanation in detail

1. Establish the Corpus

- The first task activity aims to collect all documents related to the context (domain of interest), including textual documents, XML files, e-mails, web pages, short notes, and transcribed voice recordings using speech-recognition algorithms.
- Once collected, the text documents are transformed and organized into a uniform format (e.g., ASCII text files) for processing.
- This can be a collection of digitized text stored in a folder or links to specific web pages. Many text mining software tools can convert these inputs into a flat file.
- **Examples of Flat Files:**
- **CSV Files** (.csv) — a structured flat file where columns may represent different features, such as "Document ID," "Text," "Date," etc.

2. Create the Term-Document Matrix

- In this task, the digitized and organized documents (the corpus) are used to create the **term-document matrix (TDM).**
- Build a "term-document matrix," where:
  - Rows represent documents.
  - Columns represent unique terms (words).
  - Cells contain the frequency of each term in each document.
- The goal is to convert the list of organized documents (the corpus) into a TDM where the cells are filled with the most appropriate indices



Terms Documents	Investment Risk	Project Management	Software Engineering	Development	SAP	. . .
Document 1	1			1		
Document 2		1				
Document 3			3		1	
Document 4		1				
Document 5			2	1		
Document 6	1			1		
. . .						

Figure: A Simple Term–Document Matrix.

- The assumption is that the essence of a document can be represented with a list and frequency of the terms used in that document.

Fundamental Challenges In Document Term Selection And Dimensionality Reduction Are

- **1.Methods For Identifying And Excluding Less Important Terms In Document Characterization**
- **2.Representing The Indices**
- **3.Reducing The Dimensionality Of The Matrix**

1.Methods for Identifying and Excluding Less Important Terms in Document Characterization

we can use the following methods and remove less important terms

- Exclude stop words
- Domain-specific stop words
- Include terms/dictionary
- Synonyms and phrases treated as single units
- Stemming
- Not all terms are important when characterizing documents. Here’s a summary of what should be done to refine the term-document matrix (TDM):

- **Exclude stop words**: Common terms like articles, auxiliary verbs, and frequently used words that offer no differentiation should be removed.
- **Domain-specific stop words**: Terms that hold little relevance for the analysis should be identified by domain experts and excluded.
- **Include terms/dictionary**: Preselected important terms can be used for indexing, ensuring relevance to the domain.
- **Synonyms and phrases**: Terms with similar meanings (synonyms) and specific multi-word phrases (like "Eiffel Tower") should be treated as single units to improve accuracy.
- **Stemming**: Apply stemming to reduce words to their root forms, ensuring variations like "modeled" and "modeling" are recognized as "model."
- **TDM generation**: The TDM is created with unique terms as columns, documents as rows, and term occurrence counts in the cells. However, for large corpora, this can result in a large matrix, which might complicate pattern extraction and lead to inaccuracies.
- **Documents:**
- **Document 1**: "Apple and banana are fruits. Apple is sweet."
- **Document 2**: "I like to eat a banana and an apple."
- **Document 3**: "An apple a day keeps the doctor away. The apple is healthy."

Term	Document 1	Document 2	Document 3
apple	2	1	2
banana	1	1	0
fruits	1	0	0
eat	0	1	0
doctor	0	0	1
sweet	1	0	0
healthy	0	0	1

2.REPRESENTING THE INDICES

- Once the input documents are indexed and the initial word frequencies (by document) computed, a number of additional transformations can be performed to summarize and aggregate the extracted information.
  - Raw term frequencies indicate how often a word appears in a document, suggesting its relevance.
  - Higher frequency often implies a stronger descriptor of the document's content.
- **Limitations of Raw Counts**:
  - Raw counts may not accurately reflect a term's importance. For instance:
    - If a word appears once in Document A and three times in Document B, it does not mean it's three times more important in Document B.
- In order to have a more consistent TDM for further analysis, these raw indices need to be normalized (Indices need to be normalized)
- Normalization methods
  - ***Log frequencies.***
  - ***Binary frequencies.***
  - ***Inverse document frequencies.***
- **Normalization Methods**
- To ensure a consistent and meaningful TDM, several normalization methods can be employed:
- **Log Transformation**:
  - Applies a logarithmic function to dampen the effect of raw frequencies.

$$f(wf) = 1 + \log(wf) \quad \text{for} \quad wf > 0$$

- **Binary Frequencies**:
  - Represents terms as present (1) or absent (0) in documents, disregarding frequency.

$$f(wf) = 1 \quad \text{for} \quad wf > 0$$

TDM

Term	Document 1	Document 2	Document 3
apple	2	1	2
banana	1	1	0
fruits	1	0	0
eat	0	1	0
doctor	0	0	1
sweet	1	0	0
healthy	0	0	1

Log frequencies.

Term	Document 1	Document 2	Document 3
apple	1.301	1	1.301
banana	1	1	0
fruits	1	0	0
eat	0	1	0
doctor	0	0	1
sweet	1	0	0
healthy	0	0	1

Binary frequencies.

Term	Document 1	Document 2	Document 3
apple	1	1	1
banana	1	1	0
fruits	1	0	0
eat	0	1	0
doctor	0	0	1
sweet	1	0	0
healthy	0	0	1

- ***Inverse document frequencies***

Term	Document Frequency (df)	IDF (log10(N/df))	IDF
apple	3	log10(3/3) = log10(1) = 0	0
banana	2	log10(3/2) = 0.176	0.176
fruits	1	log10(3/1) = 0.477	0.477
eat	1	log10(3/1) = 0.477	0.477
doctor	1	log10(3/1) = 0.477	0.477
sweet	1	log10(3/1) = 0.477	0.477
healthy	1	log10(3/1) = 0.477	0.477

3. REDUCING THE DIMENSIONALITY OF THE MATRIX

- **Because the TDM is often very large** and rather sparse (most of the cells filled with zeros), another important question is "**How do we reduce the dimensionality of this matrix to a manageable size?**“
- Several options are available for managing the matrix size :

--- A domain expert goes through the list of terms and eliminates those that do not make much sense for the context of the study (this is a manual, labor-intensive process).

  – Eliminate terms with very few occurrences in very few documents.

  – Transform the matrix using singular value decomposition.

– Singular Value Decomposition (SVD) is a powerful technique used to reduce the dimensionality of data, particularly useful in text mining and natural language processing.

– Specifically, assume that matrix ***A represents an m X n term*** *occurrence* matrix where *m is the number of input documents and n is the number of terms selected* for analysis.

– The SVD computes the *m X r orthogonal matrix U, n X r orthogonal matrix V,*and*r X r matrix D,*

– *so that A = UDV' and r is the number of eigen values of A'A.*

3. Extract Knowledge

- **Goal**: Discover meaningful patterns or insights from the term-document matrix.
- **Action**: Use text mining algorithms (e.g., classification, clustering, association) to uncover patterns, such as key topics or word relationships.
- **Output**: Extracted knowledge or insights, such as clusters of similar documents or terms, which can be used to solve specific problems.
- Apply text mining techniques to analyze the structured term-document matrix. Common algorithms used include:
- **Text classification** is a common task in analyzing complex data, where the goal is to categorize data instances into predefined classes. In **text mining process** involves assigning documents to relevant topics or subjects using models trained on labeled data.
- **Applications** of automated text classification include:
  - Spam filtering, Web page categorization
- **Approaches**:
- **Knowledge Engineering**: Experts manually encode rules for classification.
- **Machine Learning**: Models learn from labeled examples

CLUSTERING

- **Clustering is an unsupervised process whereby objects are classified into** "natural" groups called *clusters.*
- In clustering the problem is to group an unlabelled collection of objects (e.g., documents, customer comments, Web pages) into meaningful clusters without any prior knowledge.

1. Scatter/Gather Clustering:

• Dynamically groups documents as you browse, helping user

to explore without searching.

• Example :Browsing an online library that automatically sorts

books into new categories as you explore.

2. Query-Specific Clustering organizes documents based on how relevant they are to your search.

• The most important ones are placed in small, focused groups (smaller clusters), while less important ones are in larger groups (larger clusters).

• Smaller clusters contain documents closely related to your query, and larger clusters contain documents that are still

relevant but not as closely matched.

- **Classification**: Assigning documents to predefined categories. For example, classifying customer reviews into "positive" or "negative" sentiment.
- **Clustering**: Grouping documents with similar content. For example, clustering news articles into different topics like "sports", "politics", "technology", etc.
- **Association**: Finding relationships between terms. For instance, identifying words frequently occurring together in medical literature to highlight potential disease-symptom associations. Text mining with association rules was used to analyze published literature (news and academic articles posted on the Web) to chart the outbreak and progress of bird flu
- **Feedback Loop:**
- The process includes feedback between steps, where outputs from later steps can inform earlier steps. For example, if the term-document matrix isn’t effective, adjustments may be made to the corpus or the matrix structure.
- **Overall Flow:**

- The diagram represents how raw, unstructured text data is gradually transformed into structured insights. Starting from raw data collection, to organizing terms into a matrix, and finally extracting meaningful patterns for decision-making.