Challenges in BA

* Data quality & integration issues
* Lack of skilled personnel & analytical expertise
* Privacy & security concerns
* Resistance to adopting data driven decision making

Future Trends in BA

* AI & ML integration
* Real-Time Analytics
* Cloud-Based Analytics
* Natural language processing (NLP)
* Ethical AI & Analytics

Brief Introduction To Big Data Analytics

* Big data is data that cannot be stored in a single storage unit.
* Big Data refers to data that is arriving in many different forms, be they structured, unstructured or in a stream
* Two aspects to mag managing data on this scale: Storing & processing
* 3V's (volume, velocity, variety)

Hdfs
Hadoop
MapReduce

Unit-II

Machine versus Men on Jeopardy! The Story of Watson

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 & advanced computer science, benefiting science, business & society
* The challenge: build a real-time Jeopardy! Contestant capable of listening, understanding & responding

Competing Against the Best

* In 2011, Watson completed in Jeopardy! 's First human versus-machine match.
* Watson won a two-game match, defeating top players Brad Rutter & Ken Jennings
* Watson excelled in signaling but had trouble with short few word clues

* It processed 200 millions pages of data using four terabytes of storage without Internet access.
* watson employed multiple QA technologies like text mining, NLP & question-answering techniques.
* It computed confidence levels to decide whether to "buzz in" within 1-6 seconds, averaging 3 seconds

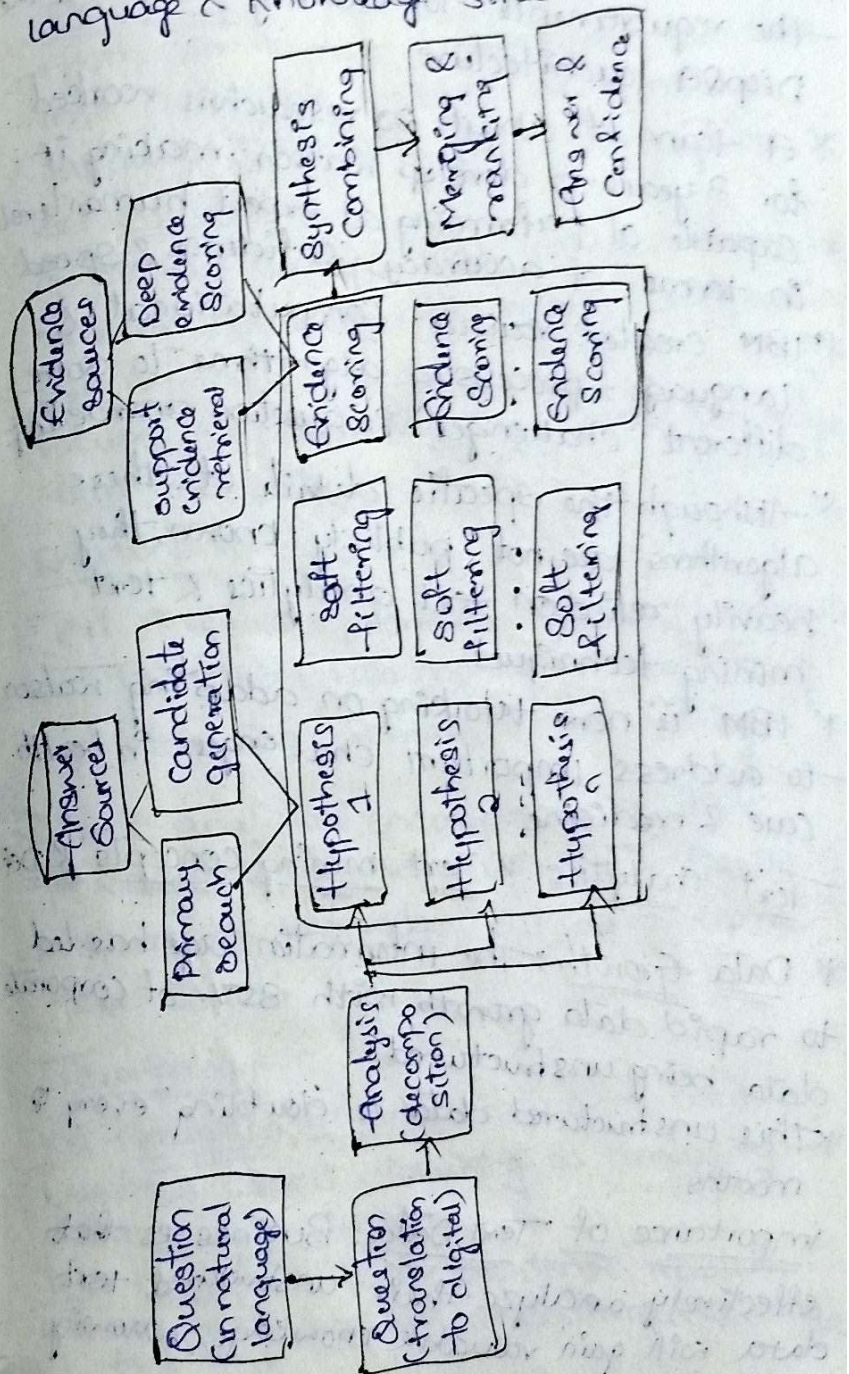
How does watson Do it?

* Watson's system called DeepQA, is a powerful, parallel-processing architecture that uses over 100 techniques to analyze language, generate & evaluate answers & rank them.
* The key to its success lies in how these techniques work together to improve accuracy, speed & confidence

Key principles of DeepQA

* Massive parallelism: It examines multiple interpretations & answers at once.
* Many experts: uses a range of probabilistic methods for analyzing questions & content
* Pervasive confidence estimation: Each part of the system scores potential answers with confidence levels, which are combined to find the best answer

* Shallow & deep knowledge: It balances between detailed & broader understanding of language & knowledge structures.



Conclusion

* The Jeopardy! challenge helped IBM define the requirements for creating Watson & 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 & speed
* IBM created various computational & language-processing algorithms to solve different challenges, ex question answering
* Although the specific details of these algorithms are not publicly known, they heavily relied on text analytics & text mining techniques
* IBM is now working on adapting Watson to address important challenges in health care & medicine

Text Analytics & Text mining concepts & Def

* Data Growth: The information age has led to rapid data growth with 85% of corporate data being unstructured
* This unstructured data is doubling every 6 months

Importance of Text Data: Businesses that effectively analyze their unstructured text data will gain valuable knowledge, allowing them to make better decisions & achieve a competitive edge.

Text Analytics & Text Mining

→ Both aim to turn unstructured text data into actionable insights using natural language processing (NLP) & analytics

Text analytics

A broader concept that includes information retrieval, extraction, data mining & web mining

Text mining

Focuses on discovering new, useful knowledge from text data.

Relationship

Text Analytics = Information Retrieval + Information Extraction + Data mining + Web mining
= Information Retrieval + Text mining

* Text analytics encompasses multiple processes, while text mining specifically focuses on knowledge discovery from text sources

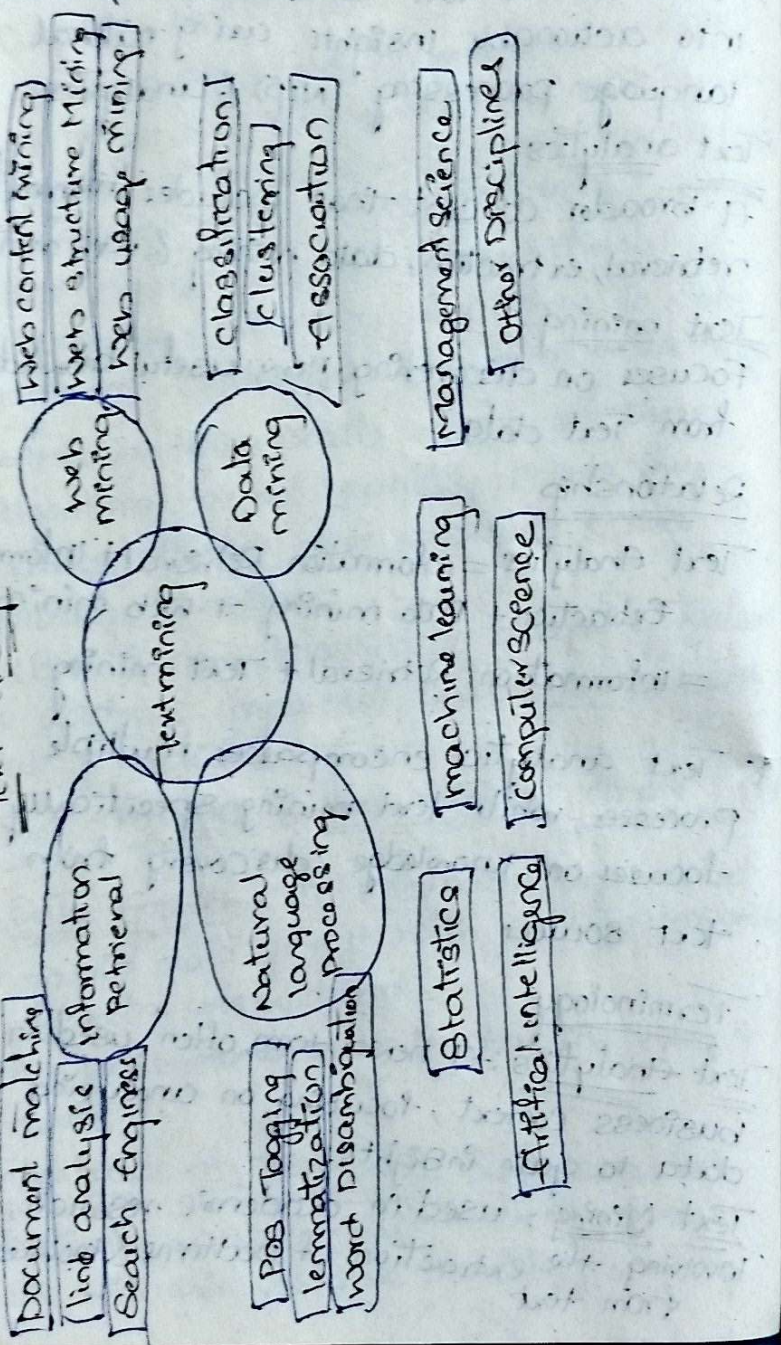
Terminology

Text Analytics: A newer term, often used in business context, focusing on analyzing text data to gain insights

Text Mining: used in academic research, involving the extraction of patterns & knowledge from text

Purpose

Both terms aim to convert unstructured text data into actionable insights using methods like NLP & analytics



Text Mining Definition

→ Semi-automated process to extract useful patterns & knowledge from large amounts of unstructured data

comparision with Data mining

DM: identifies patterns in structured data

TM: uses similar techniques but focuses on unstructured data.

Process of Text mining

1. impose structure on unstructured text data
2. use data mining techniques to extract & 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 thinking
7. Question answering

① Information extraction: It identifies key phrases & relationships in text using pattern matching. the most common method is named entity extraction, which includes:

* Named Entity Recognition: Detects entities like people, organizations, locations, dates & numbers
* Co-reference Resolution: links entities that refer to the same thing
* Relation Extraction: Identifies relationships between entities

Ex: "John works at Google in California," it extracts "John" (person), "Google" (organization) & "California" (place), plus the relationship "works at"

② Topic Tracking

* Based on a user profile & 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

Ex: If a user frequently searches for deep learning & 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.

Ex: paragraph into 2 lines

4. Categorization

Identifying the main themes of a document & then placing the documents into a predefined set of categories based on those themes

Ex: A blog post discussing healthy eating habits is categorized as "Health" & "Nutrition" based on its main themes

5. Clustering: Grouping similar documents

6. Concept linking: connects related documents by identifying their shared concepts & by doing so, helps user find information that they would not have found using traditional search methods.

Ex: If 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

Ex: "what causes rain?"
caused by condensation of water vapour in the atmosphere.

Natural Language Processing (NLP)

Def: A branch of artificial intelligence & computational linguistic that focuses on the interaction between computers & human language.

-Goal: convert human language into structured data that computers can process, aiming for a deeper understanding

Objectives

* Moves past simple word counting to include understanding of grammar, semantics & 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.

Some 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. parts of speech tagging

It is difficult to mark up terms in a text to a particular part of speech because the part of speech depends not only on the def 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."
* "light" describes the weight of the backpack

2. Text Segmentation

* Text-parsing ~~text~~ task requires the identification of word boundaries, which is often a difficult task.
* in speech segmentation emerge when analyzing spoken language, because sounds representing successive letters & words blend into each other.

Ex: "I scream," it can sound like "ice cream" when spoken quickly. The blending of sounds makes it challenging to tell where one word ends & 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 NLP is ambiguous - that is multiple possible sentence structures often need to be considered.

* choosing the most appropriate structure usually requires a fusion of semantic & contextual information
* "I looked at the man using the telescope."
* possible structures:
* "I looked at [the man using the telescope]" observed a man who was using the telescope
* "I looked at [the man] [using the telescope]" observed the man through the telescope

5. imperfect or irregular input

* Foreign or regional accents & vocal impediments in speech & typographical (or) grammatical errors in texts make the processing of the language an even more difficult task.
* Speech with Accents: "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: "I need a reserch paper" instead of "I need a research paper" can confuse text processing systems & lead to incorrect or incomplete interpretations

6. Speech acts

* It 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.
* This might be requests, warnings, promises, apologies, physical actions, greetings.

Ex: "Can you pass the class?"

These are tough for computers because understanding language involves more than just recognizing words.

Applications of NLP

1. Question answering
Producing a human language answer when given a human-language question

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

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

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 & 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 non-native language speaker to read a foreign language with correct pronunciation & accents on different parts of the words

7. Foreign language writing
A computer program that assists a non-native 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 & fix errors
* It checks for spelling, grammar, punctuation & formatting mistakes, ensuring the text is polished & error-free before final publication

11. optical character recognition
the automatic translation of images of handwritten, typewritten or printed text into machine editable textual documents

Text Mining Applications

1. Marketing Applications
2. Security "
3. Biomedical "
4. Academic "

Marketing Applications

1. Cross-Selling & up-selling
* Text mining examines unstructured data from call center notes & voice transcripts to understand customer perceptions
* This analysis helps identify opportunities for cross-seling & upselling

2. Customer Sentiment Analysis
* Text mining processes blogs, product reviews & discussion boards posts to capture customer sentiments
understanding this rich data helps enhance customer satisfaction & increase 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, enabling companies to identify at risk customers for targeted retention efforts.

4. Product Attribute Analysis
* Text mining systems can identify both explict & 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 & supplier selection.

Product: Smartphone

Explict Attribute
These are clearly defined features that can directly observed or stated about the product
Brand: Samsung
color: Black
storage: 128GB
Price: $699.

Implicit Attribute
They require interpretation often using methods like text mining, Sentiment analysis
Customer Sentiment
positive reviews about the smartphone's battery life or design
Usage context
The smartphone is often used for gaming or photography

Security Applications

a) ECHELON Surveillance system
* Echelon is major classified text mining application in the security field
* It is classified system rumored to analyze content from phone calls, faxes, emails & other communications
* It can intercept information via various channels including Satellites & public networks

b) EUROPOL's OASIS system
* It was developed in 2007 to track transnational organized crime by analyzing large volumes of structured & unstructured data
* It integrated advanced data & text mining technologies, enhancing law enforcement efforts internationally.

(c) FBI and CIA Supercomputer System
The FBI & CIA are working together to create a comprehensive data warehouse for law enforcement.

(d) Supercomputer System Goal is to improve knowledge discovery by connecting previously separate databases, enhancing data accessibility for federal, state & local agencies.

(e) Deception Detection
* Text mining is used to analyse statements from persons of interest in crimal investigations
* It developed models that achieved 70% accuracy in distinguishing between deceptive & truthful statements relying solely on textual cues
* This method can be applied to both text and transcriptions of voice recording, offering a alternative to traditional techniques like polygraphs

Biomedical Applications

(1) Growing Medical literature

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

(2) 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 & 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 & discovering geneprotein interactions

(3) Protein location prediction

* knowing a protein's location in a cell is crucial for understanding its biological role & potential as a drug target
* Shatkay (2007) developed a system that integrates sequence-based & text-based features to predict protein-locations. Their system out performanced previous models

(4) Disease-Gene Relationship Extraction

* Chun. et. al (2006). created a system to extract relationships between diseases & genes from biomedical literature using a dictionary for disease/gene names
* To reduce false positives, they used machine learning-based Named Entity Recognition filtering, which improved precision by 26.7%

(5) Gene-protein Relationship Discovery

* Nator (2005) demostrated a multilevel text analysis process to discover gene-protein & protein-protein relationships from biomedical texts
* this process involves tokenizing text, part-of-speech tagging & shallow parsing, then matching terms against domain ontologies to interpet relationships

Academic Applications

To analyse large volume of scholarly literature, identifying research trends, key topics, emerging areas of study, research gaps & patterns within a field, allowing researchers to quickly summarize & extract relevant information from published papers, ultimately accelerating the research process.

Text Mining Tools

1. Commercial software tools

(i) clear Forest: offers text analyses & visualization

(ii) IBM offers spss Mateler & data & text analytics toolkits

(iii) Megaputer Text Analyst offers semantic analysis of free-form text, summarization, clustering, navigation & natural language retrieval with search dynamic refocussing

(iv) SAS Text Miner provides a rich suite of text processing & analysis tools.

(v) KXEN Text Coder (KTC) offers a text analytics solution for automatically preparing & transforming unstructured text attributes into a structured representation for use in KXEN analytic framework.

(vi) Statistica Text mining engine provides easy-to-use text mining functionality with exceptional visualization capabilities.

(vii) Vantage point provides a variety of interactive graphical views & analysis tools with powerful capabilities to discover knowledge from text databases

(viii) Wordstat analysis module from provalis Research analyzes textual information such as responses to open-ended questions, interviews etc

2. Free software Tools

Some of which are open source, are available from a no. of nonprofit organizations.

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

2. open calais is an open source toolkit for including semantic functionality within your blog content management system, website (or) application

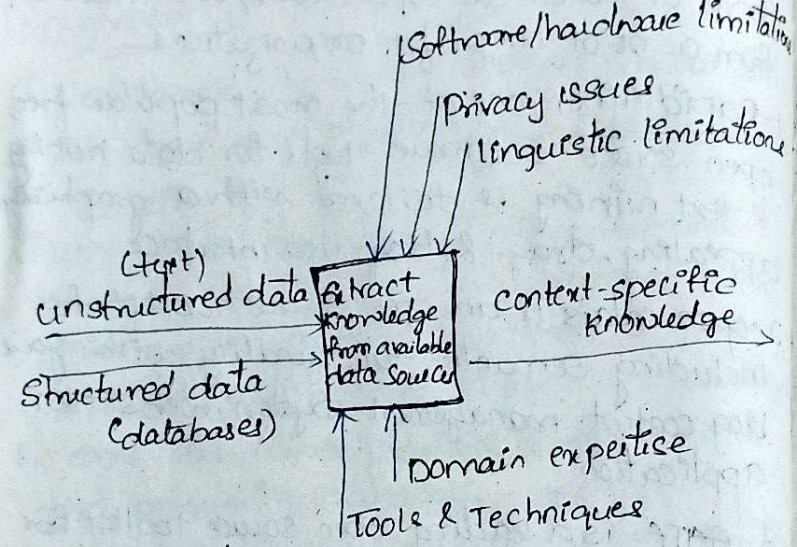
3. GATE is a leading open source toolkit for text mining. It has a free open source framework & graphical development environment

4. Lingpipe is a suite of java libraries for the linguistic analysis of human language.

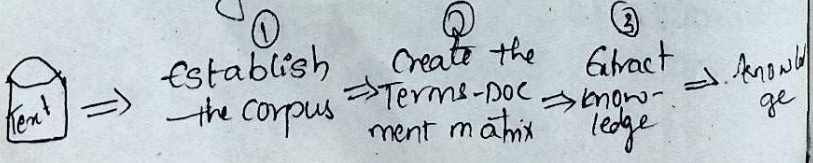
5. S-EM (spy-EM) is a text classification system that learns from positive & unlabelled examples

6. Carrot2: is a web search and text-classi clustering engine

Text Mining Process



* The text-based knowledge discovery process involves analyzing both unstructured & 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 & privacy issues & relies on techniques & expertise
* It consists of three tasks, with feedback loops for adjusting outputs if necessary.



Three step text mining process

1. **Establish the corpus:** collect & organize domain-specific unstructured data (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.

1. Establish the corpus.

The main purpose of the first task activity is to collect all of the documents related to the context being studied. This collection may include textual documents, XML files, emails, web pages & short notes, the readily available textual data. Voice recordings may also be transcribed using speech recognition algorithms & made a part of the text collection.

Once collected, the text documents are transformed & organized in a manner such that they are all in the same representation form for computer processing. The organization of the documents can be simple as a

collected of digitized text excerpts stored in a file folder or it can be a list of links to a collection of web pages in a specific domain. Many commercialy available text mining software tools could accept these as input & converts them into a flat file for processing. Alternatively, the flat file can be prepared outside the text mining software & then presented as the input to the text mining application

2. Create the Term-Document
The digitized & organized documents are used to create the term-document matrix. In the TDM rows represent the documents & columns represent the terms.
* The relationships between the terms & documents are characterized by indices

Document / Terms	investment	Project management	Software Engineering	Development	SAP	...
DOC 1	1			1		
DOC 2		1				
DOC 3			3		1	
DOC 4		1				
DOC 5			2	1		
DOC 6	1			1		
..						

Steps
1. Collect the Text Data: Gather the text documents that need to be analyzed. They could be in any formats
2. Preprocessing the Text: clean & prepare the text data by performing
→ Tokenization: Split the text into words
→ lowercasing: convert all text to lowercase to avoid treating the same word in different cases as distinct
→ Removing stop words: Remove common words that do not add much meaning
→ stemming/lemmatization: Reduce words to their base or root form
→ Removing punctuation & special characters clean up the text to focus only on words
3. Build a vocabulary: Identify all the unique terms in the corpus.
4. Create the Matrix structured
Row: each row corresponds to a document
Columns: Each column corresponds to a unique term from the vocabulary
Cells: The value in each cell represents the frequency or weight of the term in the respective document

Several ways to fill the cells

Binary encoding: Assign 1 (or) 0

Term Frequency (TF): count how many times a term appears in a document

Term Frequency-inverse document Frequency (TF-IDF): A weighted term frequency, where the TF is adjusted by how commonly the term appears across all documents.

Example

DOC 1 : "Text Mining is fun."

DOC 2 : "Text Mining involves extracting useful information"

DOC 3 : "Mining data can be fun & useful"

Step①: preprocess the Text.

* After preprocessing the tokenized & cleaned text for each document

Doc 1: ["text", "mining", "fun"]

DOC 2 : ["text", "mining", "involves", "extracting", "useful", "information"]

DOC 3 : ["mining", "data", "can", "be", "fun", "useful"]

② vocabulary

["text", "mining", "fun", "involves", "extracting", "useful", "information", "data", "can", "be"]

② Create the Term Document Matrix.

Term	DOC 1	DOC 2	DOC 3
Text	1	1	0
mining	1	1	1
fun	1	0	1
involves	0	1	0
extracting	0	1	0
useful	0	1	1
information	0	1	0
data	0	0	1
can	0	0	1
be	0	0	1

3. Extract knowledge

Goal: Discover meaningful patterns or insights from the term-document matrix

Action: use text mining algorithms 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 algorithm used include.