Advanced Social, Text and Media Analytics

By
Stuti Chug

Topics

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- **Text Mining:** Introduction, Core text mining operations, Pre-processing techniques, Categorization, Clustering, Information extraction, Probabilistic models for information extraction, Text mining applications
- **Methods and Approaches:** Content Analysis; Natural Language Processing; Clustering & Topic Detection; Simple Predictive Modelling; Sentiment Analysis; Sentiment Prediction
- **Web Analytics :** Web analytics tools, Clickstream analysis, A/B testing, online surveys; Web search and retrieval, Search engine optimization, Web crawling and Indexing, Ranking algorithms, Web traffic models
- **Social Media Analytics:** Social network and web data and methods. Graphs and Matrices. Basic measures for individuals and networks. Information visualization; Making connections: Link analysis. Random graphs and network evolution. Social contexts: Affiliation and identity; Social network analysis

Weightage

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- Quiz :- 10%
- Project :- 20 %
- MST :- 35%
- EST :- 35%

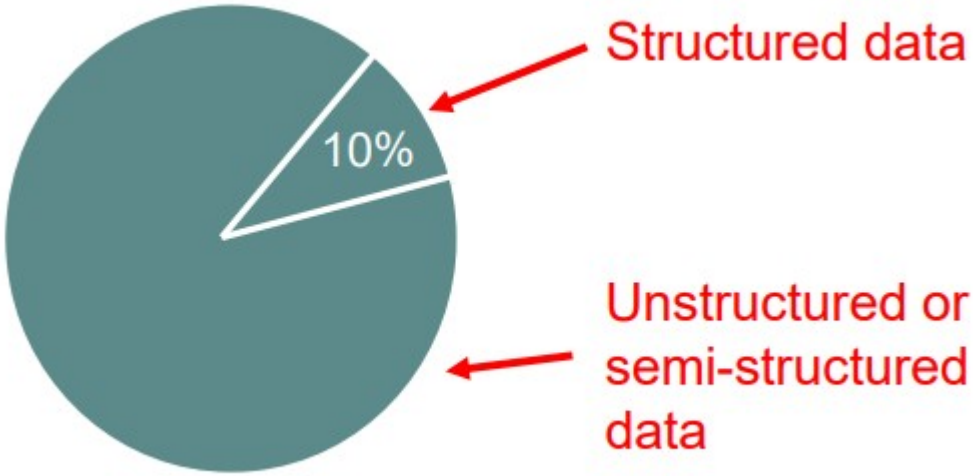
Why ?

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Approximately 90% of the world's data is held in unstructured formats

Example :

- Web pages
- Emails
- Customer complaint letters
- Scientific papers
- etc.



- **A large amount of text information can be analysed objectively and efficiently with Text mining**
- **The field of text mining has received a lot of attention due to ever increasing need for managing the information that resides in the vast amount of available text documents**

Data Mining vs Text Mining

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- Process Directly
- Identify casual Relationship
- Structured Data
- Structured numeric transaction data residing in rational data warehouse

Data Mining

- Natural Language Processing
- Discover here fore unknown information
- Semi structured and Unstructured data
- Applications deal with much more diverse and eclectic collection of

Text Mining

Information Retrieval vs Text Mining

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Information Retrieval	IR is primarily concerned with finding relevant information from a large collection of documents or data in response to a query. It's about locating and retrieving information that matches a user's search criteria. Example: When you search for "best Italian restaurants near me" on Google, IR techniques are used to provide you with a list of relevant restaurant listings based on your query.
Text Mining	Text mining, also known as text data mining, focuses on extracting useful information and patterns from text data. It involves analysing large volumes of text to uncover hidden patterns, relationships, and insights. Example: Analysing customer reviews to identify common themes and sentiments about a product or service. This might involve discovering patterns like frequent complaints about a product feature or understanding overall customer satisfaction.

Text Mining

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- In text mining, the goal is to discover unmown information, something that no one yet knows and so cold not have yet written down.
- Text mining is different from what are familiar with in web search
- In search, the user is typically looking for something that is already known and has been written by someone else.

Text Mining

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

“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. ”

For analysing vast collections of textual materials to capture key concepts, trends and hidden relationships.

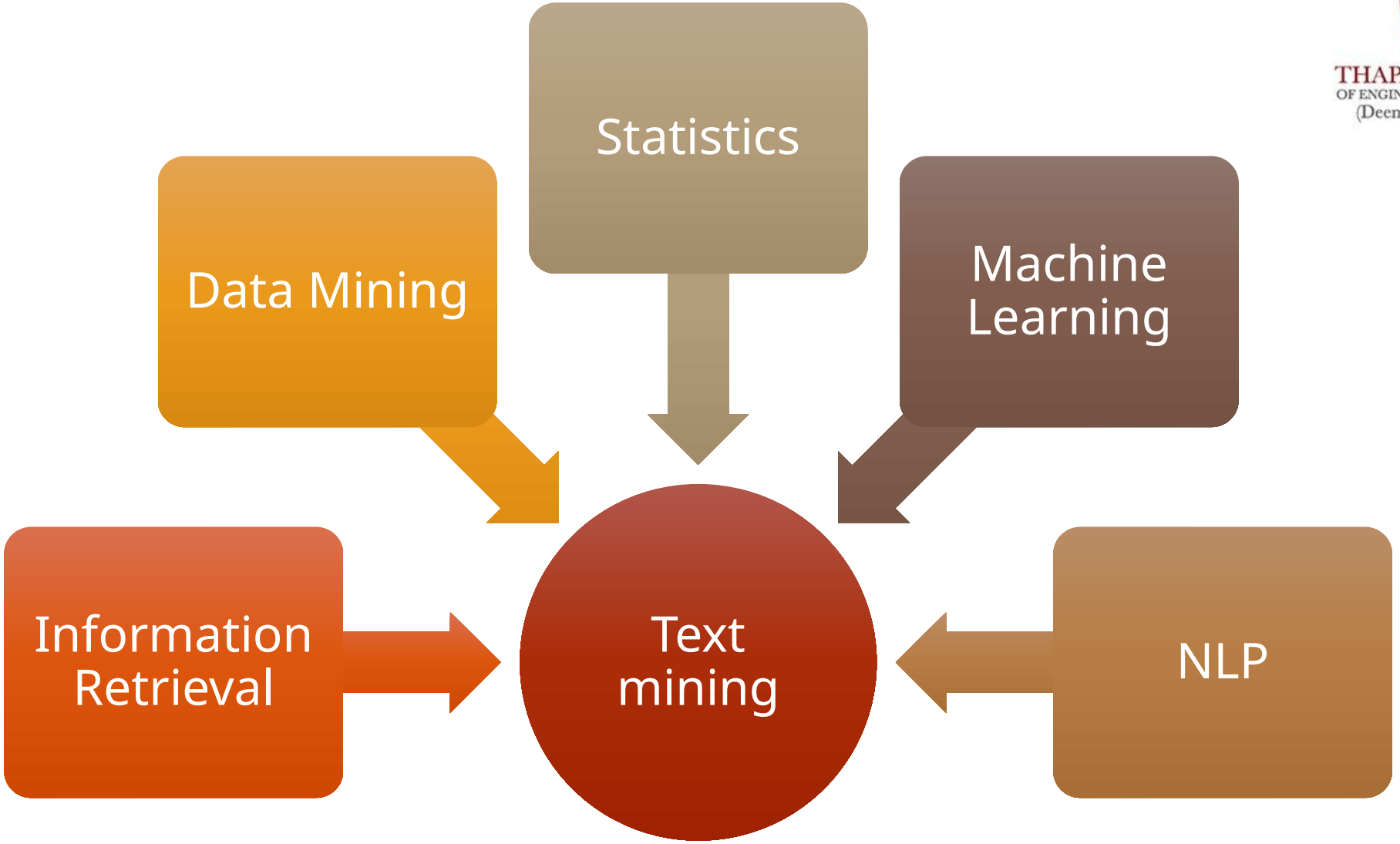
Types of Data

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- **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.

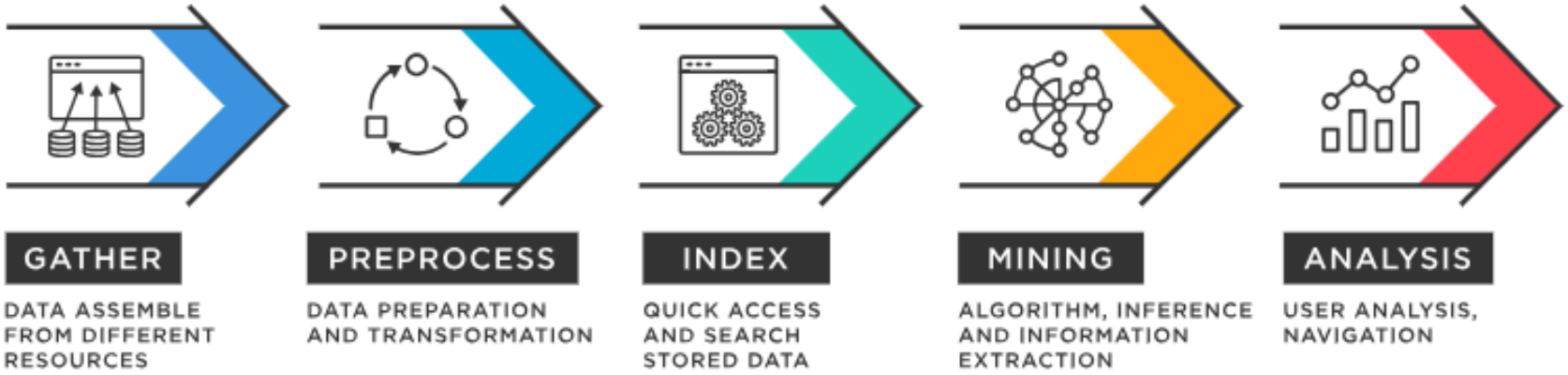
- **Semi-structured data**: As the name suggests, this data is a blend between structured and unstructured data formats. While it has some organization, it doesn’t have enough structure to meet the requirements of a relational database. Examples of semi-structured data include XML, JSON and HTML files.


Statistics
Data Mining
Machine Learning
Information Retrieval
Text mining
NLP

Steps of Text Mining

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- Pre-processing the text
- Applying text mining techniques
  - *Summarization*
  - *Classification*
  - *Clustering*
  - *Visualization*
  - *Information extraction*
- Analysing the text



Core Text Mining Operations

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- **Text Pre-processing:** This involves cleaning and preparing the text data for analysis. Steps include tokenization (breaking text into words or phrases), removing stop words (common words that add little meaning), stemming (reducing words to their root form), and lemmatization (reducing words to their base or dictionary form).
- **Feature Extraction:** Converting text into numerical features that can be used for analysis. Common techniques include:
  - ***Bag of Words (BoW):*** *Representing text by the frequency of words, disregarding grammar and word order.*
  - ***Term Frequency-Inverse Document Frequency (TF-IDF):*** *A statistical measure that evaluates the importance of a word in a document relative to a collection of documents.*
  - ***Word Embeddings:*** *Using dense vector representations (e.g., Word2Vec, GloVe) to capture semantic meaning and relationships between words.*

Cont..

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- **Text Classification:** Categorizing text into predefined categories using machine learning algorithms. Examples include spam detection, sentiment analysis, and topic classification.
- **Named Entity Recognition (NER):** Identifying and classifying entities (e.g., names, dates, locations) mentioned in the text. (The main task in NER is to extract and classify these named entities from unstructured text into predefined categories.)
- **Topic Modelling:** Discovering the underlying themes or topics within a collection of documents. Techniques include Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF).
- **Sentiment Analysis:** Determining the sentiment expressed in the text, such as positive, negative, or neutral, often used to gauge public opinion or customer feedback.

Text Pre-processing

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Before mine text, its needs to be cleaned and prepared. Pre-processing operations include:

- **Tokenization:** Splitting text into individual words or tokens. This is often the first step in preparing text for analysis.
- **Stop Word Removal:** Removing common words (e.g., "and", "the", "is") that may not contribute much meaning or value to the analysis.
- **Stemming and Lemmatization:** Reducing words to their root forms (e.g., "running" to "run") to ensure that different forms of a word are treated as the same entity.
- **Normalization:** Converting text to a standard format (e.g., lowercasing all characters, removing punctuation).
- **Named Entity Recognition (NER):** Identifying and classifying entities such as names of people, organizations, locations, dates, etc.
- **Part-of-Speech Tagging:** Assigning grammatical tags to each word (e.g., noun, verb) to understand its role in the sentence.

Text Extraction

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

This involves identifying and extracting specific pieces of information from the text:

- **Information Extraction:** Extracting structured information from unstructured text, such as extracting names, dates, and relationships between entities.
- **Keyword Extraction:** Identifying the most important words or phrases in a document that capture its key themes.
- **Phrase Extraction:** Extracting meaningful phrases or multi-word expressions from the text.
- **Concept Extraction:** Identifying key concepts or ideas from the text, which may involve recognizing and mapping terms to a predefined ontology or knowledge base.

Text Analysis

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

After pre-processing and extraction, you analyse the text to gain insights:

- **Text Classification:** Assigning predefined categories or labels to text based on its content (e.g., spam vs. non-spam emails, sentiment classification).
- **Sentiment Analysis:** Determining the sentiment expressed in the text (e.g., positive, negative, neutral). This is often used to gauge public opinion or customer feedback.
- **Topic Modeling:** Identifying themes or topics present in a collection of documents. Techniques such as Latent Dirichlet Allocation (LDA) are commonly used.
- **Clustering:** Grouping similar documents or text segments together based on their content. This can reveal underlying patterns or themes within the text data.
- **Text Summarization:** Creating a concise summary of the text while retaining its main ideas. This can be extractive (selecting key sentences) or

Pattern Discovery

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Discovering patterns and relationships within the text data:

- **Association Rule Mining:** Finding interesting relationships between different terms or entities in the text (e.g., "purchase of a laptop often leads to the purchase of a laptop bag").

- **Trend Analysis:** Identifying trends over time by analysing how topics or sentiments evolve in text data.

Visualization

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Presenting the results of text mining in a meaningful way:

- **Word Clouds:** Visual representations of word frequency, where more frequent words appear larger.
- **Topic Maps:** Visualizations showing the relationships between different topics or themes.
- **Network Graphs:** Representing relationships between entities or terms as networks to understand their connections.

Applications

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Text mining is crucial in various domains due to the exponential growth of textual data. Applications include:

- **Business Intelligence:** Analysing customer reviews, feedback, and social media to gain insights into customer sentiment and market trends.
- **Healthcare:** Extracting relevant information from medical records, research papers, and clinical notes to support decision-making and research.
- **Finance:** Monitoring news, financial reports, and social media for market sentiment analysis and risk management.
- **Legal:** Reviewing legal documents, case law, and contracts to identify relevant information and trends.

Challenges

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- **Ambiguity and Polysemy:** Words can have multiple meanings depending on context, which can complicate analysis.
- **Language Variability:** Variations in language use, including slang, jargon, and regional differences, can affect the accuracy of text mining.
- **Volume of Data:** The sheer amount of text data can be overwhelming and requires efficient processing techniques.

Text Pre-Processing

Text pre-processing is an essential step in natural language processing (NLP) that involves cleaning and transforming unstructured text data to prepare it for analysis. It includes tokenization, stemming, lemmatization, stop-word removal, and part-of-speech tagging

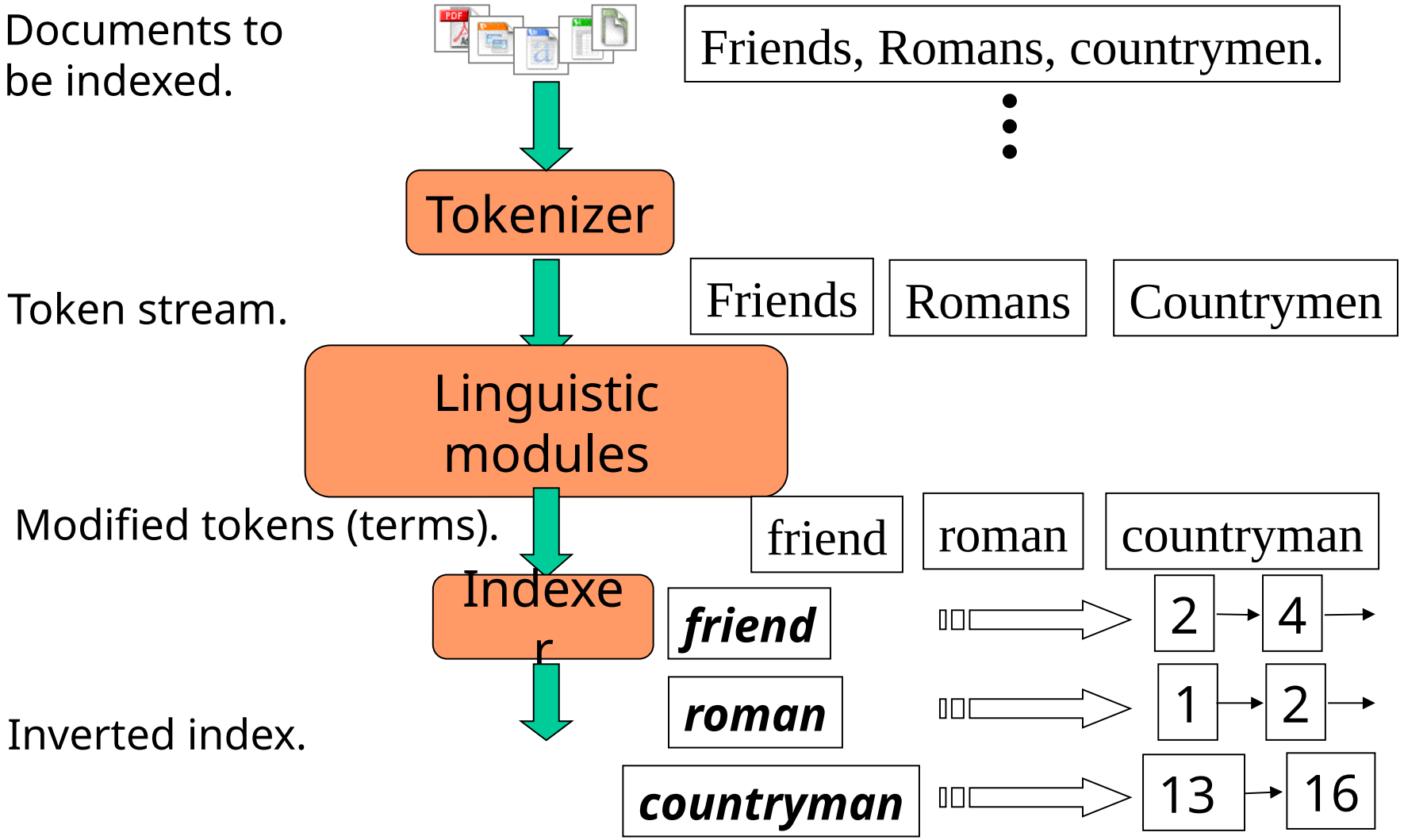
Text Mining Terminology

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- Lexicon (word dictionary)
  - *WordNet, SentiWordNet*
- Term-by-document matrix
  - *Word/document frequency*
  - *Numeric, binary, TF/IDF*
- Dimensional reduction: Singular value decomposition (SVD)
- Topic Modeling
  - *Latent Dirichlet Allocation (LDA)*
- Word Embedding
  - *Word2Vec*

- Unstructured data
- Corpus (and corpora)
- Words and Terms
- Concepts
- Stemming vs. Lemmatization
- Stop [include] words/terms
- Synonyms (and homonyms)
- Morphology
- Tokenization
- Part-of-speech tagging
- Bag of Words

Basic Indexing Pipeline



Lowercasing

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- Purpose: Normalize the text by converting all characters to lowercase. This helps in reducing the complexity of text data by treating "Word" and "word" as the same token.
- Example: "Hello World!" becomes "hello world!".

Tokenization

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- Purpose: Break down text into smaller units such as words, sentences, or subwords. This helps in analyzing individual components of the text
- .Example: "Hello, world!" is tokenized into ["Hello", ",", "world", "!"].

Removing Punctuation

- Purpose: Remove punctuation marks as they often do not contribute meaningful information to text analysis.
- Example: "Hello, world!" becomes "Hello world".

Removing Stop Words

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- Purpose: Remove common words (like "and", "the", "is") that are considered to be of little value in text analysis.
- A stop list is a list of such words that are removed during lexical analysis.
- Example: "This is a sample sentence." becomes "sample sentence.".

- However, common words are sometimes significant in information retrieval, which is an argument for a short stop list. (Consider the query, "To be or not to be?")

Suggestions for Including Words in a Stop List

- Include the most common words in the English language (perhaps 50 to 250 words).
- Do not include words that might be important for retrieval (Among the 200 most frequently occurring words in general literature in English are time, war, home, life, water, and world).
- In addition, include words that are very common in context (e.g., computer, information, system in a set of computing documents).

Stemming

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- Purpose: Reduce words to their root or base form by stripping affixes. This helps in grouping different forms of a word.
- Example: "running", "runner", and "ran" are reduced to "run".

Lemmatization

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- Purpose: Reduce words to their base or dictionary form, considering the context and meaning of the word. Unlike stemming, lemmatization ensures that the resulting words are actual words.
- Example: "running" becomes "run", and "better" becomes "good".

Removing Numbers

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- Purpose: Depending on the task, numbers may be irrelevant and can be removed to focus on textual content
- Example: "My address is 123 Main St." becomes "My address is Main St.".

- **Handling White Spaces**
  - *Purpose: Remove or normalize unnecessary white spaces to ensure consistency in the text.*
  - *Example: "Hello world " becomes "Hello world".*

Text Normalization

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- Purpose: Standardize text by correcting misspellings, expanding contractions, and normalizing abbreviations.
- Example: "don’t" becomes "do not", and "info" becomes "information".

Named Entity Recognition (NER)

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- Purpose: Identify and classify entities such as names of people, organizations, locations, etc., if relevant to the analysis.
- Example: "Barack Obama was the President of the United States." extracts ["Barack Obama": PERSON, "United States": LOCATION].

Part-of-Speech Tagging

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- Purpose: Assign parts of speech (e.g., noun, verb) to each word to understand the grammatical structure.
- Example: "The quick brown fox jumps over the lazy dog." might be tagged as ["The: DET", "quick: ADJ", "brown: ADJ", "fox: NOUN", "jumps: VERB", "over: PREP", "the: DET", "lazy: ADJ", "dog: NOUN"].

Handling Synonyms and Antonyms

- Purpose: Standardize synonyms and antonyms to a common term to reduce redundancy and improve consistency.
- Example: "happy" and "joyful" might be standardized to "happy".

Text Vectorization

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- Purpose: Convert text into numerical representations for machine learning models, such as Bag of Words, TF-IDF, or word embedding.
- Example: "I love machine learning" might be represented as a vector of word frequencies or embedding in a high-dimensional space.

Bag of Words

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- A bag of words is a representation of text that describes the occurrence of words within a document.
- keep track of word counts and disregard the grammatical details and the word order.
- It is called a “bag” of words because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not where in the document.

Why ?

- One of the biggest problems with text is that it is messy and unstructured, and machine learning algorithms prefer structured, well defined fixed-length inputs and by using the Bag-of-Words technique we can convert variable-length texts into a fixed-length vector.

Example :

- Sentence 1: ”Welcome to Great Learning, Now start learning”
- Sentence 2: “Learning is a good practice”

Sentence 1	Sentence 2
Welcome	Learning
to	is
Great	a
Learning	good
,	practice
Now	
start	
learning	

Cont…

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- Go through all the words in the above text and make a list of all of the words in our model vocabulary.
  - *Welcome*
  - *To*
  - *Great*
  - *Learning*
  - *,*
  - *Now*
  - *start*
  - *learning*
  - *is*
  - *a*
  - *good*
  - *practice*

Cont…

Word	Frequency
Welcome	1
to	1
Great	1
Learning	1
,	1
Now	1
start	1
learning	1
is	0
a	0
good	0
practice	0

Sentence 1 → **[1,1,1,1,1,1,1,1,0,0,0]**

Word	Frequency
Welcome	0
to	0
Great	0
Learning	1
,	0
Now	0
start	0
learning	0
is	1
a	1
good	1
practice	1

Sentence 2 → **[0,0,0,0,0,0,0,1,1,1,1,1]**

With pre-processing:

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- Sentence 1: ”welcome great learning now start learning”
- Sentence 2: “learning good practice”
- model vocabulary.
  - *welcome*
  - *great*
  - *learning*
  - *now*
  - *start*
  - *good*
  - *practice*

Cont…

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Word	Frequency
welcome	1
great	1
learning	2
now	1
start	1
good	0
practice	0

Sentence 1 → **[1,1,2,1,1,0,0]**

Word	Frequency
welcome	0
great	0
learning	1
now	0
start	0
good	1
practice	1

Sentence 2 → **[0,0,1,0,0,1,1]**

TF-IDF

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- TF-IDF (Term Frequency-Inverse Document Frequency) is a widely used technique in text mining and information retrieval that helps to evaluate the importance of a word in a document relative to a collection of documents or corpus. It combines two metrics:
- Term Frequency (TF): Measures how frequently a term occurs in a document. It's typically computed as:

$$TF(t,d)=$$

- **Inverse Document Frequency (IDF):** Measures how important a term is across the entire corpus. It's computed as:

$$IDF(t,D)=\log$$

where D is the total number of documents in the corpus.

Cont...

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- **TF-IDF Score:** Combines TF and IDF to weigh the term's importance in the document relative to the entire corpus:

$$\text{TF-IDF}(t,d,D)=\text{TF}(t,d)\times\text{IDF}(t,D)$$

Example

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Suppose we have a small corpus consisting of three documents:

- **Document 1:** "The cat in the hat."
- **Document 2:** "The quick brown fox."
- **Document 3:** "The cat and the hat are great.“

Step 1: Calculate Term Frequency (TF)

- **Document 1: "The cat in the hat."**
- **Total terms:** 5
- Term Frequency (TF) for each term:
  - *TF("the", Doc1) = 3/5 = 0.6*
  - *TF("cat", Doc1) = 1/5 = 0.2*
  - *TF("in", Doc1) = 1/5 = 0.2*
  - *TF("hat", Doc1) = 1/5 = 0.2*

Cont...

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- **Document 2: "The quick brown fox."**
- **Total terms:** 4
- Term Frequency (TF) for each term:
  - *TF("the", Doc2) = 1/4 = 0.25*
  - *TF("quick", Doc2) = 1/4 = 0.25*
  - *TF("brown", Doc2) = 1/4 = 0.25*
  - *TF("fox", Doc2) = 1/4 = 0.25*

- **Document 3: "The cat and the hat are great."**
- **Total terms:** 7
- Term Frequency (TF) for each term:
  - *TF("the", Doc3) = 2/7 ≈ 0.286*
  - *TF("cat", Doc3) = 1/7 ≈ 0.143*
  - *TF("and", Doc3) = 1/7 ≈ 0.143*
  - *TF("hat", Doc3) = 1/7 ≈ 0.143*
  - *TF("are", Doc3) = 1/7 ≈ 0.143*
  - *TF("great", Doc3) = 1/7 ≈ 0.143*

Cont...

Step 2: Calculate Inverse Document Frequency (IDF)

- Total number of documents in the corpus (D): 3
- IDF Calculation for each term:
  - *IDF("the") = log(3/3) = log(1) = 0*
  - *IDF("cat") = log(3/2) ≈ log(1.5) ≈ 0.176*
  - *IDF("in") = log(3/1) ≈ log(3) ≈ 0.477*
  - *IDF("hat") = log(3/2) ≈ 0.176*
  - *IDF("quick") = log(3/1) ≈ 0.477*
  - *IDF("brown") = log(3/1) ≈ 0.477*
  - *IDF("fox") = log(3/1) ≈ 0.477*
  - *IDF("and") = log(3/1) ≈ 0.477*
  - *IDF("are") = log(3/1) ≈ 0.477*
  - *IDF("great") = log(3/1) ≈ 0.477*

Cont...

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Step 3: Calculate TF-IDF

- Example: TF-IDF Calculation for "cat" in Document 1
- TF("cat", Doc1) = 0.2
- IDF("cat") ≈ 0.176
- TF-IDF("cat", Doc1) = TF("cat", Doc1) * IDF("cat") = 0.2 * 0.176 ≈ 0.035

Interpretation

- The TF-IDF score for "cat" in Document 1 is approximately 0.035. This score indicates that "cat" is relatively less significant in Document 1 compared to other terms in the corpus, likely because "cat" appears in Document 3 as well.
- TF-IDF scores help in identifying the importance of terms within a document relative to a corpus. Terms with higher TF-IDF scores are more important or unique to the document, while terms with lower scores are either common across many documents or less significant within the document.

One-Hot Encoding

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- One-hot encoding is a simple method for representing words in natural language processing (NLP).
- In this encoding scheme, each word in the vocabulary is represented as a unique vector, where the dimensionality of the vector is equal to the size of the vocabulary. The vector has all elements set to 0, except for the element corresponding to the index of the word in the vocabulary, which is set to 1.

Example

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- Vocabulary: {'mat', 'the', 'bird', 'hat', 'on', 'in', 'cat', 'tree', 'dog'}
- Word to Index Mapping: {'mat': 0, 'the': 1, 'bird': 2, 'hat': 3, 'on': 4, 'in': 5, 'cat': 6, 'tree': 7, 'dog': 8}
- One-Hot Encoded Matrix:
  - *cat: [0, 0, 0, 0, 0, 0, 1, 0, 0]*
  - *in: [0, 0, 0, 0, 0, 1, 0, 0, 0]*
  - *the: [0, 1, 0, 0, 0, 0, 0, 0, 0]*
  - *hat: [0, 0, 0, 1, 0, 0, 0, 0, 0]*
  - *dog: [0, 0, 0, 0, 0, 0, 0, 0, 1]*
  - *on: [0, 0, 0, 0, 1, 0, 0, 0, 0]*
  - *mat: [1, 0, 0, 0, 0, 0, 0, 0, 0]*
  - *bird: [0, 0, 1, 0, 0, 0, 0, 0, 0]*
  - *in: [0, 0, 0, 0, 0, 1, 0, 0, 0]*
  - *tree: [0, 0, 0, 0, 0, 0, 0, 1, 0]*

Cont...

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

While one-hot encoding is a simple and intuitive method for representing words in NLP, it has several disadvantages, which may limit its effectiveness in certain applications.

- One-hot encoding results in high-dimensional vectors, making it computationally expensive and memory-intensive, especially with large vocabularies.
- It does not capture semantic relationships between words; each word is treated as an isolated entity without considering its meaning or context.

Word Embedding

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- Word Embedding is an approach for representing words and documents. Word Embedding or Word Vector is a numeric vector input that represents a word in a lower-dimensional space. It allows words with similar meanings to have a similar representation.
- The methods such as Bag of Words (BOW) and TFIDF rely on the word count in a sentence but do not save any syntactical or semantic information.
- **Need for Word Embedding?**
  - *To reduce dimensionality*
  - *To use a word to predict the words around it.*
  - *Inter-word semantics must be captured.*

- It is a neural approach for generating word embedding. It belongs to the family of neural word embedding techniques and specifically falls under the category of distributed representation models.
- It is a popular technique in natural language processing (NLP) that is used to represent words as continuous vector spaces. Developed by a team at Google, Word2Vec aims to capture the semantic relationships between words by mapping them to high-dimensional vectors. The underlying idea is that words with similar meanings should have similar vector representations. In Word2Vec every word is assigned a vector. We start with either a random vector or one-hot vector..
- There are two neural embedding methods for Word2Vec, Continuous Bag of Words (CBOW) and Skip-gram.

- These are basically shallow neural networks that have an input layer, an output layer, and a projection layer. It reconstructs the linguistic context of words by considering both the order of words in history as well as the future.
- The method involves iteration over a corpus of text to learn the association between the words. It relies on a hypothesis that the neighboring words in a text have semantic similarities with each other. It assists in mapping semantically similar words to geometrically close embedding vectors.
- It uses the cosine similarity metric to measure semantic similarity. Cosine similarity is equal to Cos(angle) where the angle is measured between the vector representation of two words/documents.
- So if the cosine angle is one, it means that the words are overlapping.
- And if the cosine angle is a right angle or 90°, It means words hold no contextual similarity and are independent of each other.

Continuous Bag of Words(CBOW)

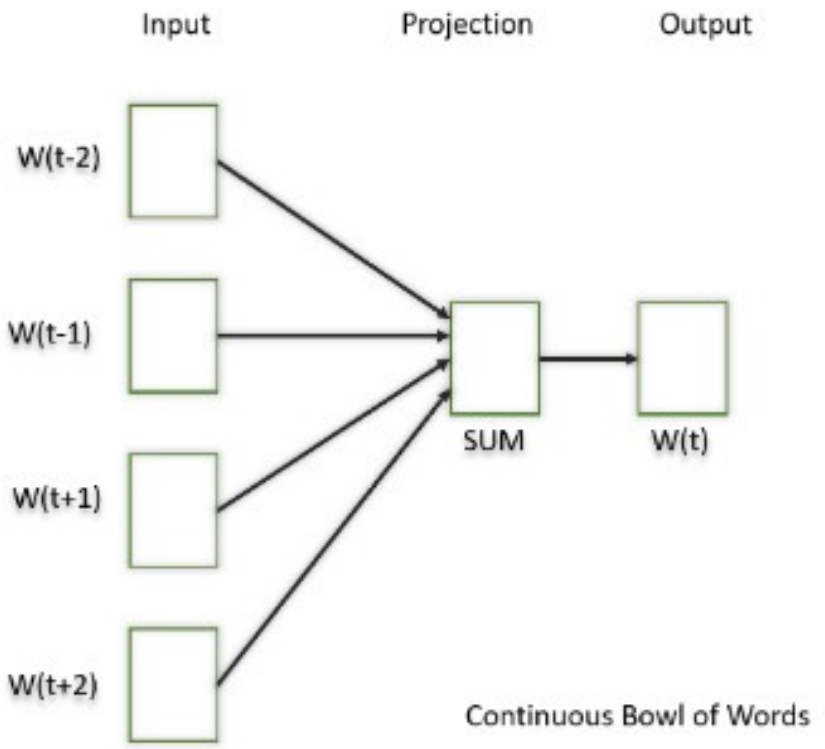
- It is a type of neural network architecture used in the Word2Vec model. The primary objective of CBOW is to predict a target word based on its context, which consists of the surrounding words in a given window. Given a sequence of words in a context window, the model is trained to predict the target word at the center of the window.
- CBOW is a feedforward neural network with a single hidden layer. The input layer represents the context words, and the output layer represents the target word. The hidden layer contains the learned continuous vector representations (word embeddings) of the input words.

Cont...

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

The hidden layer contains the continuous vector representations (word embeddings) of the input words.

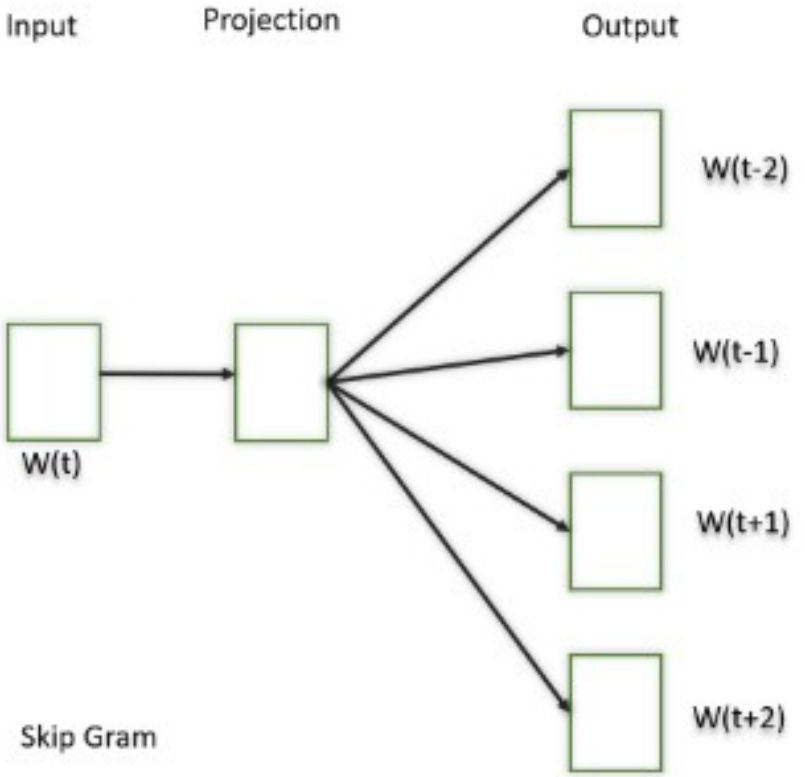
- The weights between the input layer and the hidden layer are learned during training.
- The dimensionality of the hidden layer represents the size of the word embeddings (the continuous vector space).



Skip-Gram

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- It learns distributed representations of words in a continuous vector space. The main objective of Skip-Gram is to predict context words (words surrounding a target word) given a target word. This is the opposite of the Continuous Bag of Words (CBOW) model, where the objective is to predict the target word based on its context. It is shown that this method produces more meaningful embedding.



THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- After applying the above neural embedding methods we get trained vectors of each word after many iterations through the corpus. These trained vectors preserve syntactical or semantic information and are converted to lower dimensions. The vectors with similar meaning or semantic information are placed close to each other in space.
- CBOW might be preferred when training resources are limited, and capturing syntactic information is important. Skip-gram, on the other hand, might be chosen when semantic relationships and the representation of rare words are crucial.

Pretrained Word-Embedding

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- Pre-trained word embeddings are representations of words that are learned from large corpora and are made available for reuse in various natural language processing (NLP) tasks. These embeddings capture semantic relationships between words, allowing the model to understand similarities and relationships between different words in a meaningful way.

GloVe

- GloVe is trained on global word co-occurrence statistics. It leverages the global context to create word embeddings that reflect the overall meaning of words based on their co-occurrence probabilities. this method, we take the corpus and iterate through it and get the co-occurrence of each word with other words in the corpus. We get a co-occurrence matrix through this. The words which occur next to each other get a value of 1, if they are one word apart then 1/2, if two words apart then 1/3 and so on.

Example

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Corpus:
It is a nice evening.
Good Evening!
Is it a nice evening?

	it	is	a	nice	evening	good
it	0					
is	1+1	0				
a	1/2+1	1+1/2	0			
nice	1/3+1/2	1/2+1/3	1+1	0		
evening	1/4+1/3	1/3+1/4	1/2+1/2	1+1	0	
good	0	0	0	0	1	0

Cont...

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- Initially, the vectors for each word is assigned randomly. Then we take two pairs of vectors and see how close they are to each other in space. If they occur together more often or have a higher value in the co-occurrence matrix and are far apart in space then they are brought close to each other. If they are close to each other but are rarely or not frequently used together then they are moved further apart in space.
- After many iterations of the above process, we'll get a vector space representation that approximates the information from the co-occurrence matrix. The performance of GloVe is better than Word2Vec in terms of both semantic and syntactic capturing.