TEXT MINING

Guide:-G. Appa Rao Sir
Srinivasa Chakravarthy Sir
Navya Sree Mam
SRI VIBHAV MADDAMASETTY
MADHU SPURTHI SURAPELLI
SURI VENKATA DEEPAK KUMAR REDDY
RUDRARAJU JEETENDRA VARMA

INTRODUCTION

- Text Mining is also known as text data mining.
- Text Mining is the process of transforming unstructured text into a structured format to identify meaningful patterns and new insights.
- By applying advanced analytical techniques, such as Naïve Bayes, Support Vector Machines (SVM), and other deep learning algorithms, companies are able to explore and discover hidden relationships within their unstructured data.
- Since 80% of data in the world resides in an unstructured format, text mining is an extremely valuable practice within organizations.

TEXT MINING

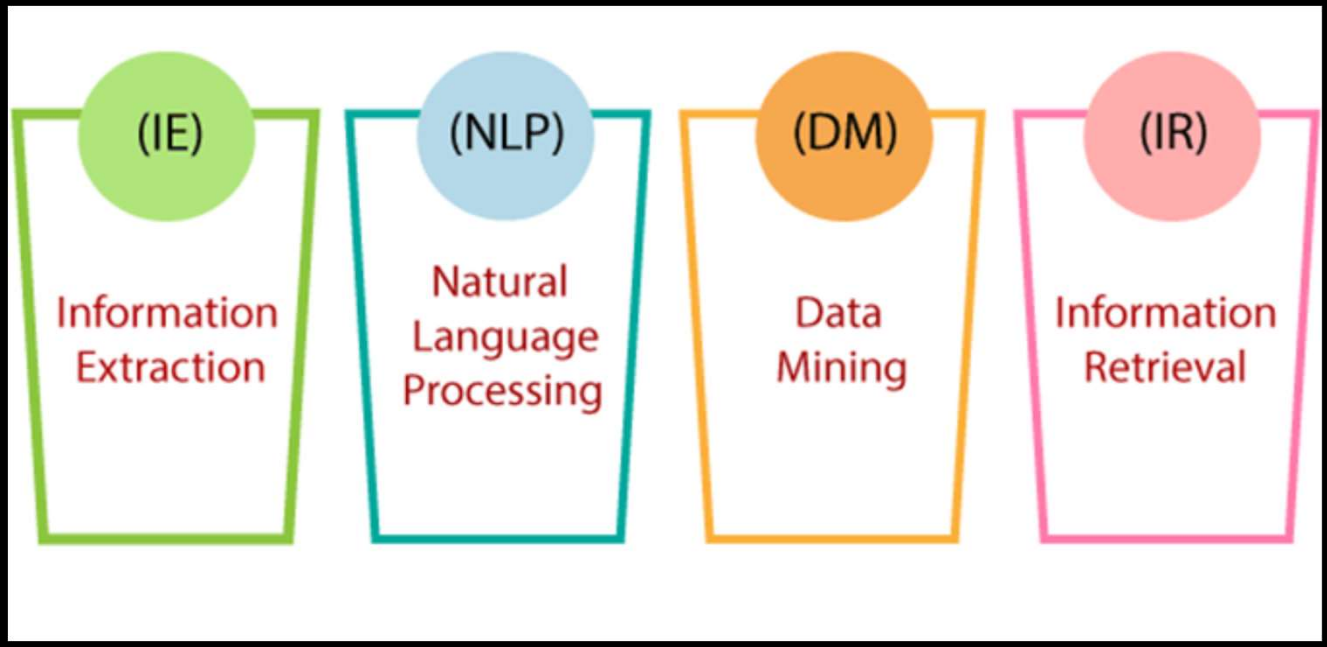

developed
features
Data
user
erent
order
techniques
field
data
visual
within
loss
human
language
novel
leverages
contexts
information
Text
Mining
Abstraction
leverage
patterns
knowledge
algorithms
learning
linguistics
problem
machine
common
another
method

AREAS OF TEXT MINING

The following are the four main areas of text mining: -

1. Information Extraction
2. Natural Language Processing
3. Data Mining
4. Information Retrieval

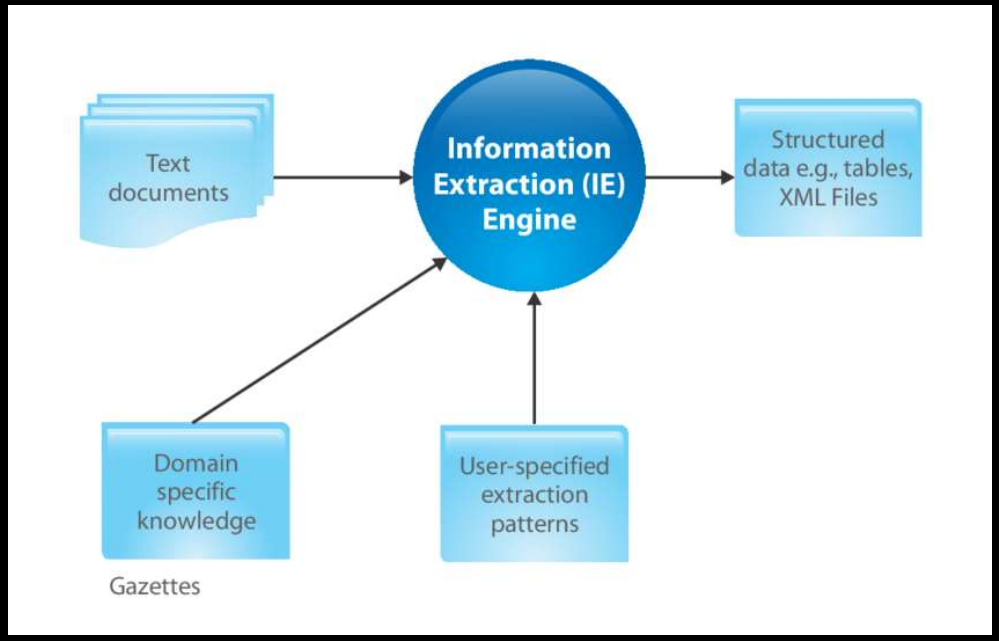
AREAS OF TEXT MINING



INFORMATION EXTRACTION

- Information extraction (IE) surfaces the relevant pieces of data when searching various documents.
- It also focuses on extracting structured information from free text and storing these entities, attributes, and relationship information in a database.
- Common information extraction sub-tasks include:
    1. **Feature selection,** or attribute selection, is the process of selecting the important features (dimensions) to contribute the most to output of a predictive analytics model.
    2. **Feature extraction** is the process of selecting a subset of features to improve the accuracy of a classification task. This is particularly important for dimensionality reduction.
    3. **Named-entity recognition (NER)** also known as entity identification or entity extraction, aims to find and categorize specific entities in text, such as names or locations. For example, NER identifies “California” as a location and “Mary” as a woman’s name.

INFORMATION EXTRACTION



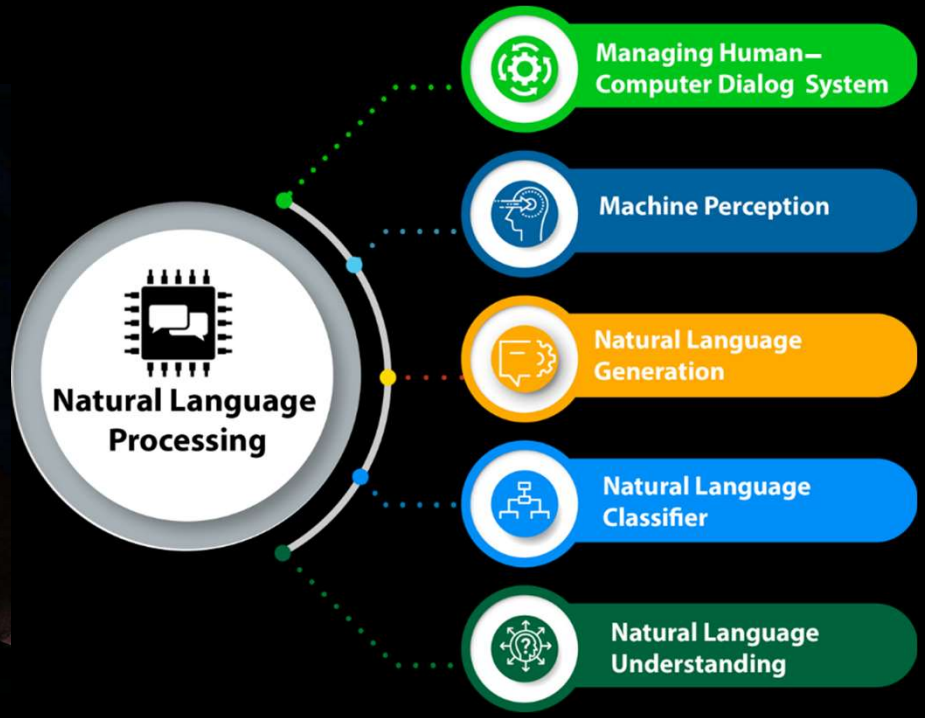
NATURAL LANGUAGE PROCESSING

- Natural language processing uses methods from various disciplines, such as computer science, artificial intelligence, linguistics, and data science, to enable computers to understand human language in both written and verbal forms.
- By analyzing sentence structure and grammar, NLP sub-tasks allow computers to “read”. Common sub-tasks include:
    1. **Summarization:** This technique provides a synopsis of long pieces of text to create a concise, coherent summary of a document’s main points.
    2. **Part-of-Speech (PoS) tagging:** This technique assigns a tag to every token in a document based on its part of speech—i.e. denoting nouns, verbs, adjectives, etc. This step enables semantic analysis on unstructured text.

NATURAL LANGUAGE PROCESSING

3. **Text categorization**: This task, which is also known as text classification, is responsible for analyzing text documents and classifying them based on predefined topics or categories. This sub-task is particularly helpful when categorizing synonyms and abbreviations.
4. **Sentiment analysis**: This task detects positive or negative sentiment from internal or external data sources, allowing you to track changes in customer attitudes over time. It is commonly used to provide information about perceptions of brands, products, and services. These insights can propel businesses to connect with customers and improve processes and user experiences.

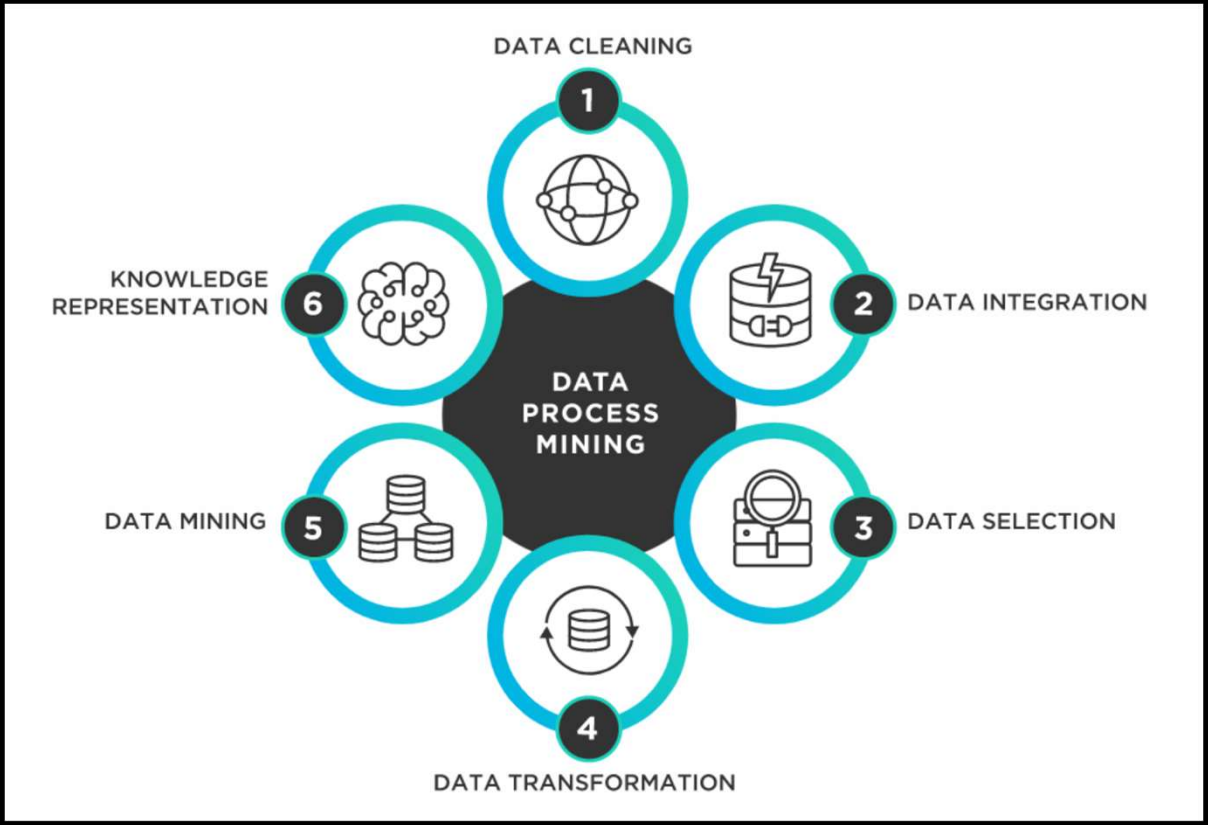
NATURAL LANGUAGE PROCESSING



DATA MINING

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

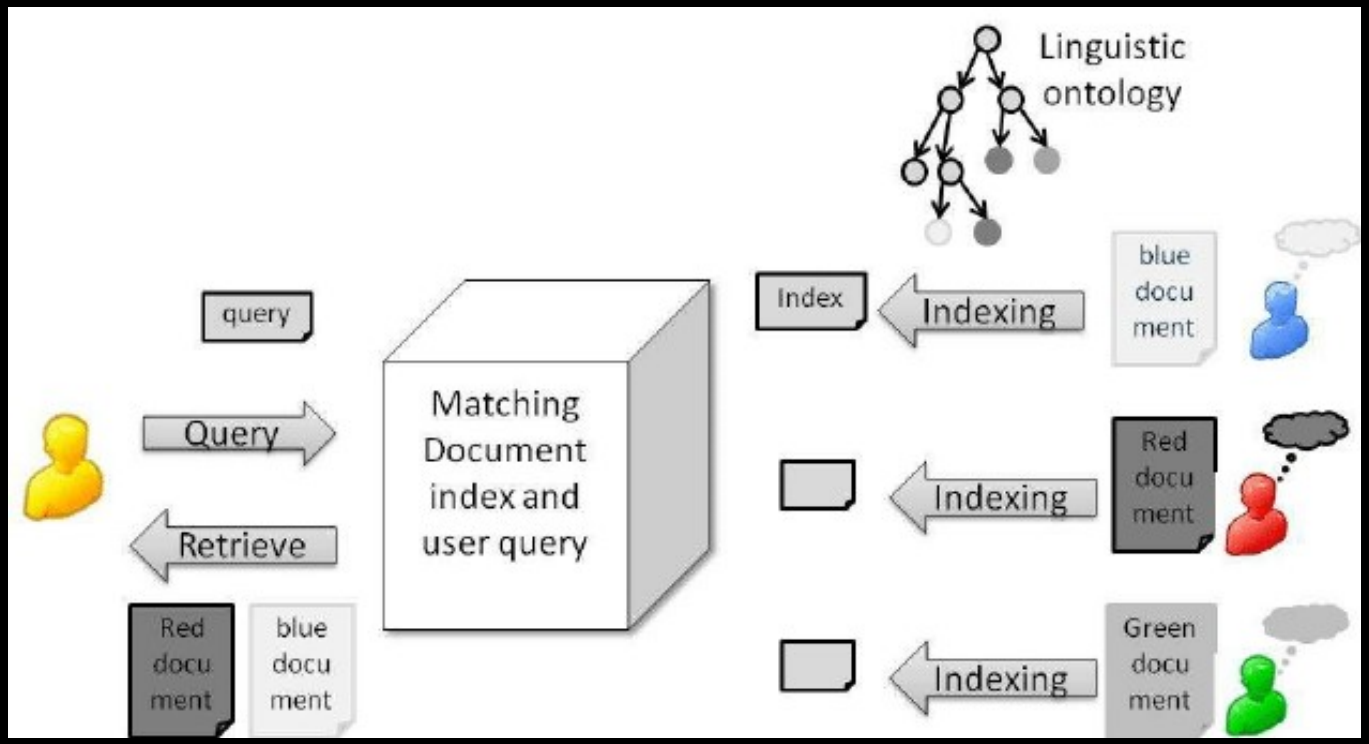
DATA MINING


DATA CLEANING
1
2
DATA INTEGRATION
3
DATA SELECTION
4
DATA TRANSFORMATION
5
DATA MINING
6
KNOWLEDGE REPRESENTATION
DATA PROCESS MINING

INFORMATION RETRIEVAL

- Information retrieval (IR) returns relevant information or documents based on a pre-defined set of queries or phrases.
- IR systems utilize algorithms to track user behaviors and identify relevant data.
- Information retrieval is commonly used in library catalogue systems and popular search engines like Google.
- Some common IR sub-tasks include:
    1. **Tokenization:** This is the process of breaking out long-form text into sentences and words called “tokens”. These are, then, used in the models, like bag-of-words, for text clustering and document matching tasks.
    2. **Stemming:** This refers to the process of separating the prefixes and suffixes from words to derive the root word form and meaning. This technique improves information retrieval by reducing the size of indexing files.

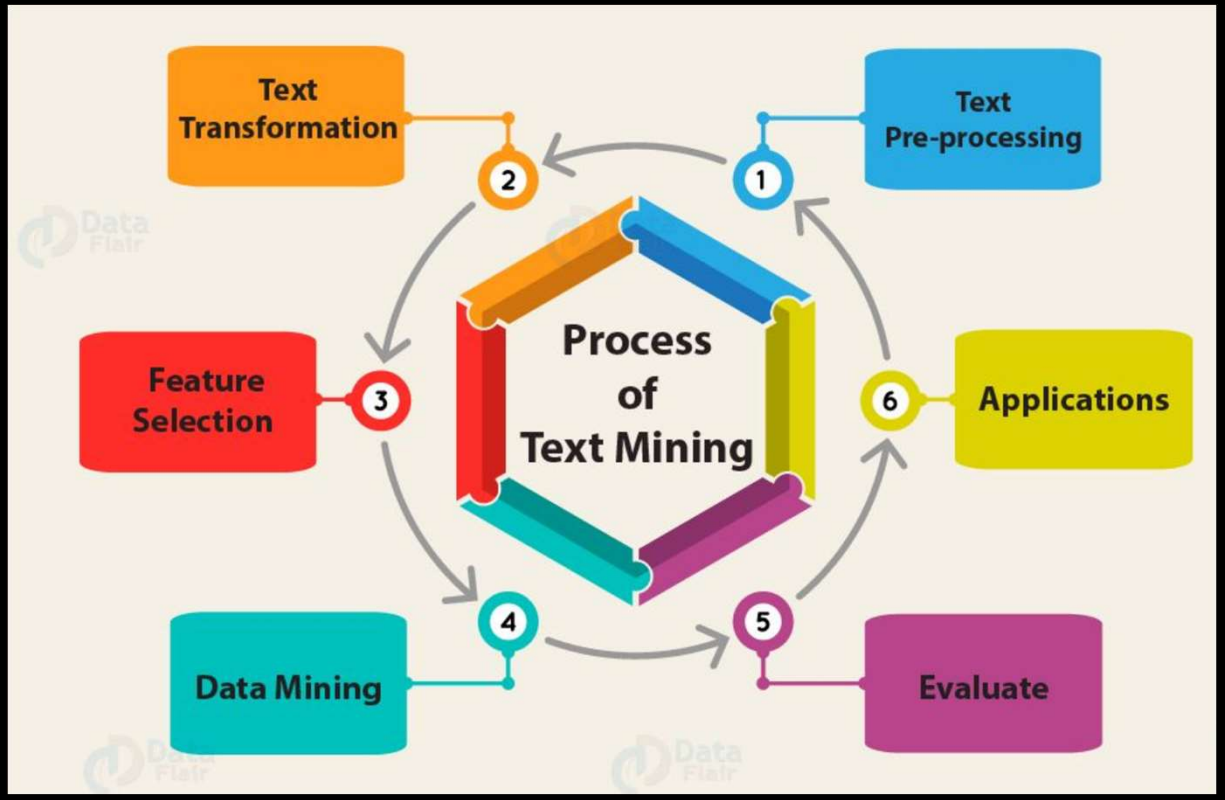
INFORMATION RETRIEVAL



TEXT MINING PROCESS

- A process of Text mining involves a series of activities to perform to mine the information. These activities are:
    1. Text Pre-Processing
    2. Text Transformation
    3. Feature Selection
    4. Data Mining
    5. Evaluate
    6. Applications

TEXT MINING PROCESS



TEXT PRE-PROCESSING

- It involves a series of steps as shown in below:
  1. **Text Cleanup-** Text Cleanup means removing any unnecessary or unwanted information. Such as remove ads from web pages, normalize text converted from binary formats.
  2. **Tokenization-** Tokenizing is simply achieved by splitting the text into white spaces.
  3. **Part of Speech Tagging**- Part-of-Speech (POS) tagging means word class assignment to each token. Its input is given by the tokenized text. Taggers have to cope with unknown words and ambiguous word-tag mappings.

TEXT TRANSFORMATOIN

- A text document is represented by the words it contains and their occurrences. Two main approaches to document representation are:
    1. Bag of words-For a given document, you extract only the unigram words to create an unordered list of words. Only the unigram words themselves, making for a bunch of words to represent the document is called Bag of words.
    2. Vector Space-Given the bag of words that you extracted from the document, you create a feature vector for the document, where each feature is a word and the feature's value is a term weight. The term weight might be a binary value (with 1 indicating that the term occurred in the document and 0 indicating that it did not), a term frequency value (indicating how many times the term occurred in the document) or a TF-IDF value (a small floating-point number like 1.23).

FEATURE SELECTION

- Feature selection also is known as variable selection.
- It is the process of selecting a subset of important features for use in model creation.
- Redundant features are the ones which provides no extra information.
- Irrelevant features provide no useful or relevant information in any context.
- In the feature selection method, only the relevant features are selected.

DATA MINING

- At this point, the Text mining process merges with the traditional process.
- Classic Data Mining techniques are used in the structured database.
- Several methods can be used to classify the documents based on different categories like keywords, author name etc.
- The decision tree text classifier is based on a tree structure in which the inner vertices are labelled by terms, their branches are obtained by checking the word weights in the test dataset, the leaf vertices are the categories.
- Because deep learning approaches are among the most important and highly used in NLP (Natural Language Processing), the algorithm H2O's Deep Learning can also be applied.

TEXT MINING APPLICATIONS

1. **Web Mining**

   These days web contains a treasure of information about subjects such as persons, companies, organizations, products, etc. that may be of wide interest. Web Mining is an application of data mining techniques. That need to discover hidden and unknown patterns from the Web. Web mining is an activity of identifying term implied in a large document collection.

TEXT MINING APPLICATIONS

2. **Resume Filtering**

Big enterprises and headhunters receive thousands of resumes from job applicants every day. Extracting information from resumes with high precision and recall is not easy. Automatically extracting this information can the first step in filtering resumes. Hence, automating the process of resume selection is an important task.

TEXT MINING APPLICATIONS

3. Fraud Detection

Text analytics backed up by text mining techniques provides a tremendous opportunity for domains that gather a majority of data in the text format. Insurance and finance companies are harnessing this opportunity. By combining the outcomes of text analyses with relevant structured data these companies are now able to process claims swiftly as well as to detect and prevent frauds.

THANK YOU