

Text Mining

- extracting **essential information** from standard language **text data**.

What is Text Mining?

- Text data mining can be described as the process of **extracting** essential data from standard language text.
- All the data that we generate via :
 - **Text messages**
 - **Text documents**
 - **Emails**
 - **Files**
- The primary source of data is *e-commerce websites, social media platforms, published articles, survey, and many more.*
- The larger part of the generated data is **unstructured**, which makes it **challenging** and **expensive** for the organizations to analyze with the help of the people.

What is Text Mining?

- Text mining, also referred to as text data mining, roughly equivalent to **text analytics**, is the process of deriving **high-quality** information from text.
- Text mining uses natural language processing (**NLP**), allowing machines to understand the human language and process it automatically.
- Natural language processing (**NLP**) is the ability of a computer program to understand human language as it is spoken and written. It is a component of artificial intelligence (**AI**).

Why is Text Mining Important?

- Individuals and organizations generate tons of data every day. Stats claim that almost **80% of the existing text data is unstructured**, meaning it's not organized in a predefined way, it's not searchable, and it's almost impossible to manage. In other words, it's just not useful.

How Does Text Mining Work?

- Text mining helps to analyze large amounts of raw data and find relevant insights. Combined with machine learning, it can create text analysis models that learn to classify or extract specific information based on previous training.
- The **first step** in text mining is **collecting** or **gathering the data**.
- Data can be **internal** (interactions through chats, emails, surveys, spreadsheets, databases, etc) or **external** (information from social media, review sites, news outlets, and any other websites).

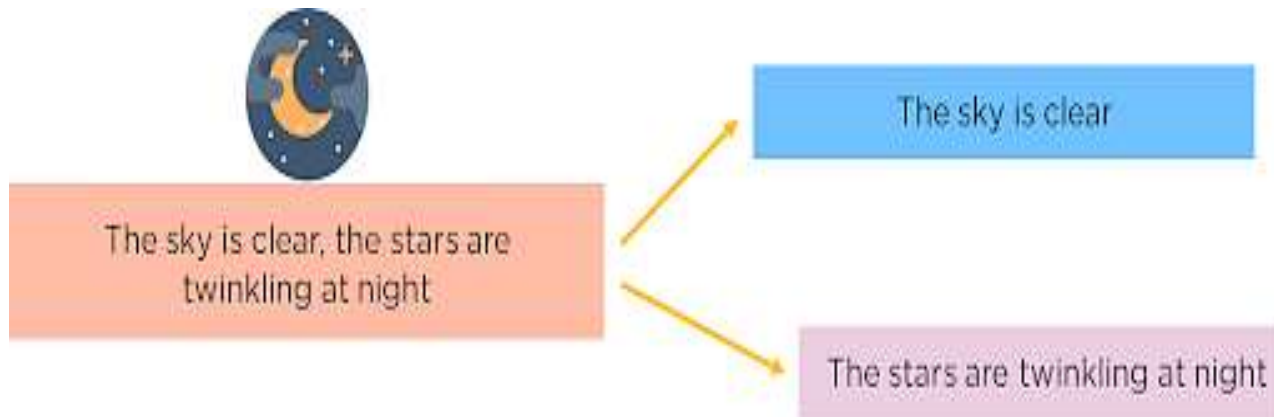
How Does Text Mining Work?

- The **second step** is preparing(**preprocessing**) your data. Text mining systems use several **NLP techniques** — like Segmentation, tokenization, lemmatization, stemming and stop removal — to build the inputs of your **machine learning model**.

The steps to perform **preprocessing** of data :

- **Segmentation:**

- Break the entire document/article into its component sentences by its punctuations like full stops and commas.



The steps to perform **preprocessing** of data :

- **Tokenizing:**

- Tokenization is a process of splitting a text / sentence into smaller units(words) which are also called tokens.



The star are twinkling at night

The

stars

are

twinkling

at

night

The steps to perform **preprocessing** of data :

- **Stemming:**

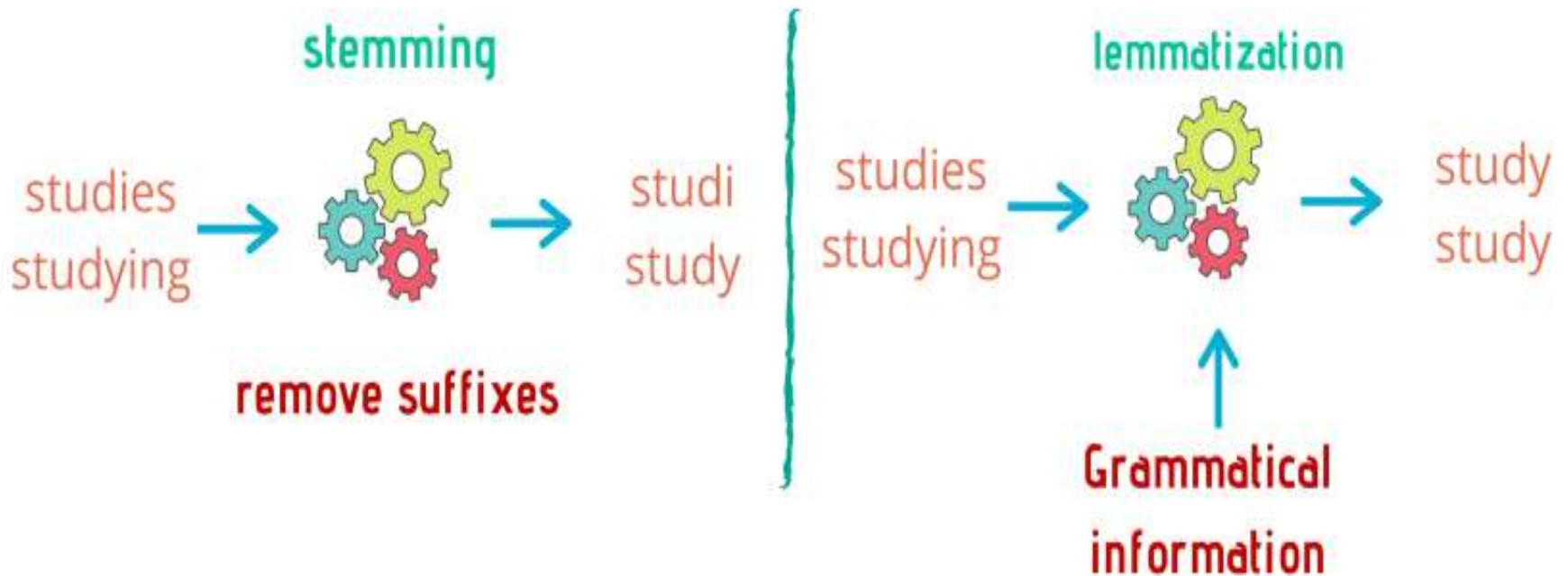
- Stemming is a technique used to extract the base form of the words by removing affixes from them. For example, the stem of the words **eating, eats, eaten** is **eat**.

- **Lemmatization:**

- Lemmatization considers the context and converts the word to its meaningful base form, which is called ***Lemma***. For instance, stemming the word '**Caring**' would return '**Car**'. For instance, lemmatizing the word '**Caring**' would return '**Care**'.

The steps to perform **preprocessing** of data :

STEMMING VS. LEMMATIZATION



The steps to perform **preprocessing** of data :

- **Filtering (or) Removing Stop Words:**

- It is a process of removing non-essential words,i.e Words such as **was, in, is, and, the,** are called stop words and can be removed.

The star are twinkling at night

stars

twinkling

night



How Does Text Mining Work?

- The **third step** is **Feature Extraction** from pre processed data.
- The mapping from textual data to real-valued vectors is called **feature extraction**.
- One of the commonly used technique to extract the features from textual data is **calculating the frequency of words/tokens** in the document/corpus.

Bag of Words (BOW)

- One of the simplest technique is **Bag of Words (BOW)** to represent the text in numerical format(**Vectors**).
- In **BOW**, we make a list of unique words in the text corpus called vocabulary. Then we can represent each **sentence** or **document** as a **vector**, with each word represented as **1** for **presence** and **0** for **absence**.
- The **bag-of-words** (BOW) model is a representation that turns arbitrary text into **fixed-length vectors** by counting how many times each word appears. This process is often referred to as **vectorization**.

Bag of Words (BOW)

- Let's understand this with an example. Suppose we wanted to vectorize the following:

Document 1 : **the cat sat**

Document 2: **the cat sat in the hat**

Document 3: **the cat with the hat**

Step 1: Determine the Vocabulary

- We first define our **vocabulary**, which is the set of all unique words found in our document set.
- The Words are :

the, cat, sat, in, hat, with

Bag of Words (BOW)

Step 2: Count

Document	the	cat	sat	in	hat	with
<i>the cat sat</i>	1	1	1	0	0	0
<i>the cat sat in the hat</i>	2	1	1	1	1	0
<i>the cat with the hat</i>	2	1	0	0	1	1

Bag of Words (BOW)

Step 3: Vector Representation

Now we have length-6 vectors for each document.

- *the cat sat*: [1, 1, 1, 0, 0, 0]
 - *the cat sat in the hat*: [2, 1, 1, 1, 1, 0]
 - *the cat with the hat*: [2, 1, 0, 0, 1, 1]
- Notice that we lose **contextual information**, e.g. where in the document the word appeared, when we use **BOW**. It's like a literal **bag**-of-words: it only tells you *what* words occur in the document, not *where* they occurred.



- Term frequency — Inverse document frequency (**TFIDF**)
- Word2Vec (**W2V**)

Text mining process

- Text preprocessing
 - Syntactic/Semantic text analysis
- Features Generation
 - Bag of words
- Features Selection
 - Simple counting
 - Statistics
- Text/Data Mining
 - Classification-Supervised learning
 - Clustering-Unsupervised learning
- Analyzing results

