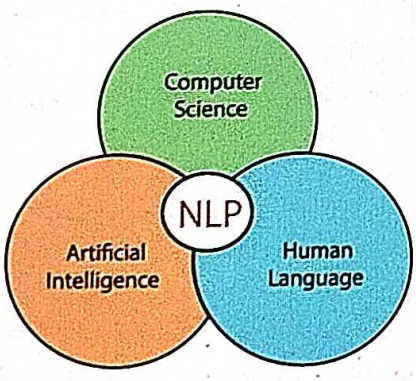
1a Natural Language Processing / Text mining:

NLP stands for **Natural Language Processing**, which is a part of **Computer Science, Human language,** and **Artificial Intelligence.** It is the technology that is used by machines to understand, analyse, manipulate, and interpret human's languages. It helps developers to organize knowledge for performing tasks such as **translation, automatic summarization, Named Entity Recognition (NER), speech recognition, relationship extraction,** and **topic segmentation.**



Advantages of NLP

- NLP helps users to ask questions about any subject and get a direct response within seconds.
- NLP offers exact answers to the question means it does not offer unnecessary and unwanted information.
- NLP helps computers to communicate with humans in their languages.
- It is very time efficient.
- Most of the companies use NLP to improve the efficiency of documentation processes, accuracy of documentation, and identify the information from large databases.

Disadvantages of NLP

A list of disadvantages of NLP is given below:

- NLP may not show context.
- NLP is unpredictable
- NLP may require more keystrokes.
- NLP is unable to adapt to the new domain, and it has a limited function that's why NLP is built for a single and specific task only.

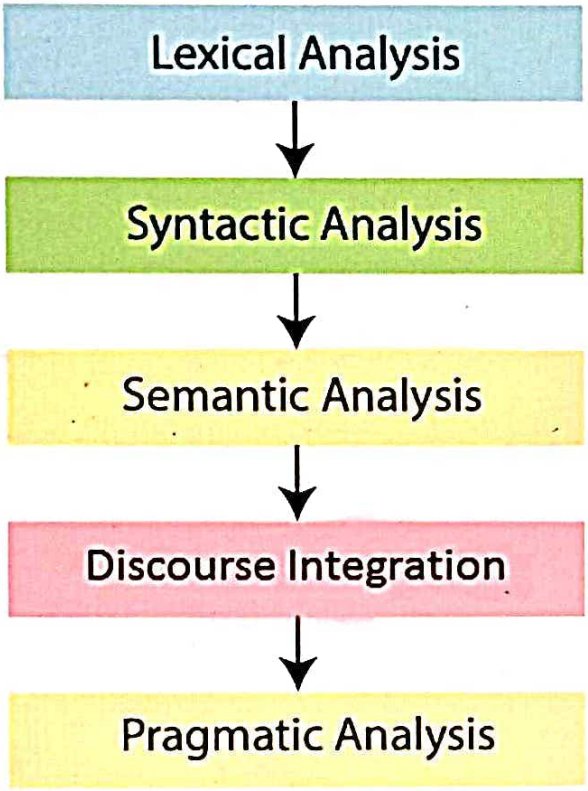
Common Natural Language Processing (NLP) Task:

- **Text and speech processing:** This includes Speech recognition, text-&-speech processing, encoding(i.e converting speech or text to machine-readable language), etc.
- **Text classification:** This includes Sentiment Analysis in which the machine can analyze the qualities, emotions, and sarcasm from text and also classify it accordingly.
- **Language generation:** This includes tasks such as machine translation, summary writing, essay writing, etc. which aim to produce coherent and fluent text.
- **Language interaction:** This includes tasks such as dialogue systems, voice assistants, and chatbots, which aim to enable natural communication between humans and computers.

NLP techniques are widely used in a variety of applications such as search engines, machine translation, sentiment analysis, text summarization, question answering, and many more. NLP research is an active field and recent advancements in deep learning have led to significant improvements in NLP performance. However, NLP is still a challenging field as it requires an understanding of both computational and linguistic principles.

1b Phases of NLP

There are the following five phases of NLP:



1. Lexical Analysis and Morphological

The first phase of NLP is the Lexical Analysis. This phase scans the source code as a stream of characters and converts it into meaningful lexemes. It divides the whole text into paragraphs, sentences, and words.

2. Syntactic Analysis (Parsing)

Syntactic Analysis is used to check grammar, word arrangements, and shows the relationship among the words.

Example: Agra goes to the Poonam

In the real world, Agra goes to the Poonam, does not make any sense, so this sentence is rejected by the Syntactic analyzer.

3. Semantic Analysis

Semantic analysis is concerned with the meaning representation. It mainly focuses on the literal meaning of words, phrases, and sentences.

4. Discourse Integration

Discourse Integration depends upon the sentences that proceeds it and also invokes the meaning of the sentences that follow it.

5. Pragmatic Analysis

Pragmatic is the fifth and last phase of NLP. It helps you to discover the intended effect by applying a set of rules that characterize cooperative dialogues.

For Example: "Open the door" is interpreted as a request instead of an order.

2 Word2Vec is a popular natural language processing (NLP) technique that's used to convert words or phrases into numerical vectors (mathematical representations) in a way that captures semantic meaning. Developed by Tomas Mikolov and his team at Google in 2013, Word2Vec has had a significant impact on various NLP tasks, such as text classification, sentiment analysis, and machine translation. The primary idea behind Word2Vec is to learn vector representations of words in a way that similar words have similar vector representations.

There are two main architectures for Word2Vec: Continuous Bag of Words (CBOW) and Skip-Gram. Let's explore both in detail:

1. **Continuous Bag of Words (CBOW):**
   - In CBOW, the model tries to predict a target word based on the context of surrounding words. For instance, if you have the sentence "I am learning to ___ deep learning," CBOW will try to predict the missing word "understand" using the context provided by "I am learning to" and "deep learning."
   - The input to the model is a fixed-sized context window around the target word, typically a few words on either side.
   - CBOW aims to learn the word vectors that minimize the difference between the predicted word and the actual word. It optimizes the likelihood of the target word given the context.
2. **Skip-Gram:**
   - Skip-Gram, on the other hand, reverses the task. It takes a target word and tries to predict the context words around it.
   - Given the word "learning" as input, Skip-Gram will try to predict the context words "I," "am," "to," and "deep learning."
   - Skip-Gram also optimizes the likelihood of the context words given the target word.

The training process for both CBOW and Skip-Gram typically involves a neural network with a hidden layer. This network learns to map words to continuous vector representations (word embeddings). These word embeddings capture semantic relationships between words. In other words, words with similar meanings will have similar vector representations.

Here's a simplified overview of the Word2Vec training process:

1. **Data Preparation**: You need a large corpus of text data. The larger, the better, as it allows for more accurate word representations.
2. **Data Preprocessing**: Tokenize the text, remove punctuation, and transform words to lowercase to create a clean text corpus.
3. **Building the Model**: Create a neural network with one hidden layer. The input and output layer sizes depend on the vocabulary size. The hidden layer size is typically chosen to be the size of the word vectors (e.g., 100, 300 dimensions).
4. **Training**: Train the model using your text data. During training, the weights connecting the input and hidden layer represent the word vectors.
5. **Word Vectors**: Once training is complete, you can extract the weights of the hidden layer, which represent the word vectors.

Key benefits of Word2Vec:

- **Semantic Meaning**: Words with similar meanings are close in vector space, allowing for semantic relationships to be captured.
- **Efficiency**: Word2Vec produces relatively compact word vectors, which are computationally efficient for NLP tasks.
- **Generalization**: Pre-trained Word2Vec models can be used for downstream tasks without the need for extensive domain-specific training data.

In practice, pre-trained Word2Vec models are often used in various NLP applications to improve the performance of models on specific tasks by leveraging the learned semantic relationships between words.

3

Bag of Words:

→ The NLP is used for text modeling. The text modeling also known as bag of words model.

→ NLP algorithms works in the modeling based on the numbers.

→ Bag of words model is used to Preprocess the text by converting it into a bag of words, which keep a count of the total occurances of most frequently used words.

Steps to create Bag of words:

Step-1:

a) Convert text to lower case.
b) Remove all non word characters.
c) Remove all puntuations.

Step-2:

a) Obtaining most frequent words in our text.
b) We declare a dictionary to hold our bag of words.
b) Next we tokenize each sentence to word.
c) Now for each word in sentence we check if the word exist in our dictionary or not.
d) If it does, then we increment its count by 1. If it doesn't we add it to our dictionary and set its count as 0.

Step 3: Building the bag of words model, in this step we construct a vector which would tell whether a word in each sentence is a frequent

word or not. If a word in a sentence is a frequent word we set it as 1. else we set it as 0.

Eg:- St1 : The cat sat on the mat. [100 000 000]

St2 : The dog chased the cat. [100000001]

St3 : The mat was soft and fluffy. [1100000]

4 Inverse document frequency (IDF):

→ IDF of a term reflects the proportion of documents in the corpus (collection of language models) that contain the term, words unique to small percentage of documents receive higher imp values than words common acerоss all documents.

$$IDF = \log\left(\frac{\text{No of docs in corpus}}{\text{No of docs in corpus contains the term}}\right)$$

→ Score is called as TF-IDF score.

TF-IDF score = TF * IDF.

This score is used to measure doc frequency.

Application:

1) Used in search engines (to rank the relevance of a document)
2) Text classification, summarization and topic modeling.

2) A collection of related doc contain 10,000 doc. If 100 doc out of 10,000 doc contain the term t. Calculate the value of IDF of t.

Sol:-

$$IDF = \log_{10}\left(\frac{10,000}{100}\right)$$

$$= \log_{10} 100$$

$$= 2\log_{10} 10$$

$$= 2$$

Score of TF-IDF of term t is = 0.2×2

$= 0.4$

5 Term frequency : It is a widely used statistical model or method in NLP and information retrival

Def: TF of a term or word is the no of time the term appears in a document compared to the total no of words in the document.

$$TF = \frac{\text{no of times the term appears in the doc}}{\text{Total no of terms in the document.}}$$

→ The words within a text are transformed into important numbers by a text vectorization Process.

Eg:- 1) Imagine the term t appears 20 times in a doc that contains a total of 100 words. What is the value of TF of t.

Sol:-

$$TF = \frac{\text{Total no of time t appears}}{\text{Total no of terms}}$$

$$TF = \frac{20}{100} = \frac{1}{5} = 0.2$$

6 One hot encoding

-> It is the process of turning catagorical features into a numerical structure that ML algorithms can readily process. Each catagory in a feature is a binary vector of 1's and 0's.

→ This encoding is used in NLP to encode categorical factors as binary vectors such as words or PoS identifiers.

→ This method transforms the input data text to numerical values.

1. It allows the use of categorical variables in models that require numerical input.
2. It can improve model performance by providing more information to the model about the categorical variable.
3. It can help to avoid the problem of ordinality, which can occur when a categorical variable has a natural ordering (e.g. "small", "medium", "large").

The disadvantages of using one hot encoding include:

1. It can lead to increased dimensionality, as a separate column is created for each category in the variable. This can make the model more complex and slow to train.

2. It can lead to sparse data, as most observations will have a value of 0 in most of the one-hot encoded columns.

3. It can lead to overfitting, especially if there are many categories in the variable and the sample size is relatively small.

4. One-hot-encoding is a powerful technique to treat categorical data, but it can lead to increased dimensionality, sparsity, and overfitting. It is important to use it cautiously and consider other methods such as ordinal encoding or binary encoding.

One Hot Encoding Examples

In **One Hot Encoding**, the categorical parameters will prepare separate columns for both Male and Female labels. So, wherever there is a Male, the value will be 1 in the Male column and 0 in the Female column, and vice-versa. Let's understand with an example: Consider the data where fruits, their corresponding categorical values, and prices are given.

Fruit	Categorical value of fruit	Price
apple	1	5
mango	2	10
apple	1	15
orange	3	20

The output after applying one-hot encoding on the data is given as follows,

apple	mango	orange	price
1	0	0	5
0	1	0	10
1	0	0	15
0	0	1	20

Why use a One Hot Encoding?

One of the best advantages of One Hot encoding is that it represents categorical data to be more expressive. As we discussed earlier, many machine learning algorithms cannot be able to work with the categorical data directly, so that it needs to be converted into integer.

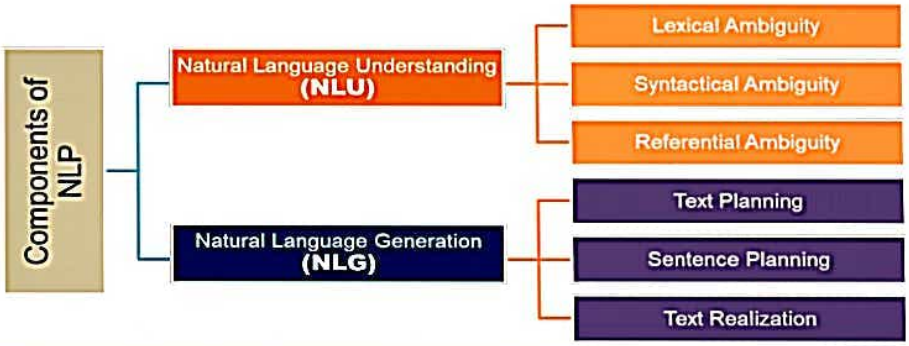
We can use the integer value directly or where it is needed. It can solve the problem where the natural ordinal has a relationship between the categories. For example - We can assign the integer values to **"weather"** label, such as 'winter', 'summer' and 'Monsoon'.

But there may be problems if no ordinal relationship find. If we allow the representation to lean or any such relationship, it might be damaged the learning to solve problems.

7 NLP can be divided into two basic components.

Natural Language Understanding(NLU)

Natural Language Generation(NLG)



Natural Language Processing (NLP)

Components of NLP

Natural language understanding (NLU)

Natural Language Understanding (NLU) helps the machine to understand and analyse human language by extracting the metadata from content such as concepts, entities, keywords, emotion, relations, and semantic roles.

There are lot of ambiguity while learning or trying to interpret a language.

He is looking for a match

What do you understand by 'match'?
Partner
Or Cricket/Football Match

Natural Language Processing (NLP)

Components of NLP

Natural language understanding (NLU)

if you know what is ambiguity (different meaning of any particular thing) then this term has a direct relation to this word.

Lexical (word level) – Lexical work at the word level, imagine any word that is used as a verb and also used as a noun. These are crucial to deciding for NLP

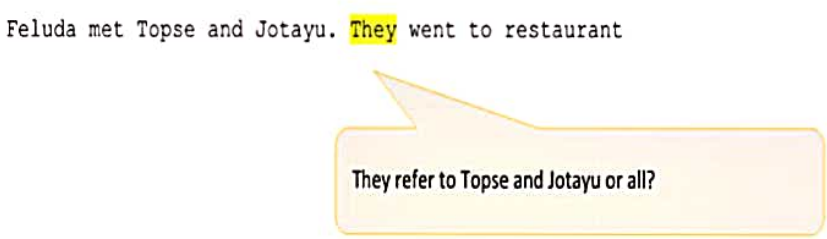
The chicken is ready to eat.

Is the chicken ready to eat his food or the chicken is ready for someone else to it? You never know.

Components of NLP

Natural language understanding (NLU)

Syntactical Ambiguity means when we see more than one meaning in a sequence of words. It is also termed as grammatical ambiguity.



Components of NLP

Natural language understanding (NLU)

Referential – Let see a new scenario to understand this better. **"Alex went to Dave; he said that he was hungry".**

This is just an explanation statement to demonstrate how complex the interpretations can be for the computers to understand in their initial NLP phase.

So, in the above statement the confusion for a computer to understand two he's is meant for which person **(means Alex or Dave).**

Components of NLP

Natural language Generation: So the machine has understood that we asked them to do something, now come to their turn to provide a proper **response or feedback. NLG does the same thing.**

Text Planning – This means to plain text from the knowledge base, just like we humans have a vocabulary which helps us to frame sentences.

Sentence making – To arrange all the words and make an arrangement in a meaningful pattern.

Text Realization – To process all the sentences in a proper sequence or order and give the output is called text realization

8 Applications of Natural Language Processing

1. Chatbots

Chatbots are a form of artificial intelligence that are programmed to interact with humans in such a way that they sound like humans themselves. Depending on the complexity of the chatbots, they can either just respond to specific keywords or they can even hold full conversations that make it tough to distinguish them from humans. Chatbots are created using Natural Language Processing and Machine Learning, which means that they understand the complexities of the English language and find the actual meaning of the sentence and they also learn from their conversations with humans and become better with time. Chatbots work in two simple steps. First, they identify the meaning of the question asked and collect all the data from the user that may be required to answer the question. Then they answer the question appropriately.

2. Autocomplete in Search Engines

Have you noticed that search engines tend to guess what you are typing and automatically complete your sentences? For example, On typing "game" in Google, you may get further suggestions for "game of thrones", "game of life" or if you are interested in maths then "game theory". All these suggestions are provided using autocomplete that uses Natural Language Processing to guess what you want to ask. Search engines use their enormous data sets to analyze what their customers are probably typing when they enter particular words and suggest the most common possibilities. They use Natural Language Processing to make sense of these words and how they are interconnected to form different sentences.

3. Voice Assistants

These days voice assistants are all the rage! Whether its Siri, Alexa, or Google Assistant, almost everyone uses one of these to make calls, place reminders, schedule meetings, set alarms, surf the internet, etc. These voice assistants have made life much easier. But how do they work? They use a complex combination of speech recognition, natural language understanding, and natural language processing to understand what humans are saying and then act on it. The long term goal of voice assistants is to become a bridge between humans and the internet and provide all manner of services based on just voice interaction. However, they are still a little far from that goal seeing as Siri still can't understand what you are saying sometimes!

4. Language Translator

Want to translate a text from English to Hindi but don't know Hindi? Well, Google Translate is the tool for you! While it's not exactly 100% accurate, it is still a great tool to convert text from one language to another. Google Translate and other translation tools as well as use Sequence to sequence modeling that is a technique in Natural Language Processing. It allows the algorithm to convert a sequence of words from one language to another which is translation. Earlier, language translators used Statistical machine translation (SMT) which meant they analyzed millions of documents that were already translated from one language to another (English to Hindi in this case) and then looked for the common patterns and basic vocabulary of the language. However, this method was not that accurate as compared to Sequence to sequence modeling.

5. Sentiment Analysis

Almost all the world is on social media these days! And companies can use sentiment analysis to understand how a particular type of user feels about a particular topic, product, etc. They can use natural language processing, computational linguistics, text analysis, etc. to understand the general sentiment of the users for their products and services and find out if the sentiment is good, bad, or neutral. Companies can use sentiment analysis in a lot of ways such as to find out the emotions of their target audience, to understand product reviews, to gauge their brand sentiment, etc. And not just private companies, even governments use sentiment analysis to find popular opinion and also catch out any threats to the security of the nation.

6. Grammar Checkers

Grammar and spelling is a very important factor while writing professional reports for your superiors even assignments for your lecturers. After all, having major errors may get you fired or failed! That's why grammar and spell checkers are a very important tool for any professional writer. They can not only correct grammar and check spellings but also suggest better synonyms and improve the overall readability of your content. And guess what, they utilize natural language processing to provide the best possible piece of writing! The NLP algorithm is trained on millions of sentences to understand the correct format. That is why it can suggest the correct verb tense, a better synonym, or a clearer sentence structure than what you have written. Some of the most popular grammar checkers that use NLP include Grammarly, WhiteSmoke, ProWritingAid, etc.

7. Email Classification and Filtering

Emails are still the most important method for professional communication. However, all of us still get thousands of promotional Emails that we don't want to read. Thankfully, our emails are automatically divided into 3 sections namely, Primary, Social, and Promotions which means we never have to open the Promotional section! But how does this work? Email services use natural language processing to identify the contents of each Email with text classification so that it can be put in the correct section. This method is not perfect since there are still some Promotional newsletters in Primary, but its better than nothing. In more advanced cases, some companies also use specialty anti-virus software with natural language processing to scan the Emails and see if there are any patterns and phrases that may indicate a phishing attempt on the employees.

Steps to get text data into workable format

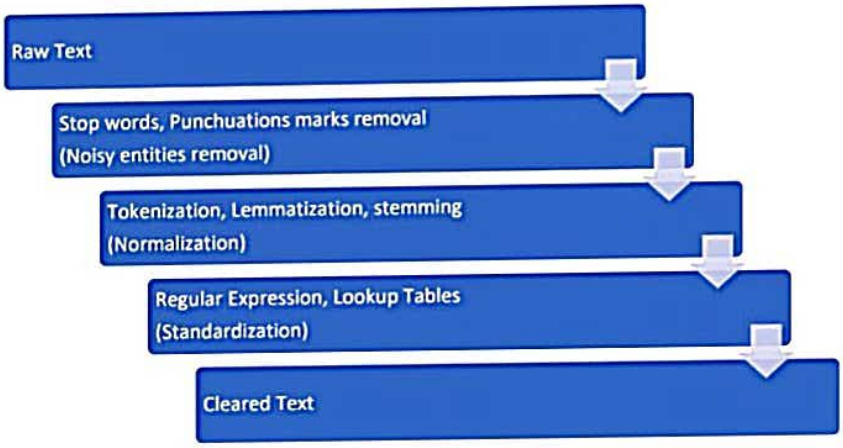


Fig: Approximate order of steps for preprocessing text data.

Steps to get text data into workable format

Tokenization: Converting the sentences in text into individual units (or) words is called tokenization.

For example, Consider a text:

"Vishnu Institute of Technology is a good college. I am studying there".

After applying tokenization,

"Vishnu", "institute", "of", "Technology", "is"," a"," good",
"college","." "I", "am", "Studying"," there","."

Steps to get text data into workable format

Noisy Entities removal:

Any piece of text which is **not relevant to the context** of the data can be specified as the noise.

Generally, the noisy entities are **Stop words, punctuation marks.**

1) Stop Words removal

It is a process of removing common language articles, Pronouns and propositions such as "and", "the" or "to" in English.

These words provide little or no value to NLP.

Stop words are removed /filtered from text for better efficiency.

Steps to get text data into workable format

Some common stop words in English are:

i	himself	whom	having	of	from	where	so
me	she	this	do	at	up	why	than
my	her	that	does	by	down	how	too
myself	hers	these	did	for	in	all	very
we	herself	those	doing	with	out	any	s
our	it	am	a	about	on	both	t
ours	its	is	an	against	off	each	can
ourselves	itself	are	the	between	over	few	will
you	they	was	and	into	under	more	just
your	them	were	but	through	again	most	should
yours	their	be	if	during	further	other	now
yourself	theirs	been	or	before	then	some	
yourselves	themselves	being	because	after	once	such	
he	what	have	as	above	here	only	
him	which	has	until	below	there	own	
his	who	had	while	to	when	same	

After removing stop words, our output look like this.

"Vishnu", "institute", "Technology"," good", "college", "Studying"

Note that the above list changes from time to time.

Natural Language Processing (NLP)

Steps to get text data into workable format

2) Normalization

Stemming: It is a process of transforming a word to its root form.

Stemming is a rude rule-based process of stripping the suffixes **("ing", "ly", "es", "s" etc)** from a word.

Natural Language Processing (NLP)

Steps to get text data into workable format

After Appling stemming the output looks like this.

"Vishnu", "institute", "Technology"," good", "college", "Study"

Stemming fails some times.

The **big problem** with stemming is that sometimes it produces the root word which may not have any meaning.

For Example, intelligence, intelligent, and intelligently, all these words are originated with a single root word **"intelligen."** In English, the word **"intelligen" do not have any meaning.**

Steps to get text data into workable format

Lemmatization:

1. Lemmatization is quite similar to the Stamming.
2. It is used to group different inflected forms of the word, called **Lemma.**
3. The **main difference** between Stemming and lemmatization is that it produces the root word, which has a meaning.

For example: In lemmatization, the words intelligence, intelligent, and intelligently has a root word intelligent, which has a meaning.

As we can see above, stemming fails sometimes.

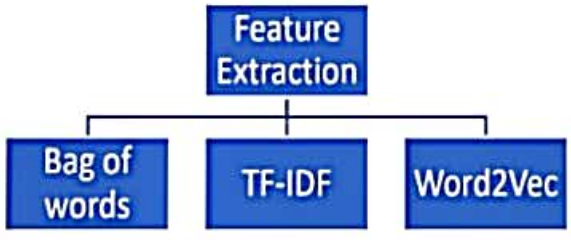
Steps to get text data into workable format

3) Object standardization:

Machine learning algorithms **cannot work with the raw text directly;** the text must be converted into numbers. Specifically, vectors of numbers.

Popular and simple method of feature extraction **(Feature engineering of text data)** with text data which are currently used are:

Popular Feature Extraction methods:



Steps to get text data into workable format

Bag of Words (BoW) Model

1. The Bag of Words (BoW) model is the simplest form of **text representation in numbers.**
2. Like the term itself, we can **represent a sentence as a bag of words vector** (a string of numbers).

Bag of words is a two-step process.

1) A vocabulary of known words:

2. A measure of the presence of known words:

Steps to get text data into workable format

Bag of Words (BoW) Model

1) A vocabulary of known words:

In this step, all unique words in the document/text is collected.

Here's a sample of reviews about a particular horror movie:

1. Review 1: This movie is very scary and long
2. Review 2: This movie is not scary and is slow
3. Review 3: This movie is spooky and good

Steps to get text data into workable format

1) A vocabulary of known words:

We will first build a vocabulary from all the unique words in the above three reviews.

The vocabulary consists of these 11 words: **'This', 'movie', 'is', 'very', 'scary', 'and', 'long', 'not', 'slow', 'spooky', 'good'.**

Steps to get text data into workable format

1) A vocabulary of known words:

We can now take each of these words and mark their occurrence in the three movie reviews above **with 1s and 0s.**

This will give us 3 vectors for 3 reviews:

	1 This	2 movie	3 is	4 very	5 scary	6 and	7 long	8 not	9 slow	10 spooky	11 good	Length of the review(in words)
Review 1	1	1	1	1	1	1	1	0	0	0	0	7
Review 2	1	1	2	0	0	1	1	0	1	0	0	8
Review 3	1	1	1	0	0	0	1	0	0	1	1	6

Vector of Review 1: [1 1 1 1 1 1 1 0 0 0 0]

Vector of Review 2: [1 1 2 0 0 1 1 0 1 0 0]

Vector of Review 3: [1 1 1 0 0 0 1 0 0 1 1]

10 **Notion of corpus:**

In machine learning, the notion of a "corpus" is similar to its use in linguistics and natural language processing (NLP), but it is broader and can refer to structured collections of data used for training, evaluation, and testing machine learning models. A corpus in the context of machine learning can consist of various types of data, not limited to text or speech, and serves as a fundamental resource for model development. Here are some key aspects of the notion of a corpus in machine learning:

1. **Data Collection**: A corpus in machine learning is typically a dataset collected for a specific task or problem. This dataset can include various types of data, such as:
    - **Text Data**: Text corpora used for tasks like text classification, sentiment analysis, and natural language understanding.
    - **Image Data**: Image corpora for tasks like image classification, object detection, and computer vision.
    - **Audio Data**: Audio corpora for tasks like speech recognition, music genre classification, and sound event detection.
    - **Tabular Data**: Structured corpora that contain data in tabular form, commonly used in machine learning for tasks like regression and classification.
    - **Time Series Data**: Temporal data used in tasks like time series forecasting and anomaly detection.
    - **Biological Data**: Genomic, proteomic, or other biological data for tasks like sequence analysis and bioinformatics.
    - **Sensor Data**: Data collected from sensors, IoT devices, and other sources for various machine learning applications.
2. **Size and Representativeness**: The size and representativeness of a machine learning corpus depend on the specific problem it aims to address. It should be large enough to capture meaningful patterns and variations relevant to the task.

3. **Annotation**: Depending on the task, a corpus may be annotated with labels, ground truth data, or other metadata to facilitate supervised learning. Annotation can be especially important in tasks like object detection, named entity recognition, and sentiment analysis.
4. **Quality**: Data quality is crucial in machine learning. A well-constructed corpus should be free from errors and biases that can impact model performance. Data cleaning and preprocessing are often required to ensure data quality.
5. **Splitting**: The corpus is typically divided into training, validation, and test sets for machine learning model development. This allows for model training, hyperparameter tuning, and performance evaluation.
6. **Open and Proprietary Corpora**: Just as in NLP, some machine learning corpora are publicly available, while others are proprietary and may require licensing. The choice of corpus depends on the problem and the resources available.
7. **Ethical and Privacy Considerations**: When working with corpora, especially those containing sensitive or personally identifiable information, ethical and privacy considerations are essential. Compliance with data protection laws and ethical guidelines is crucial.

Machine learning corpora serve as the foundation for training and testing machine learning models across various domains and applications, including image recognition, speech understanding, recommendation systems, autonomous vehicles, and more. They are instrumental in building and fine-tuning models to make predictions, classify data, and solve complex problems.