1.1 Introduction to NLP

What are the four types of NLP?

The four types of Natural Language Processing (NLP) are:

- Natural Language Understanding (NLU)
- Natural Language Generation (NLG)
- Natural Language Processing (NLP) itself, which encompasses both NLU and NLG.
- Natural Language Interaction (NLI)

- **What are different NLP techniques?**
- **NLP techniques and methods**

Here are some fundamental techniques used in NLP: Tokenization. This is the process of breaking text into words, phrases, symbols, or other meaningful elements, known as tokens. Parsing. Parsing involves analyzing the grammatical structure of a sentence to extract meaning.

Why use NLP techniques?

Natural language processing (NLP) is critical to fully and efficiently analyze text and speech data. It can work through the differences in dialect, slang, and grammatical irregularities typical in day-to-day conversations.

What is natural language processing?

Natural language processing is a branch of artificial intelligence that aims to help computers to understand human language input in the form of text or speech.

NLP combines multiple disciplines, including computation linguistics, machine learning, deep learning, and statistics.

These technologies work together to essentially give computer software the ability to process and understand human language in the way that another human could, including its meaning, intent, and sentiment.

NLP technology is used in a variety of applications including:

- Digital assistants such as Siri.
- Speech-to-text dictation software.
- Voice-operated GPS systems.
- Customer service chatbots.
- Predictive text.
- Digital voicemail.
- Autocorrect.
- Search autocomplete.

- Email filters.

Additionally, companies are increasingly using NLP to create enterprise solutions that help businesses simplify processes, increase productivity, and streamline operations.

The benefits of employing natural language processing

It's standard these days for companies to collect, store, process, and analyze large quantities of numerical data in order to generate valuable insights that can improve results.

Natural language processing opens up and empowers businesses to make smarter decisions that are based on larger sets of data. Further, this collection and analysis process happens quickly, especially compared to traditional methods.

For this reason, natural language processing has a number of relevant advantages.

When working with so much data, you'll be able to generate insights to improve customer experience with the launch of new products.

On top of that, using NLP helps businesses become more efficient by automating work processes that require reviewing or analyzing texts. This frees up employees to work on other needle-moving tasks.

Taken together, you're bound to see improved productivity, reduced costs, and an uplift in revenue.

Now that we've learned about how natural language processing works, it's important to understand what it can do for businesses.

Enhanced Data Analysis

While NLP and other forms of AI aren't perfect, natural language processing can bring objectivity to data analysis, providing more accurate and consistent results.

Faster Insights

With the Internet of Things and other advanced technologies compiling more data than ever, some data sets are simply too overwhelming for humans to comb through. Natural language processing can quickly process massive volumes of data, gleaning insights that may have taken weeks or even months for humans to extract.

Increased Employee Productivity

NLP handles mundane tasks like sifting through data sets, sorting emails and assessing customer responses. With these repetitive responsibilities out of the way, workers are freed up to focus on more complex and pressing matters.

Higher-Quality Customer Experience

In the form of chatbots, natural language processing can take some of the weight off customer service teams, promptly responding to online queries and redirecting customers when needed. NLP can also analyze customer surveys and feedback, allowing teams to gather timely intel on how customers feel about a brand and steps they can take to improve customer sentiment.

NLP Use Cases

Keeping the advantages of natural language processing in mind, let's explore **how different industries are applying this technology.**

Customer Service

While NLP-powered chatbots and callbots are most common in customer service contexts, companies have also relied on natural language processing to power virtual assistants. These assistants are a form of conversational AI that can carry on more sophisticated discussions. And if NLP is unable to resolve an issue, it can connect a customer with the appropriate personnel.

Marketing

Gathering market intelligence becomes much easier with natural language processing, which can analyze online reviews, social media posts and web forums. Compiling this data can help marketing teams understand what consumers care about and how they perceive a business' brand.

Human Resources

Recruiters and HR personnel can use natural language processing to sift through hundreds of resumes, picking out promising candidates based on keywords, education, skills and other criteria. In addition, NLP's data analysis capabilities are ideal for reviewing employee surveys and quickly determining how employees feel about the workplace.

E-Commerce

Natural language processing can help customers book tickets, track orders and even recommend similar products on e-commerce websites. Teams can also use data on customer purchases to inform what types of products to stock up on and when to replenish inventories.

Finance

In finance, NLP can be paired with machine learning to generate financial reports based on invoices, statements and other documents. Financial analysts can also employ natural language processing to predict stock market trends by analyzing news articles, social media posts and other online sources for market sentiments.

Insurance

Insurance companies can assess claims with natural language processing since this technology can handle both structured and unstructured data. NLP can also be trained to pick out unusual information, allowing teams to spot fraudulent claims.

Education

NLP-powered apps can check for spelling errors, highlight unnecessary or misapplied grammar and even suggest simpler ways to organize sentences. Natural language processing can also translate text into other languages, aiding students in learning a new language.

Healthcare

Healthcare professionals can develop more efficient workflows with the help of natural language processing. During procedures, doctors can dictate their actions and notes to an app, which produces an accurate transcription. NLP can also scan patient documents to identify patients who would be best suited for certain clinical trials.

Manufacturing

With its ability to process large amounts of data, NLP can inform manufacturers on how to improve production workflows, when to perform machine maintenance and what issues need to be fixed in products. And if companies need to find the best price for specific materials, natural language processing can review various websites and locate the optimal price.

Cybersecurity

IT and security teams can deploy natural language processing to filter out suspicious emails based on word choice, sentiment and other factors. This makes it easier to protect different departments from spam, phishing scams and other cyber attacks. With its ability to understand data, NLP can also detect unusual behavior and alert teams of possible threats.

the core components, techniques, applications, and challenges of NLP.

Natural Language Processing (NLP) stands as a pivotal technology in the realm of artificial intelligence, bridging the gap between human communication and computer understanding. It is a multidisciplinary domain that empowers computers to interpret, analyze, and generate human language, enabling seamless interaction between humans and machines. The significance of NLP is evident in its widespread applications, ranging from automated customer support to real-time language translation.

This article aims to provide newcomers with a comprehensive overview of NLP, its workings, applications, challenges, and future outlook.

What is Natural Language Processing?

Natural Language Processing (NLP) is a branch of **artificial intelligence** that focuses on the interaction between computers and humans through natural language. The objective is to program computers to process and analyze large amounts of natural language data.

NLP involves enabling machines to understand, interpret, and produce human language in a way that is both valuable and meaningful. OpenAI, known for developing advanced language models like **ChatGPT**, highlights the importance of NLP in creating intelligent systems that can understand, respond to, and generate text, making technology more user-friendly and accessible.

How Does NLP Work?

Components of NLP

Natural Language Processing is not a monolithic, singular approach, but rather, it is composed of several components, each contributing to the overall understanding of language. The main components that NLP strives to understand are Syntax, Semantics, Pragmatics, and Discourse.

Syntax

- Definition: Syntax pertains to the arrangement of words and phrases to create well-structured sentences in a language.
- Example: Consider the sentence "The cat sat on the mat." Syntax involves analyzing the grammatical structure of this sentence, ensuring that it adheres to the grammatical rules of English, such as subject-verb agreement and proper word order

Semantics

- Definition: Semantics is concerned with understanding the meaning of words and how they create meaning when combined in sentences.
- Example: In the sentence "The panda eats shoots and leaves," semantics helps distinguish whether the panda eats plants (shoots and leaves) or is involved in a violent act (shoots) and then departs (leaves), based on the meaning of the words and the context.

Pragmatics

- Definition: Pragmatics deals with understanding language in various contexts, ensuring that the intended meaning is derived based on the situation, speaker's intent, and shared knowledge.
- Example: If someone says, "Can you pass the salt?" Pragmatics involves understanding that this is a request rather than a question about one's ability to pass the salt, interpreting the speaker's intent based on the dining context.

Discourse

- Definition: Discourse focuses on the analysis and interpretation of language beyond the sentence level, considering how sentences relate to each other in texts and conversations.
- Example: In a conversation where one person says, "I'm freezing," and another responds, "I'll close the window," discourse involves understanding the coherence between the two statements, recognizing that the second statement is a response to the implied request in the first.

Understanding these components is crucial for anyone delving into NLP, as they form the backbone of how NLP models interpret and generate human language.

NLP techniques and methods

To analyze and understand human language, NLP employs a variety of techniques and methods. Here are some fundamental techniques used in NLP:

- **Tokenization.** This is the process of breaking text into words, phrases, symbols, or other meaningful elements, known as tokens.
- **Parsing.** Parsing involves analyzing the grammatical structure of a sentence to extract meaning.
- **Lemmatization.** This technique reduces words to their base or root form, allowing for the grouping of different forms of the same word.
- **Named Entity Recognition (NER).** NER is used to identify entities such as persons, organizations, locations, and other named items in the text.
- **Sentiment analysis.** This method is used to gain an understanding of the sentiment or emotion conveyed in a piece of text.

Each of these techniques plays a vital role in enabling computers to process and understand human language, forming the building blocks of more advanced NLP applications.

What is NLP Used For?

Now that we have some of the basic concepts defined, let's take a look at how natural language processing is used in the modern world.

Industry applications

Natural Language Processing has found extensive applications across various industries, revolutionizing the way businesses operate and interact with users. Here are some of the key industry applications of NLP.

Healthcare

NLP assists in **transcribing and organizing clinical notes**, ensuring accurate and efficient documentation of patient information. For instance, a physician might dictate their notes, which NLP systems transcribe into text. Advanced NLP models can further categorize the information, identifying symptoms, diagnoses, and prescribed treatments, thereby streamlining the documentation process, minimizing manual data entry, and enhancing the accuracy of electronic health records.

Finance

Financial institutions leverage NLP to **perform sentiment analysis** on various text data like news articles, financial reports, and social media posts to

gauge market sentiment regarding specific stocks or the market in general. Algorithms analyze the frequency of positive or negative words, and through machine learning models, predict potential impacts on stock prices or market movements, aiding traders and investors in making informed decisions.

Customer Service

NLP-powered chatbots have revolutionized customer support by providing instant, 24/7 responses to customer inquiries. These chatbots understand customer queries through text or voice, interpret the underlying intent, and provide accurate responses or solutions. For instance, a customer might inquire about their order status, and the chatbot, integrating with the order management system, retrieves and delivers the real-time status, enhancing customer experience and reducing support workload.

E-Commerce

NLP significantly enhances on-site search functionality in e-commerce platforms by understanding and interpreting user queries, even if they are phrased in a conversational manner or contain typos. For example, if a user searches for “blu jeens,” NLP algorithms correct the typos and understand the intent, providing relevant results for “blue jeans,” thereby ensuring that users find what they are looking for, even with imprecise queries.

Legal

In the legal sector, NLP is utilized to automate document review processes, significantly reducing the manual effort involved in sifting through vast volumes of legal documents. For instance, during litigation, legal professionals need to review numerous documents to identify relevant information. NLP algorithms can scan through these documents, identify and highlight pertinent information, such as specific terms, dates, or clauses, thereby expediting the review process and ensuring that no critical information is overlooked.

Everyday applications

Beyond industry-specific applications, NLP is ingrained in our daily lives, making technology more accessible and user-friendly. Here are some everyday applications of NLP:

- **Search engines.** NLP is fundamental to the functioning of search engines, enabling them to understand user queries and provide relevant results.

- **Virtual assistants.** Siri, Alexa, and Google Assistant are examples of virtual assistants that use NLP to understand and respond to user commands.
- **Translation services.** Services like Google Translate employ NLP to provide real-time language translation, breaking down language barriers and fostering communication.
- **Email filtering.** NLP is used in email services to filter out spam and categorize emails, helping users manage their inboxes more effectively.
- **Social media monitoring.** NLP enables the analysis of social media content to gauge public opinion, track trends, and manage online reputation.

The applications of NLP are diverse and pervasive, impacting various industries and our daily interactions with technology. Understanding these applications provides a glimpse into the transformative potential of NLP in shaping the future of technology and human interaction.

Challenges and The Future of NLP

Although natural language processing is an incredibly useful tool, it's not without it flaws. Here, we look at some of the challenges we need to overcome, as well as what the future holds for NLP.

Overcoming NLP challenges

Natural Language Processing, despite its advancements, faces several challenges due to the inherent complexities and nuances of human language. Here are some of the challenges in NLP:

- **Ambiguity.** Human language is often ambiguous, with words having multiple meanings, making it challenging for NLP models to interpret the correct meaning in different contexts.
- **Context.** Understanding the context in which words are used is crucial for accurate interpretation, and it remains a significant challenge for NLP.
- **Sarcasm and irony.** Detecting sarcasm and irony is particularly challenging as it requires understanding the intended meaning, which may be opposite to the literal meaning.
- **Cultural nuances.** Language is deeply intertwined with culture, and understanding cultural nuance and idioms is essential for effective NLP.

Researchers and developers are continually working to overcome these challenges, employing advanced machine learning and deep learning

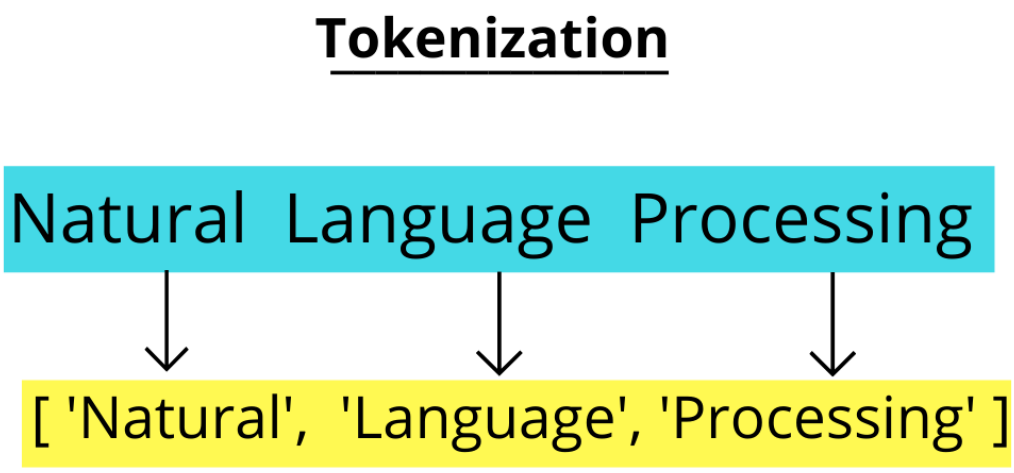
NLP Techniques

NLP is a rich field requiring the use of a number of different techniques in order to successfully process and understand human language. Below, we review and define a selection of the techniques commonly used in NLP technology.

Tokenization

Also called word segmentation, tokenization is one of the simplest and most important techniques involved in NLP.

It's a crucial preprocessing step in which a long string of text is broken down into smaller units called tokens. Tokens include words, characters, and sub words. They are the building blocks of natural language processing, and most NLP models process raw text on the token level.



An example from Medium of how a simple phrase can be broken down into tokens.

Stemming & lemmatization

After tokenization, the next preprocessing step is either stemming or lemmatization. These techniques generate the root word from the different existing variations of a word.

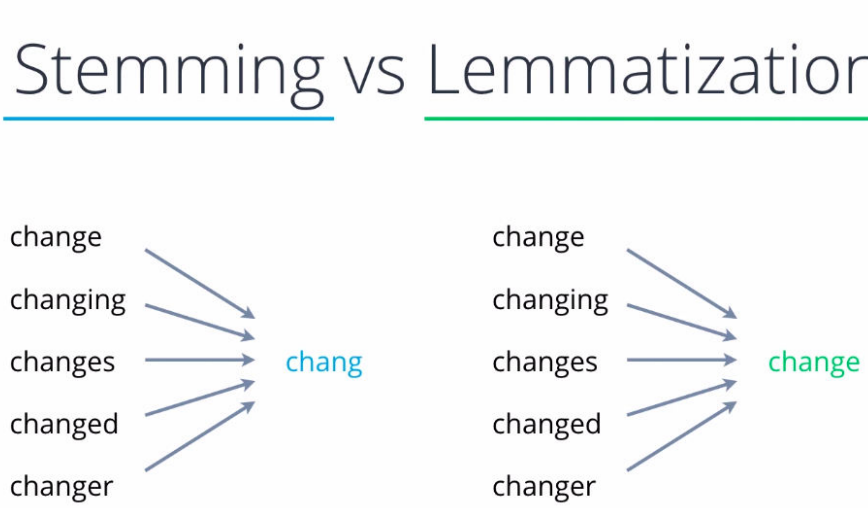
For example, the root word "stick" can be written in many different variations, like:

- Stick
- Stuck

- Sticker
- Sticking
- Sticks
- Unstick

Stemming and lemmatization are two different ways to try to identify a root word. Stemming works by removing the end of a word. This NLP technique may or may not work depending on the word. For example, it would work on “sticks,” but not “unstick” or “stuck.”

Lemmatization is a more sophisticated technique that uses morphological analysis to find the base form of a word, also called a lemma.



The difference between how stemming and lemmatization work is illustrated in this image from itnext, using different forms of the word “change.”

Morphological segmentation

Morphological segmentation is the process of splitting words into the morphemes that make them up. A morpheme is the smallest unit of language that carries meaning. Some words such as “table” and “lamp” only contain one morpheme.

But other words can contain multiple morphemes. For example, the word “sunrise” contains two morphemes: sun and rise. Like stemming and lemmatization, morphological segmentation can help preprocess input text.

John Hopkins shows morphological segmentation by breaking the word "unachievability" into its morphemes.

Stop words removal

Stop words removal is another preprocessing step of NLP that removes filler words to allow the AI to focus on words that hold meaning. This includes conjunctions such as "and" and "because," as well as prepositions such as "under" and "in."

By removing these unhelpful words, NLP systems are left with less data to process, allowing them to work more efficiently. It isn't a necessary step of every NLP use case, but it can help with things such as text classification.

Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

Examples from GeeksforGeeks *of what short phrases look like with the stop words removed.*

Text classification

Text classification is an umbrella term for any technique used to organize large quantities of raw text data. Sentiment analysis, topic modeling, and keyword extraction are all different types of text classification. And we'll talk about them shortly.

Text classification essentially takes unstructured text data and structures it, preparing it for further analysis. It can be used on nearly every text type and help with a number of different organization and categorization applications.

In this way, text classification is an essential part of natural language processing, used to help with everything from detecting spam to monitoring brand sentiment.

Some possible applications of text classification include:

- Grouping product reviews into categories based on sentiment.
- Flagging customer emails as more or less urgent.
- Organizing content by topic.

Sentiment analysis

Sentiment analysis, also known as emotion AI or opinion mining, is the process of analyzing text to determine whether it is generally positive, negative, or neutral.

As one of the most important NLP techniques for text classification, sentiment analysis is commonly used for applications such as analyzing user-generated content. It can be used on a variety of text types, including reviews, comments, tweets, and articles.

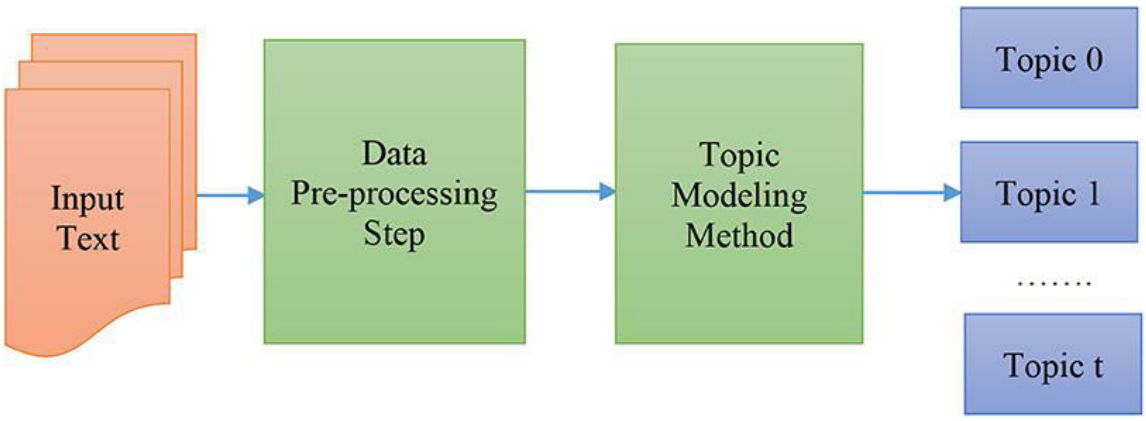
The Revuze platform employs sentiment analysis to understand how customers feel about various aspects of products. This allows companies to gain insights about consumers' needs in real-time, and act accordingly to improve overall CX.

In this example from the Revuze platform, you can see how customers rate different aspects of the product.

Topic modeling

Topic modeling is a technique that scans documents to find themes and patterns within them, clustering related expressions and word groupings as a way to tag the set.

It’s an unsupervised machine learning process, meaning that it doesn’t require the documents it is processing to have previously been categorized by humans.



A sample NLP workflow from ***Frontiersin*** *demonstrates how Input text is proprocessed before undergoing topic modeling, which breaks it into several topics.*

Keyword extraction

Keyword extraction is a technique that skims a document, ignoring the filler words and honing in on the important keywords. It is used to automatically extract the most frequently used and essential words and phrases from a document, helping to summarize it and identify what it’s about.

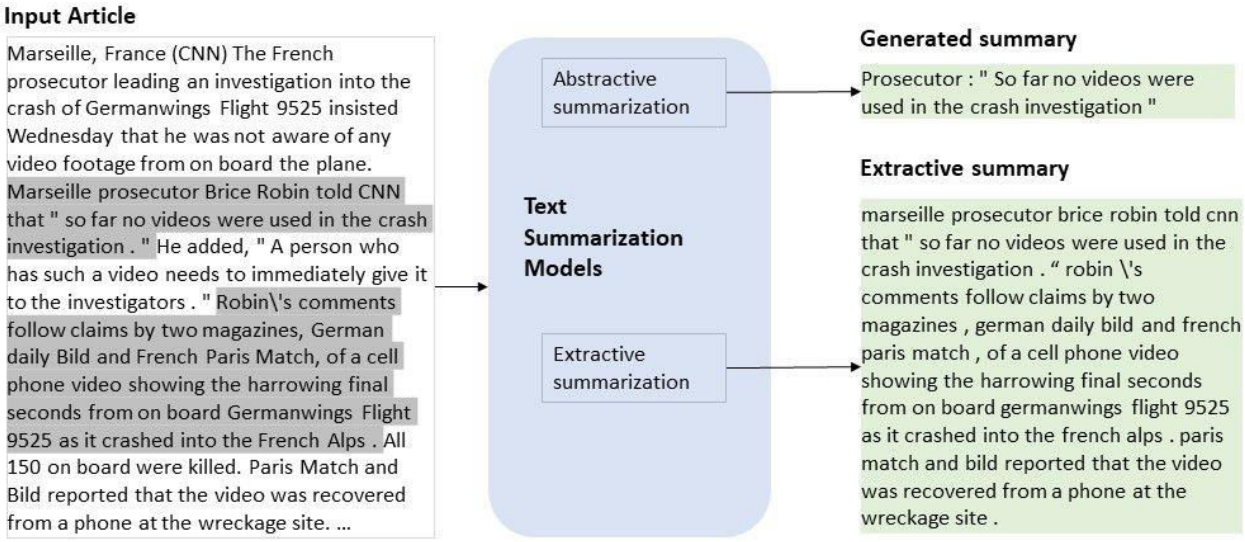
This is highly useful for any situation in which you want to identify a topic of interest in a textual dataset, such as whether there is a problem that comes up again and again in customer emails.

Text summarization

This NLP technique summarizes a text in a coherent way, and it’s great for extracting useful information from a source. While a human would have to read an entire document in order to write an accurate summary of it, which takes quite a bit of time, automatic text summarization can do it much more quickly.

There are two types of text summarization:

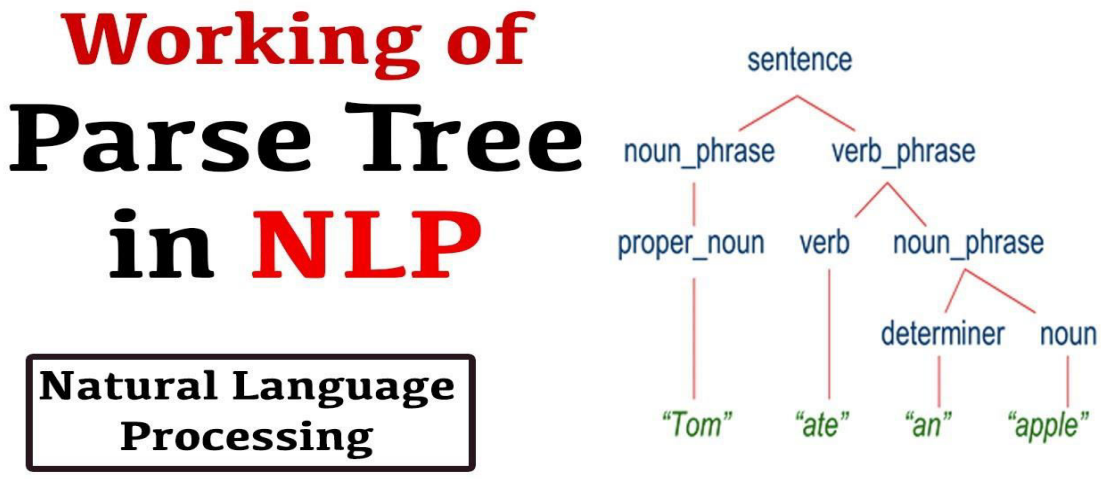
- **Extraction-based** – This technique pulls key phrases and words from the document to make a summary without changing the original text.
- **Abstraction-based** – This technique creates new phrases and sentences based on the original document, essentially paraphrasing it.



An example from the Microsoft tech community of how the two types of text summarization work.

Parsing

Parsing is the process of figuring out the grammatical structure of a sentence, determining which words belong together as phrases and which are the subject or object of a verb. This NLP technique offers additional context about a text in order to help with processing and analyzing it accurately.



This is how parsing might work on a short sentence.

Named entity recognition

Named entity recognition (NER) is a type of information extraction that locates and tags "named entities" with predefined keywords such as names, locations, dates, events, and more.

In addition to tagging a document with keywords, NER also keeps track of how many times a named entity is mentioned in a given dataset. NER is similar to keyword extraction, but the extracted keywords are put into predefined categories.

NER can be used to identify how often a certain term or topic is mentioned in a given data set. For example, it might be used to identify that a certain issue, tagged as a word like “slow” or “expensive,” comes up again and again in customer reviews.

A sample by Shaip of how named entity recognition works.

TF-IDF

TD-IDF, which stands for term frequency-inverse document frequency, is a statistical technique that determines the relevance of a word to one document in a collection of documents. It works by looking at two metrics: the number of times a word appears in a given document and the number of times the same word appears in a set of documents.

If a word is common in every document, it won’t receive a high score, even if it appears many times. But if a word frequently repeats in one document while rarely appearing in the rest of the documents in a set, it will rank high, suggesting it is highly relevant to that one document in particular.

Natural language processing applications

NLP is a quickly developing technology with many different applications for organizations of every kind. Some of the different ways a business can benefit from NLP include:

- **Machine translation** – Using NLP, computers can translate large amounts of text from a target to a source language, which can be used for customer support, data mining, and even publishing multilingual content.
- **Information retrieval** – NLP can be used to quickly access and retrieve information based on a user's query from text repositories such as file servers, databases, and the internet.
- **Sentiment analysis** – This NLP technique can be used to monitor brand and product sentiment to help with customer service and product sentiment, among other applications.
- **Information extracting** – This process, which includes retrieving information from unstructured data and extracting it into structured, editable formats, can be used for business intelligence, including competitive intelligence.
- **Question answering** – Question answering uses NLP to give an answer to a question asked in natural human language and can be used for chatbots and customer support.

Natural language processing examples

Here are just a few more concrete examples of ways an organization might apply NLP to its business processes.

NLP in ChatGPT

One of the most popular recent applications of NLP technology is ChatGPT, the trending AI chatbot that's probably all over your social media feeds. ChatGPT is fueled by NLP technology, using a multi-layer transformer network to generate human-like written responses to inquiries submitted in natural human language. ChatGPT uses unsupervised learning, which means it can generate responses without being told what the correct answer is.

ChatGPT is an exciting step forward in the application of NLP technology for businesses and individuals alike, with many saying it can rival even Google. Possible uses for ChatGPT include customer service, translation, summarization, and even content writing.

NLP for customer experience analytics

Using NLP for social listening and customer review analysis can lead to tremendous insight into what customers are thinking and saying about a brand and its products. With sentiment analysis and text classification, companies can:

- Understand general sentiment about the brand – Does the public feel positively or negatively about us?
- Identify what customers like and dislike about a service or product.
- Learn what new products customers might be interested in.
- Know which products to scale and which to pull back on.

- Discover insights that can be used to improve customer experience and boost customer satisfaction.

For example, let's say spicy chocolate brand Shock-O just released a new Popping Jalapeno Chocolate and wants to get a sense of whether or not customers like it. Shock-O can use an NLP-powered tool to analyze customer sentiment and learn what people are saying about the Popping Jalapeno Chocolate, whether they speak about it positively or negatively, and what themes come up again and again in reviews of this product.

All of this information can then be used to determine whether to continue producing Popping Jalapeno Chocolate, whether to increase or decrease its production of it, whether to make it spicier or less spicy, etc.

NLP for customer service

90% of customers believe that it is essential or very important to receive an immediate response when they have a question. Yet human customer service representatives are limited in availability and bandwidth.

This is just one reason why NLP-powered chatbots are growing in popularity. By being able to properly understand and analyze customer inquiries, chatbots can offer the necessary answers to questions, helping to improve customer satisfaction while cutting down on agents' workload.

NLP can also be used to process and analyze customer service surveys and tickets in order to better understand what issues customers are having, what they're happy with, what they're unhappy with and more. All of this serves as crucial data for boosting customer happiness, which will, in turn, increase customer retention and improve word-of-mouth.

NLP for recruitment

HR professionals spend countless hours reviewing resumes in order to identify suitable candidates. NLP can make this process much more efficient by taking over the screening process and analyzing resumes for certain keywords.

For example, you might set up an NLP system to flag any resume that uses the word "Python" or "leadership" for a human to review later on.

This can increase the likelihood of finding strong candidates, helping an organization fill open positions more quickly and with better talent. What's more, it can also free up HR professionals' time to focus on tasks that require more strategic thinking.

1.2 Regular Expressions

•Formally, a regular expression is an algebraic notation for characterizing a set of string.

•Regular expressions are particularly useful for searching in texts, when we have a pattern to search for and a corpus of texts to search through.

•The corpus can be a single document or a collection.

Regular Expressions: Disjunctions

•Letters inside square brackets []

•Ranges [A-Z]

Pattern	**Matches**
[wW]oodchuck	Woodchuck, woodchuck
[1234567890]	Any digit

1.3 Words

•we need to decide what counts as a word.

•He stepped out into the hall, was delighted to encounter a water brother

•13 words if we don't count punctuation marks as words.

•15 if we count punctuation

•Counting words like comma, period, etc are depends on the task.

The Switchboard corpus of American English telephone conversations between strangers was collected in the early 1990s;

•an utterance is the spoken correlate of a sentence:

I do uh main- mainly business data processing

•This utterance has two kinds of disfluencies.

–The broken-off word main- is called a fragment.

–Words like uh and um are called fillers or filled pauses.

1.4 What is a corpus?

A corpus is a collection of authentic text or audio organized into datasets. Authentic here means text written or audio spoken by a native of the language or dialect. A corpus can be made up of everything from newspapers, novels, recipes, radio broadcasts to television shows, movies, and tweets. In natural language processing, a corpus contains text and speech data that can be used to train AI and machine learning systems. If a user has a specific problem or objective they want to address, they'll need a collection of data that supports, or at least is a representation of, what they're looking to achieve with machine learning and NLP.

What are the features of a good corpus?

- **Large corpus size:**Generally, the larger the size of a corpus, the better. Large quantities of specialized datasets are vital to training algorithms designed to perform sentiment analysis.
- **High-quality data:**High quality is crucial when it comes to the data within a corpus. Due to the large volume of data required for a corpus, even minuscule errors in the training data can lead to large-scale errors in the machine learning system's output.
- **Clean data:**Data cleansing is also vital for creating and maintaining a high-quality corpus. Data cleansing allows identifying and eliminating any errors or duplicate data to create a more reliable corpus for NLP.
- **Balance:**A high-quality corpus is a balanced corpus. While it can be tempting to fill a corpus with everything and anything available, if one doesn't streamline and structure the data collection process, it could unbalance the relevance of the dataset.

What are the challenges regarding creating a corpus?

Deciding the type of data needed to solve the problem statement

Availability of data

Quality of the data

Adequacy of the data in terms of the amount

NLP CUSTOM CORPUS

What is a corpus?

A corpus can be defined as a collection of text documents. It can be thought as just a bunch of text files in a directory, often alongside many other directories of text files.

1.5 Text Normalization

Text normalization is a key step in natural language processing (NLP). It involves cleaning and preprocessing text data to make it consistent and usable for different NLP tasks. The process includes a variety of techniques, such as case normalization, punctuation removal, stop word removal, stemming, and lemmatization.

Steps to carry out text normalization in NLP

1. Case Normalization

Case normalization is converting all text to lowercase or uppercase to standardize the text. This technique is useful when working with text data that contains a mix of uppercase and lowercase letters.

Example text normalization

Input: “The quick BROWN Fox Jumps OVER the lazy dog.”

Output: “the quick brown fox jumps over the lazy dog.”

Advantages

- It eliminates case sensitivity, making text data consistent and easier to process.
- It reduces the dimensionality of the data, which can improve the performance of NLP algorithms.

Disadvantages

- It can lead to loss of information, as capitalization can indicate proper nouns or emphasis.

Text normalization code in Python

```
text = "The quick BROWN Fox Jumps OVER the lazy dog."
text = text.lower()
print(text)
```

2. Punctuation Removal

Punctuation removal is the process of removing special characters and punctuation marks from the text. This technique is useful when working with text data containing many punctuation marks, which can make the text harder to process.

Example text normalization

Input: “The quick BROWN Fox Jumps OVER the lazy dog!!!”

Output: “The quick BROWN Fox Jumps OVER the lazy dog”

Advantages

- It removes unnecessary characters, making the text cleaner and easier to process.
- It reduces the dimensionality of the data, which can improve the performance of NLP algorithms.

Disadvantages

- It can lead to loss of information, as punctuation marks can indicate sentiment or emphasis.

Text normalization code in Python

```
import string

text = "The quick BROWN Fox Jumps OVER the lazy dog!!!"

text = text.translate(text.maketrans("", "", string.punctuation))

print(text)
```

3. Stop Word Removal

Stop word removal is the process of removing common words with little meaning, such as “the” and “a”. This technique is useful when working with text data containing many stop words, which can make the text harder to process.

Example text normalization

Input: “The quick BROWN Fox Jumps OVER the lazy dog.”

Output: “quick BROWN Fox Jumps OVER lazy dog.”

Advantages

- It removes unnecessary words, making the text cleaner and easier to process.
- It reduces the dimensionality of the data, which can improve the performance of NLP algorithms.

Disadvantages

- It can lead to loss of information, as stop words can indicate context or sentiment.

Text normalization code in Python

```
from nltk.corpus import stopwords
text = "The quick BROWN Fox Jumps OVER the lazy dog."
stop_words = set(stopwords.words("english"))
words = text.split()
filtered_words = [word for word in words if word not in stop_words]
text = " ".join(filtered_words)
print(text)
```

4. Stemming

Stemming is reducing words to their root form by removing suffixes and prefixes, such as “running” becoming “run”. This method is helpful when working with text data that has many different versions of the same word, which can make the text harder to process.

Example text normalization

Input: “running,runner,ran”

Output: “run,run,run”

Advantages

- It reduces the dimensionality of the data, which can improve the performance of NLP algorithms.
- It makes it easier to identify the core meaning of a word.

Disadvantages

- It can lead to loss of information, as the root form of a word may not always be the correct form.
- It may produce non-existent words.

Text normalization code in Python

```
from nltk.stem import PorterStemmer
stemmer = PorterStemmer()
text = "running,runner,ran"
words = text.split(",")
stemmed_words = [stemmer.stem(word) for word in words]
text = ",".join(stemmed_words)
print(text)
```

5. Lemmatization

Lemmatization is reducing words to their base form by considering the context in which they are used, such as “running” becoming “run”. This technique is similar to stemming, but it is more accurate as it considers the context of the word.

Example text normalization

Input: “running,runner,ran”

Output: “run,runner,run”

Advantages

- It reduces the dimensionality of the data, which can improve the performance of NLP algorithms.
- It makes it easier to identify the core meaning of a word while preserving context.

Disadvantages

- It can be more computationally expensive than stemming.
- It may not be able to handle all words or forms.

Text normalization code in Python

```
from nltk.stem import WordNetLemmatizer
lemmatizer = WordNetLemmatizer()
text = "running,runner,ran"
words = text.split(",")
lemmatized_words = [lemmatizer.lemmatize(word) for word in words]
text = ",".join(lemmatized_words)
print(text)
```

6. Tokenization

Tokenization is the process of breaking text into individual words or phrases, also known as “tokens”. This technique is useful when working with text data that needs to be analyzed at the word or phrase level, such as in text classification or language translation tasks.

Example text normalization

Input: “The quick BROWN Fox Jumps OVER the lazy dog.”

Output: [“The”, “quick”, “BROWN”, “Fox”, “Jumps”, “OVER”, “the”, “lazy”, “dog.”]

Advantages

- It allows for analysing and manipulating individual words or phrases in the text data.
- It can improve the performance of NLP algorithms that rely on word or phrase-level analysis.

Disadvantages

- It can lead to the loss of information, as the meaning of a sentence or text can change based on the context of words.
- It may not be able to handle all forms of text.

Text normalization code in Python

```
from nltk.tokenize import word_tokenize

text = "The quick BROWN Fox Jumps OVER the lazy dog."

tokens = word_tokenize(text)

print(tokens)
```

7. Replacing synonyms and Abbreviation to their full form to normalize the text in NLP

This technique is useful when working with text data that contains synonyms or abbreviations that need to be replaced by their full form.

Example text normalization

Input: “I’ll be there at 2pm”

Output: “I will be there at 2pm”

Advantages

- It makes text data more readable and understandable.
- It can improve the performance of NLP algorithms that rely on word or phrase-level analysis.

Disadvantages

- It can lead to the loss of information, as the meaning of a sentence or text can change based on the context of words.
- It may not be able to handle all forms of text.

Text normalization code in Python

```
text = "I'll be there at 2pm"
synonyms = {"I'll": "I will", "2pm": "2 pm"}
for key, value in synonyms.items():
    text = text.replace(key, value)
print(text)
```

8. Removing numbers and symbol to normalize the text in NLP

This technique is useful when working with text data that contain numbers and symbols that are not important for the NLP task.

Example text normalization

Input: “I have 2 apples and 1 orange #fruits”

Output: “I have apples and orange fruits”

Advantages

- It removes unnecessary numbers and symbols, making the text cleaner and easier to process.
- It reduces the dimensionality of the data, which can improve the performance of NLP algorithms.

Disadvantages

- It can lead to loss of information, as numbers and symbols can indicate quantities or sentiments.

Text normalization code in Python

```
import re
text = "I have 2 apples and 1 orange #fruits"
text = re.sub(r"[\d#]", "", text)
```

```
print(text)
```

9. Removing any remaining non-textual elements to normalize the text in NLP

Removing any remaining non-textual elements such as HTML tags, URLs, and email addresses This technique is useful when working with text data that contains non-textual elements such as HTML tags, URLs, and email addresses that are not important for the NLP task.

Example text normalization

Input: “Please visit example.com for more information or contact me at info@example.com”

Output: “Please visit for more information or contact me at “

Advantages

- It removes unnecessary non-textual elements, making the text cleaner and easier to process.
- It reduces the dimensionality of the data, which can improve the performance of NLP algorithms.

Disadvantages

- It can lead to loss of information, as non-textual elements can indicate context or sentiment.

Text normalization code in Python

```
import re
text = "Please visit <a href='www.example.com'>example.com</a> for more information or contact me at info@example.com"
text = re.sub(r"(<[^>]+>)|(http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[!*\(\),]|(?:%[0-9a-fA-F][0-9a-fA-F]))+)", "", text)
print(text)
```

It’s important to note that these steps should be applied depending on the specific requirements of the NLP task and the type of text data being processed.

Text normalization is an iterative process, and the steps may be repeated multiple times.

Examples:-

Lemmatization:

•Draw/Prepare tree structure for lemmatizing of root word

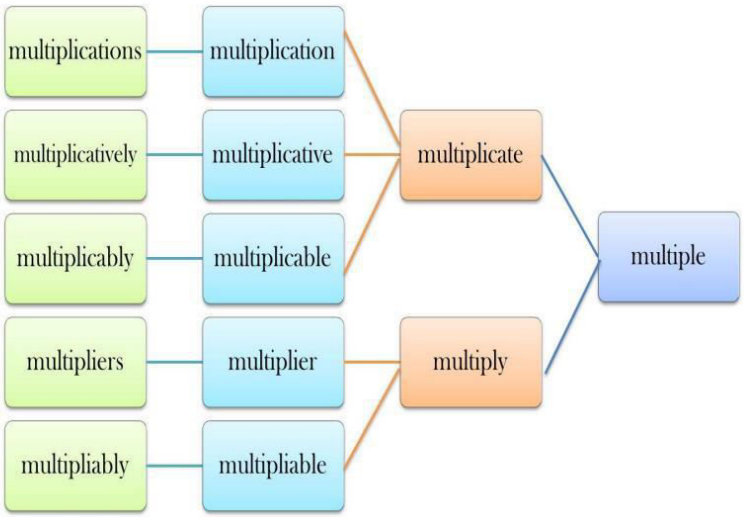
•Add

•Song

•See

•Go

NOTE: Refer here example covered in lectures as tree structures



1.6 Minimum Edit Distance

•The minimum edit distance between two strings

•Is the minimum number of editing operations

–Insertion

–Deletion

–Substitution

•Needed to transform one into the other

How to find the Min Edit Distance?

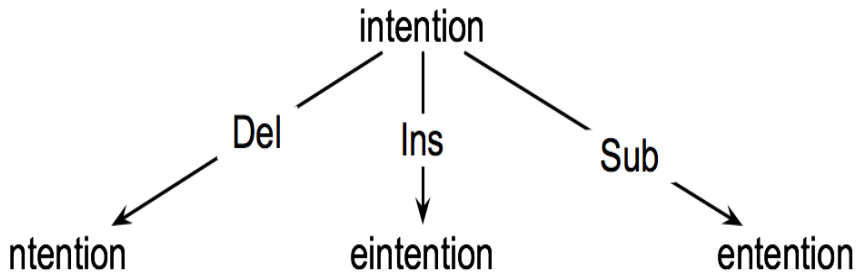
•Searching for a path (sequence of edits) from the start string to the final string:

–**Initial state:** the word we're transforming

–**Operators:** insert, delete, substitute

–**Goal state:** the word we're trying to get to

–**Path cost:** what we want to minimize: the number of edits



Minimum Edit Distance

```
I N T E * N T I O N
| | | | | | | | | |
* E X E C U T I O N
d s s   i s
```

d – deletion
s – substitution
i – insertion

If each operation has cost of 1
–Distance between these is 5

Examples:

Given two strings **str1** and **str2** of length **M** and **N** respectively and below operations that can be performed on **str1**. Find the minimum number of edits (operations) to convert '**str1**' into '**str2**'.

- **Operation 1 (INSERT)**: Insert any character before or after any index of **str1**
- **Operation 2 (REMOVE):** Remove a character of **str1**
- **Operation 3 (Replace):** Replace a character at any index of **str1** with some other character.

Note: All of the above operations are of equal cost.
Examples:
Input: *str1 = "geek", str2 = "gesek"*
Output: *1*
Explanation: *We can convert str1 into str2 by inserting a 's' between two consecutive 'e' in str2.*
Input: *str1 = "cat", str2 = "cut"*
Output: *1*
Explanation: *We can convert str1 into str2 by replacing 'a' with 'u'.*
Input: *str1 = "sunday", str2 = "saturday"*
Output: *3*
Explanation: *Last three and first characters are same. We basically need to convert "un" to "atur". This can be done using below three operations. Replace 'n' with 'r', insert t, insert a*

Illustration of Edit Distance:
Let's suppose we have ***str1="GEEXSFRGEEKKS"*** *and* ***str2="GEEKSFORGEEKS"***
Now to convert ***str1*** *into* ***str2*** *we would require* ***3*** *minimum operations:*
*Operation 1: Replace '****X****' to '****K****'*
*Operation 2: Insert '****O****' between '****F****' and '****R****'*
*Operation 3: Remove second last character i.e. '****K****'*
Refer the below image for better understanding.