NATURAL LANGUAGE PROCESSING

By Enid Roman
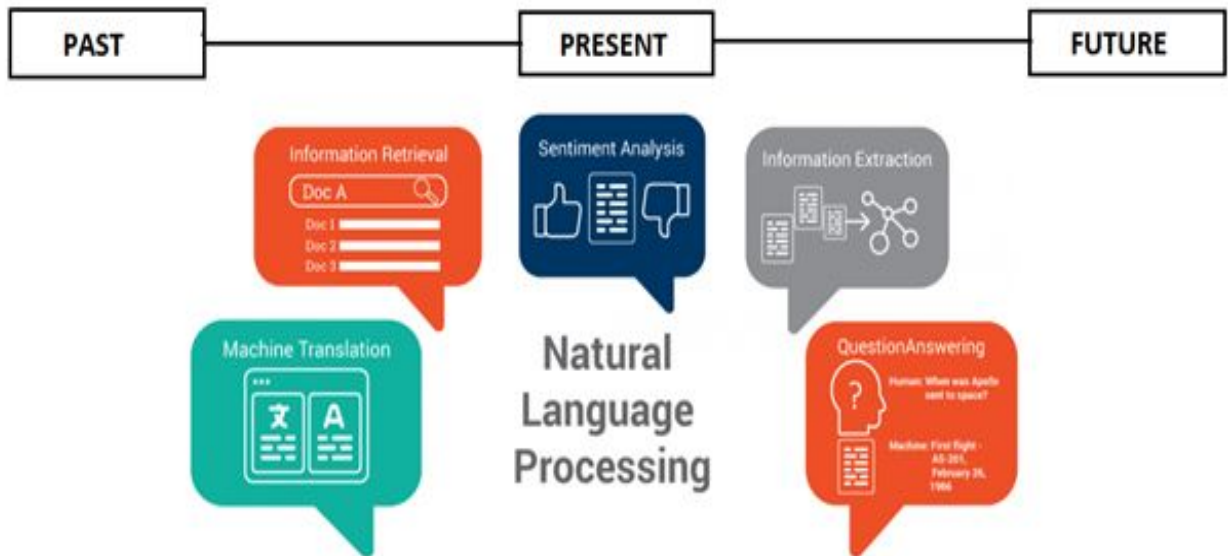
WHAT IS IT AND WHY IT MATTERS?

Natural language processing (NLP) is a branch of artificial intelligence that helps computers understand, interpret and manipulate human language. NLP draws from many disciplines, including computer science and computational linguistics, in its pursuit to fill the gap between human communication and computer understanding.
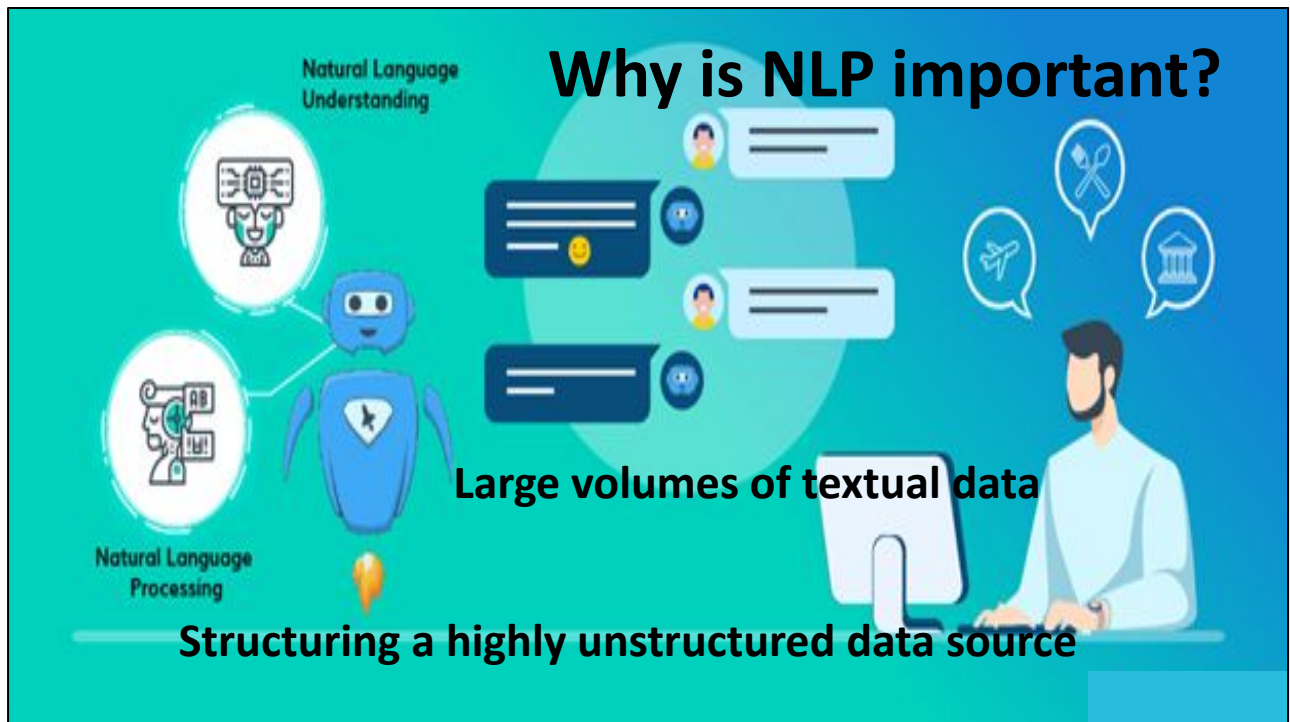
Evolution of Natural Language Processing

While natural language processing isn't a new science, the technology is rapidly advancing thanks to an increased interest in human-to-machine communications, plus an availability of big data, powerful computing and enhanced algorithms.

As a human, you may speak and write in English, Spanish or Chinese. But a computer's native language – known as machine code or machine language – is largely incomprehensible to most people. At your device's lowest levels, communication occurs not with words but through millions of zeros and ones that produce logical actions.

Why is NLP important?

Large volumes of textual data

Natural language processing helps computers communicate with humans in their own language and scales other language-related tasks.

Today's machines can analyze more language-based data than humans, without fatigue and in a consistent, unbiased way. Considering the staggering amount of unstructured data that's generated every day, from medical records to social media, automation will be critical to fully analyze text and speech data efficiently.

Structuring a highly unstructured data source

While supervised and unsupervised learning, and specifically deep learning, are now widely used for modeling human language, there's also a need for syntactic and semantic understanding and domain expertise that are not necessarily present in these machine learning approaches. NLP is important because it helps resolve ambiguity in language and adds useful numeric structure to the data for many downstream applications, such as speech recognition or text analytics.
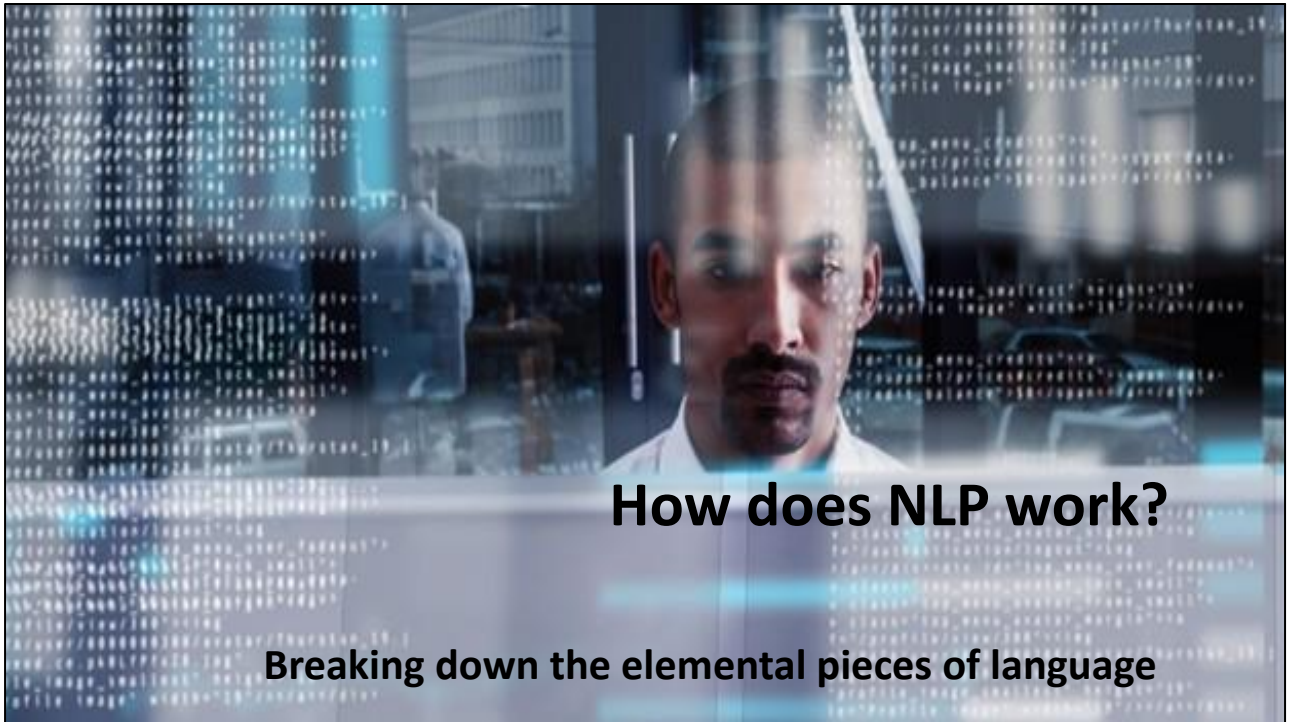
How does NLP work?

Breaking down the elemental pieces of language

Natural language processing includes many different techniques for interpreting human language, ranging from statistical and machine learning methods to rules-based and algorithmic approaches. We need a broad array of approaches because the text- and voice-based data varies widely, as do the practical applications.

Basic NLP tasks include tokenization and parsing, lemmatization/stemming, part-of-speech tagging, language detection and identification of semantic relationships.

In general terms, NLP tasks break down language into shorter, elemental pieces, try to understand relationships between the pieces and explore how the pieces work together to create meaning.

**These underlying tasks are often used in higher-level NLP capabilities, such as**:

- ● **Content categorization**. A linguistic-based document summary, including search and indexing, content alerts and duplication detection.
- ● **Topic discovery and modeling.** Accurately capture the meaning and themes in text collections and apply advanced analytics to text, like optimization and forecasting.
- ● **Corpus Analysis.** Understand corpus and document structure through output statistics for tasks such as sampling effectively, preparing data as input for further models and strategizing modeling approaches.
- ● **Contextual extraction.** Automatically pull structured information from text-based sources.
- ● **Sentiment analysis.** Identifying the mood or subjective opinions within large amounts of text, including average sentiment and opinion mining.
- ● **Speech-to-text and text-to-speech conversion.** Transforming voice commands into written text, and vice versa.
- ● **Document summarization.** Automatically generating synopses of large bodies of text and detect represented languages in multi-lingual corpora (documents).
- ● **Machine translation.** Automatic translation of text or speech from one language to another.

In all these cases, the overarching goal is to take raw language input and use linguistics and algorithms to transform or enrich the text in such a way that it delivers greater value.
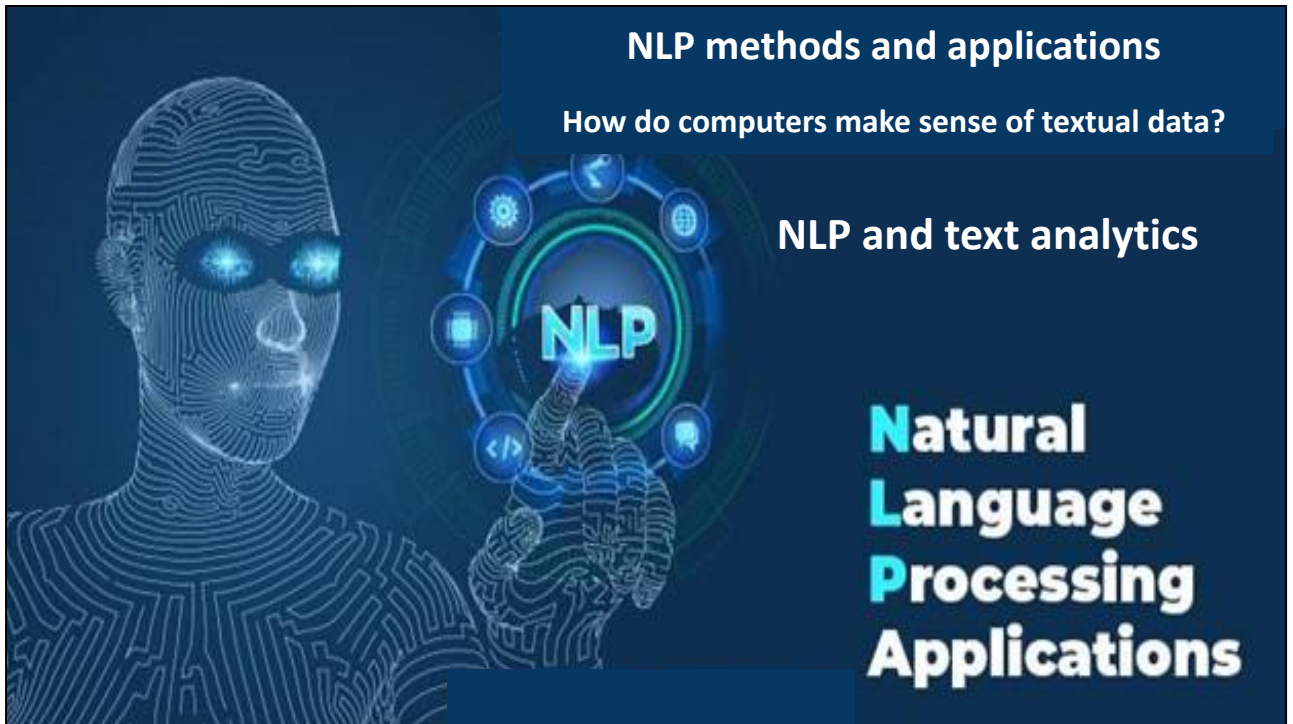
NLP methods and applications

How do computers make sense of textual data?

NLP and text analytics

Natural language processing goes hand in hand with text analytics, which counts, groups and categorizes words to extract structure and meaning from large volumes of content. Text analytics is used to explore textual content and derive new variables from raw text that may be visualized, filtered, or used as inputs to predictive models or other statistical methods.

NLP and text analytics are used together for many applications, including:

- **Investigative discovery**
- **Subject-matter expertise**
- **Social media analytics**

NLP and text analytics are used together for many applications, including:

- **Investigative discovery**. Identify patterns and clues in emails or written reports to help detect and solve crimes.
- **Subject-matter expertise**. Classify content into meaningful topics so you can take action and discover trends.
- **Social media analytics.** Track awareness and sentiment about specific topics and identify key influencers.
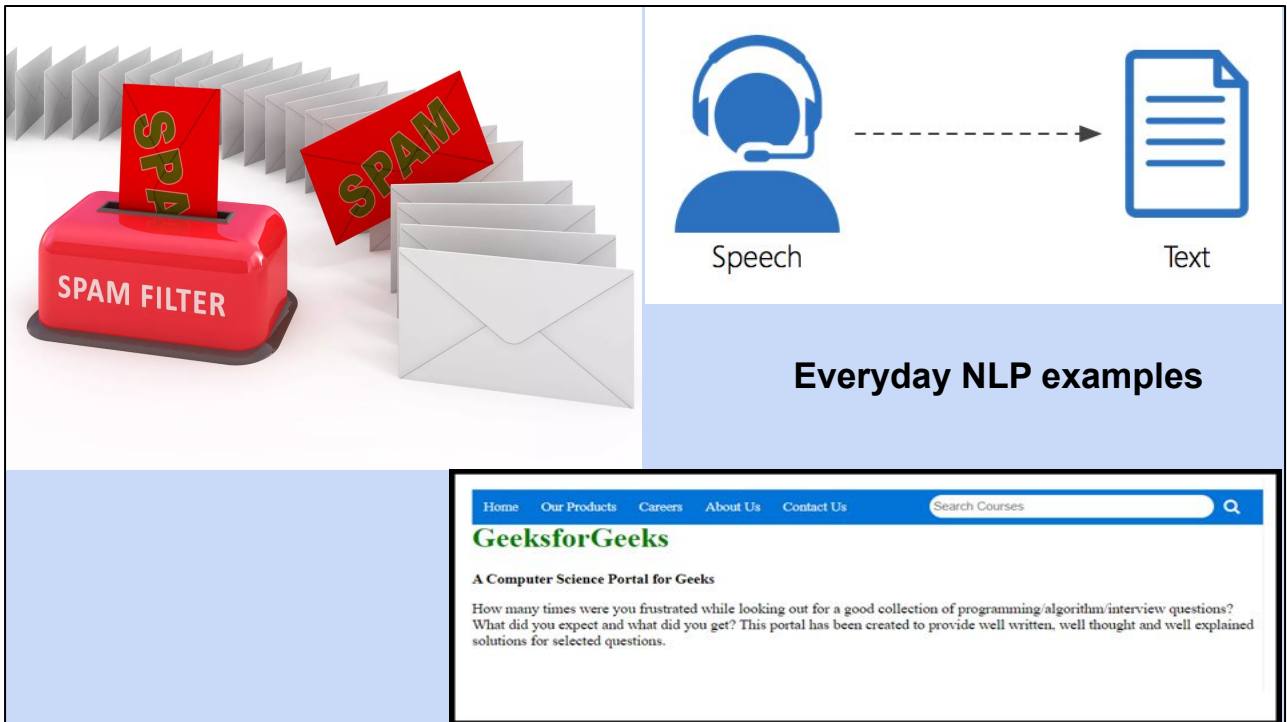
Everyday NLP examples

There are many common and practical applications of NLP in our everyday lives. Beyond conversing with virtual assistants like Alexa or Siri, here are a few more examples:

- Have you ever looked at the emails in your spam folder and noticed similarities in the subject lines? You're seeing Bayesian spam filtering, a statistical NLP technique that compares the words in spam to valid emails to identify junk mail.
- Have you ever missed a phone call and read the automatic transcript of the voicemail in your email inbox or smartphone app? That's speech-to-text conversion, an NLP capability.
- Have you ever navigated a website by using its built-in search bar, or by selecting suggested topic, entity, or category tags? Then you've used NLP methods for search, topic modeling, entity extraction and content categorization.

A subfield of NLP called natural language understanding (NLU) has begun to rise in popularity because of its potential in cognitive and AI applications. NLU goes beyond the structural understanding of language to interpret intent, resolve context, and word ambiguity, and even generate well-formed human language on its own. NLU algorithms must tackle the extremely complex problem of semantic interpretation – that is, understanding the intended meaning of spoken or written language, with all the subtleties, context, and inferences that we humans are able to comprehend.

The evolution of NLP toward NLU has a lot of important implications for businesses and consumers alike. Imagine the power of an algorithm that can understand the meaning and nuance of human language in many contexts, from medicine to law to the classroom. As the volumes of unstructured information continue to grow exponentially, we will benefit from computers' tireless ability to help us make sense of it all.