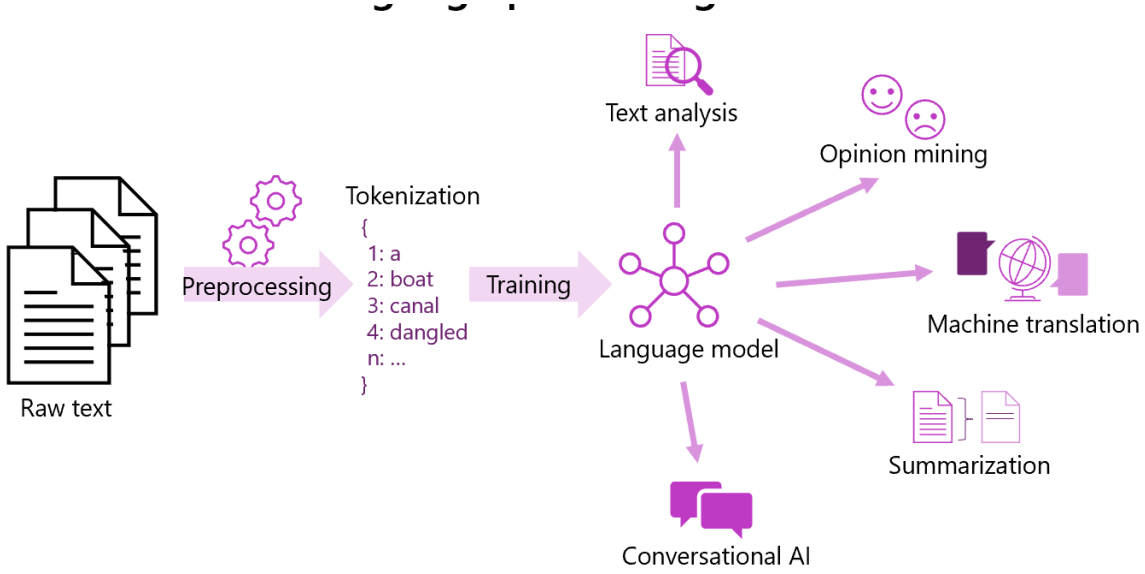
Agenda

- Natural Language Processing concepts
- Natural Language Processing capabilities in Azure

Natural Language Processing (NLP)

What is an Native Language Processing?

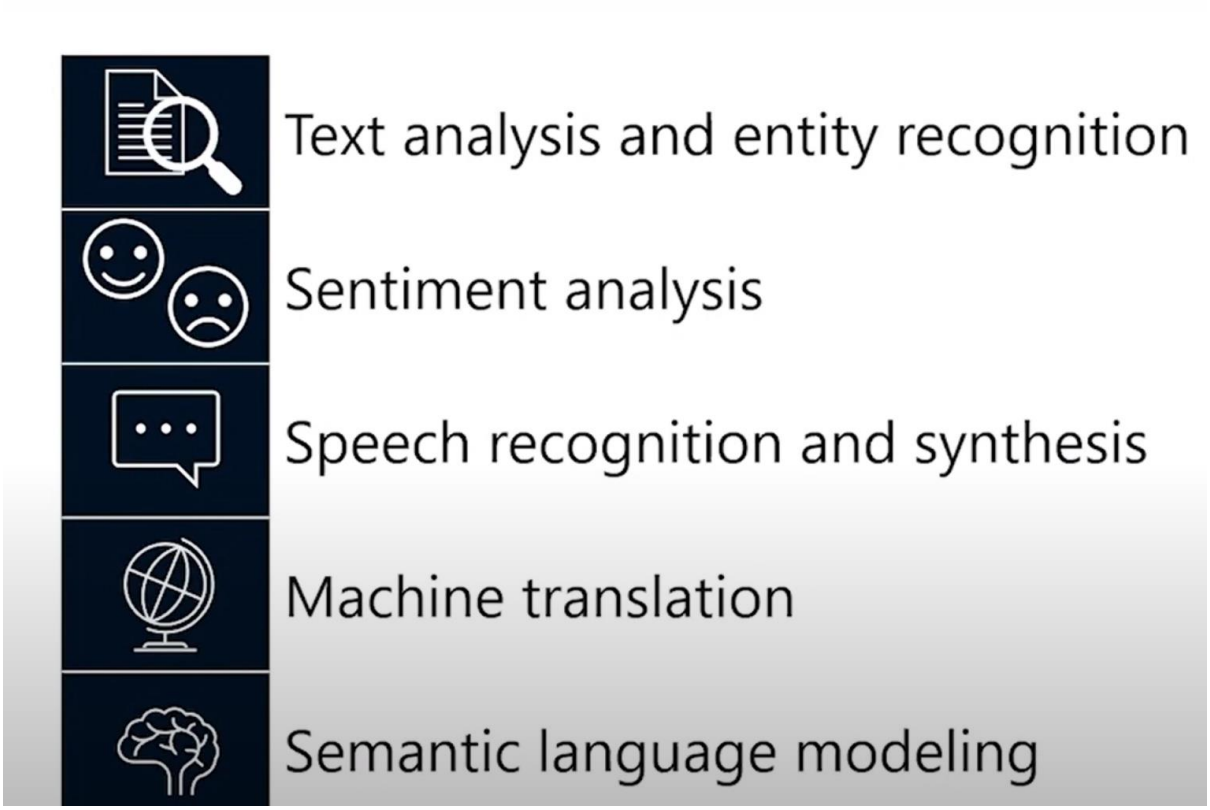


In order for computer systems to interpret the subject of a text in a similar way humans do, they use **natural language processing** (NLP), an area within AI that deals with understanding written or spoken language and responding in kind. Text analysis describes NLP processes that extract information from unstructured text.

Natural language processing might be used to create:

- A social media feed analyzer that detects sentiment for a product marketing campaign.
- A document search application that summarizes documents in a catalog.
- An application that extracts brands and company names from text.

Azure AI Language is a cloud-based service that includes features for understanding and analyzing text. **Azure AI Language** includes various features that support sentiment analysis, key phrase identification, text summarization, and conversational language understanding.



Understand Text Analytics

Principles: -

There are many pre-processing techniques that can be applied to text, including stop-word removal (eliminating common words with little semantic value, like “a”, “the”, etc.), Text normalization, stemming or lemmatization (coalescing words with the same root as a single word – so for example, “universe” and “universal” might both be represented as “univers”), and other techniques. Many of these approaches were extensively used in statistical modeling techniques that relied on counting the frequency of words in text to derive meaning, but are not so important in modern deep learning techniques. One thing that is commonly done though is to ***tokenize*** the words – representing each distinct word or phrase with a unique numeric identifier.The set of tokens for a given *corpus* of text represents the vocabulary for a language model.

Corpus

Statistical analysis of a body of text

Tokenization: -

The first step in analyzing a corpus is to break it down into *tokens*. For the sake of simplicity, you can think of each distinct word in the training text as a token, though in reality, tokens can be generated for partial words, or combinations of words and punctuation.

Example:

"*we choose to go to the moon*"

The phrase can be broken down into the following tokens, with numeric identifiers:

1. we
2. choose
3. to
4. go
5. the
6. moon

Notice that "to" (token number 3) is used twice in the corpus. The phrase "*we choose to go to the moon*" can be represented by the tokens [*1,2,3,4,3,5,6*].

We've used a simple example in which tokens are identified for each distinct word in the text. However, consider the following concepts that may apply to tokenization depending on the specific kind of NLP problem you're trying to solve:

1. **Text normalization**: Before generating tokens, you may choose to *normalize* the text by removing punctuation and changing all words to lower case. For analysis that relies purely on word frequency, this approach improves overall performance. However, some semantic meaning may be lost - for example, consider the sentence "*Mr Banks has worked in many banks.*". You may want your analysis to differentiate between the person *Mr Banks* and the *banks* in which he has worked. You may also want to consider "*banks.*" as a separate token to "*banks*" because the inclusion of a period provides the information that the word comes at the end of a sentence
2. **Stop word removal**. Stop words are words that should be excluded from the analysis. For example, "*the*", "*a*", or "*it*" make text easier for people to read but add little semantic meaning. By excluding these words, a text analysis solution may be better able to identify the important words.
3. **n-grams** are multi-term phrases such as "I have" or "he walked". A single word phrase is a *unigram*, a two-word phrase is a *bi-gram*, a three-word phrase is a *tri-gram*, and so on. By considering words as groups, a machine learning model can make better sense of the text.
4. **Stemming** is a technique in which algorithms are applied to consolidate words before counting them, so that words with the same root, like "power", "powered", and "powerful", are interpreted as being the same token.

Frequency analysis

- After tokenizing the words, you can perform some analysis to count the number of occurrences of each token.
- The most commonly used words (other than *stop words* such as "*a*", "*the*", and so on) can often provide a clue as to the main subject of a text corpus.

For example, the most common words in the entire text of the "go to the moon" speech we considered previously include "*new*", "*go*", "*space*", and "*moon*".

- If we were to tokenize the text as **bi-grams** (word pairs), the most common bi-gram in the speech is "*the moon*". From this information, we can easily surmise that the text is primarily concerned with space travel and going to the moon.

Understanding things in Easier Way

Certainly! Let's break down how n-gram analysis can be applied to the sentence "we choose to go to the moon" to understand its word relationships:

1. Tokenization and Building N-grams:

- First, we separate the sentence into single words (unigrams) and bi-grams (word pairs):
    - Unigrams: we, choose, to, go, to, the, moon
    - Bi-grams: we choose, choose to, to go, go to, to the, the moon

2. Analyzing Bi-grams:

Now, let's see what insights bi-grams might reveal:

- **"we choose":** This suggests an intentional decision or goal.
- **"choose to":** This emphasizes the act of choosing and reinforces the agency of "we."
- **"to go":** This implies movement or a journey.
- **"go to":** This strengthens the direction of the journey.
- **"to the":** This specifies the destination of the journey.
- **"the moon":** This is the most crucial bi-gram as it clearly states the target of the journey.

Understanding Importance:

While frequency analysis tells us which words appear most often ("the" and "to" might be frequent here), bi-grams reveal a deeper connection.

- **"The moon" as the Focus:** Even though individual words like "the" appear more often, "the moon" as a bi-gram stands out because it shows these two words are used together consistently. This points towards the moon being the central focus of the sentence.

Machine learning for text classification

- Another useful text analysis technique is to use a classification algorithm, such as *logistic regression*, to train a machine learning model.
- A common application of this technique is to train a model that classifies text as *positive* or *negative* in order to perform *sentiment analysis* or *opinion mining*.

For example, consider the following restaurant reviews, which are already labeled as **0** (*negative*) or **1** (*positive*):

- *The food and service were both great*: 1

• *A really terrible experience*: 0
• *Mmm! tasty food and a fun vibe*: 1
• *Slow service and substandard food*: 0