
NATURAL LANGUAGE
PROCESSING

NATURAL LANGUAGE PROCESSING

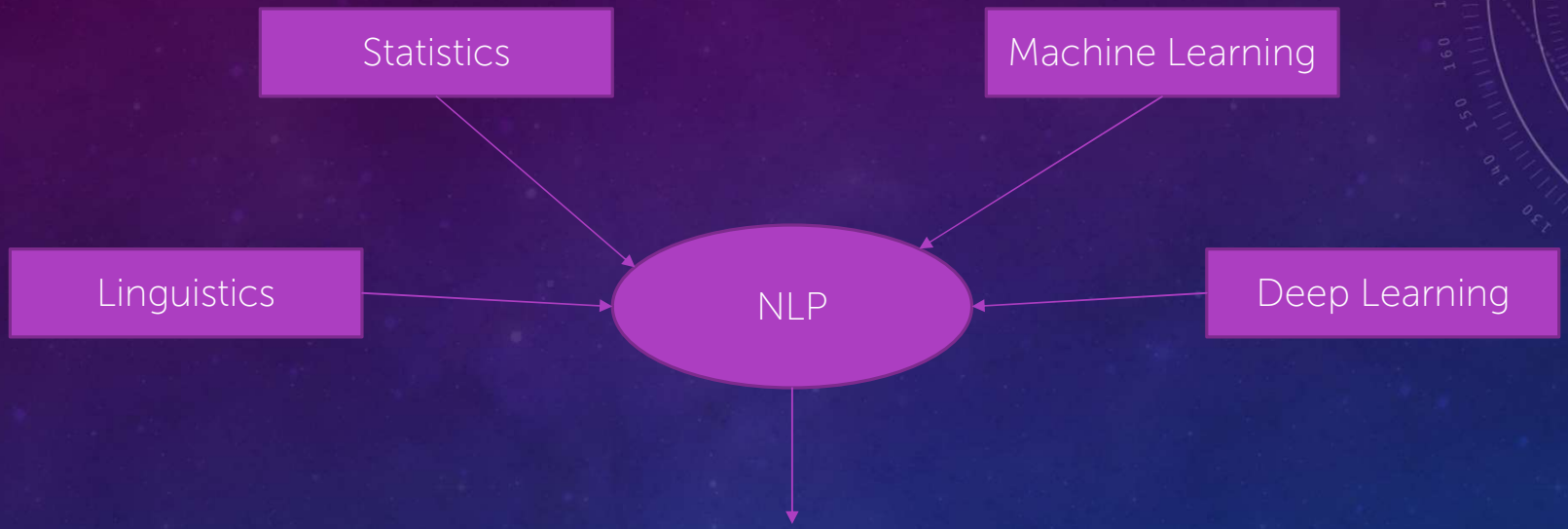
- Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, how to program computers to process and analyze large amounts of natural language data.
- The goal is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them.
- The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.

https://en.wikipedia.org/wiki/Natural_language_processing

" Natural language processing (NLP) refers to the branch of computer science - and more specifically, the branch of artificial intelligence or AI - concerned with giving computers the ability to understand text and spoken words in much the same way human beings can. "

What is natural language processing?

https://www.ibm.com/cloud/learn/natural-language-processing



1. Voice-operated GPS systems e.g., waze
2. Digital assistants e.g., Siri
3. Speech-to-Text software
4. Customer service Chatbots

NLP TASKS

Task	Description
Speech recognition	Also called speech-to-text, is the task of reliably converting voice data into text data. Speech recognition is required for any application that follows voice commands or answers spoken questions. What makes speech recognition especially challenging is the way people talk—quickly, slurring words together, with varying emphasis and intonation, in different accents, and often using incorrect grammar.
Part of speech tagging	Also called grammatical tagging, is the process of determining the part of speech of a particular word or piece of text based on its use and context. Part of speech identifies 'make' as a verb in 'I can make a paper plane,' and as a noun in 'What make of car do you own?'

NLP TASKS (CONT.)

Task	Description
Word sense disambiguation	The selection of the meaning of a word with multiple meanings through a process of semantic analysis that determine the word that makes the most sense in the given context. For example, word sense disambiguation helps distinguish the meaning of the verb 'make' in ‘make the grade’ (achieve) vs. ‘make a bet’ (place).
Named entity recognition	NEM, identifies words or phrases as useful entities. NEM identifies ‘Kentucky’ as a location or ‘Fred’ as a man's name.
Sentiment analysis	Attempts to extract subjective qualities—attitudes, emotions, sarcasm, confusion, suspicion—from text.

NLP TASKS (CONT.)

Task	Description
Co-reference resolution	The task of identifying if and when two words refer to the same entity. The most common example is determining the person or object to which a certain pronoun refers (e.g., 'she' = 'Mary'), but it can also involve identifying a metaphor or an idiom in the text (e.g., an instance in which 'bear' isn't an animal but a large hairy person).
Natural language generation	Sometimes described as the opposite of speech recognition or speech-to-text; it's the task of putting structured information into human language.

PYTHON AND THE NATURAL LANGUAGE TOOLKIT (NLTK)

- The Python programing language provides a wide range of tools and libraries for attacking specific NLP tasks.
- Many of these are found in the Natural Language Toolkit, or NLTK, an open-source collection of libraries, programs, and education resources for building NLP programs.

NATURAL LANGUAGE PROCESSING WITH PYTHON'S NLTK PACKAGE

HTTPS://REALPYTHON.COM/NLTK-NLP-PYTHON/

TOKENIZING

- Tokenizing – split the text by word or by sentence for easier processing.
  - Tokenizing by word: Words are like the atoms of natural language. They're the smallest unit of meaning that still makes sense on its own. Tokenizing your text by word allows you to identify words that come up particularly often. For example, if you were analyzing a group of job ads, then you might find that the word "Python" comes up often. That could suggest high demand for Python knowledge, but you'd need to look deeper to know more.
  - Tokenizing by sentence: When you tokenize by sentence, you can analyze how those words relate to one another and see more context. Are there a lot of negative words around the word "Python" because the hiring manager doesn't like Python? Are there more terms from the domain of herpetology than the domain of software development, suggesting that you may be dealing with an entirely different kind of python than you were expecting?

STOP WORDS

- Stop words are words that you want to ignore, so you filter them out of your text when you're processing it.
- Very common words like 'in', 'is', and 'an' are often used as stop words since they don't add a lot of meaning to a text in and of themselves.

STEMMING

- Stemming is a text processing task in which you reduce words to their root, which is the core part of a word.
- For example, the words “helping”, and “helper” share the root “help.”
- Stemming allows you to zero in on the basic meaning of a word rather than all the details of how it’s being used.

PARTS OF SPEECH

- Part of speech is a grammatical term that deals with the roles words play when you use them together in sentences.
- Tagging parts of speech, or POS tagging, is the task of labeling the words in your text according to their part of speech.
- In English, there are eight parts of speech.

Part of Speech	Role	Examples
Noun	Is a person, place, or thing	mountain, bagel, Poland
Pronoun	Replaces a noun	you, she, we
Adjective	Gives information about what a noun is like	efficient, windy, colorful
Verb	Is an action or a state of being	learn, is, go
Adverb	Gives information about a verb, an adjective, or another adverb	efficiently ,always, very
Preposition	Gives information about how a noun or pronoun is connected to another word	from, about, at
Conjunction	Connects two other words or phrases	so, because, and
Injection	Is an exclamation	yay, ow, wow

LEMMATIZING

- Like stemming, lemmatizing reduces words to their core meaning, but it will give you a complete English word that makes sense on its own instead of just a fragment of a word like 'discoveri'.
- A lemma is a word that represents a whole group of words, and that group of words is called a lexeme.
- For example, if you were to look up the word “blending” in a dictionary, then you’d need to look at the entry for “blend,” but you would find “blending” listed in that entry.
- In this example, “blend” is the lemma, and “blending” is part of the lexeme. So when you lemmatize a word, you are reducing it to its lemma.

CHUNKING

- While tokenizing allows you to identify words and sentences, chunking allows you to identify phrases.
- A phrase is a word or group of words that works as a single unit to perform a grammatical function. Noun phrases are built around a noun.
- Here are some examples:
  - "A planet"
  - "A tilting planet"
  - "A swiftly tilting planet"

CHINKING

- Chinking is used together with chunking, but while chunking is used to include a pattern, chinking is used to exclude a pattern.

NAMED ENTITY RECOGNITION (NER)

- Named entities are noun phrases that refer to specific locations, people, organizations, and so on.
- With named entity recognition, you can find the named entities in your texts and determine what kind of named entity they are.
- For a complete list visit https://www.nltk.org/book/ch07.html#sec-ner

CORPUS AND CORPORA

- Now that you've done some text processing tasks with small example texts, you're ready to analyze a bunch of texts at once.
- A group of texts is called a corpus.
- NLTK provides several corpora covering everything from novels hosted by Project Gutenberg to inaugural speeches by presidents of the United States.

CONCORDANCE

- When you use a concordance, you can see each time a word is used, along with its immediate context.
- This can give you a peek into how a word is being used at the sentence level and what words are used with it.

THE END