Glossary

ASCII

A character set with 128 characters, including lower and uppercase letters, digits, and punctuation.

Constituency Parsing

A technique that focuses on first analyzing and determining the constituent parts of a sentence, then recursively breaking each constituent into smaller constituents, until finally individual tokens (usually words) remain.

Constructed language

A form of communication that uses a purposefully developed grammar and vocabulary, such as programming languages or the written/spoken language Esperanto.

Contraction map

A list of the keys that are contractions of phrases. Each key corresponds to an expanded form of an uncontracted phrase. Contraction maps are used in specific rules for expanding contractions.

Corpus

A large set of text documents. The plural form is corpora.

Deduplication

The removal of repetitive or redundant information. Deduping helps create a smaller vocabulary without any loss of information.

Dependency Parsing

A technique that attempts to recover syntactic relationships of words in a sentence by assuming that all words in a sentence have a dependency on other words in the sentence except one (called the "root").

Dependency tree

The graphical output of dependency parsing, which depicts a root vertex (usually a verb) and directed edges connecting each head node to all dependent nodes. There is a unique path from the root to any other word or phrase. Head nodes are also called "parent" nodes. Dependent nodes are also called "child," "modifier," or "subordinate" nodes.

Generic rule in contraction expansion

When expanding contractions, a generic rule would be used to expand common suffixes regardless of the full phrase. For example, a generic rule would expand "'s" to "has" regardless of the full phrase, which would be correct only some of the time; a generic rule here would turn "Mike's car" to "Mike has car" and not the possessive.

Lemma

A root or a dictionary form of a word.

Mixed rule in contraction expansion

When expanding contractions, a mixed rule would be where a specific rule is used first and then a generic rule could be used in the second iteration to fix any missed contractions.

Natural language

A form of written and verbal communication evolved naturally by humans through use. Examples include English, Russian, Spanish, etc., along with many dialects of these languages.

Natural Language Toolkit (NLTK)

A powerful Python package for working with human language data, including libraries to help with stemming, lemmatization, spelling correction, classification, tokenization, and more. For more complete information, see nltk.org.

Parts-of-Speech (POS) tagging

A set of categories that label the syntactic function of each word in a sentence. This could be a noun, verb, adjective, article, and more.

Preprocessing

A set of techniques used to standardize textual content with minimal loss of information. Examples of preprocessing techniques might include spelling correction, removing stopwords, reducing a vocabulary, expanding contractions, and more.

Regular Expressions (regex)

Regex is a powerful pattern matching technique for strings of Unicode characters. Regexes can be used for finding patterns, as well as for finding one pattern and replacing it with another. Regex rule syntax can be used in a variety of different programming languages.

Shallow parsing

Also called “chunking,” it is a technique that analyzes sentence structure by breaking it into its smallest parts (tokens that are usually words) and then grouping them into phrases. The focus here is more on the phrase identification than any deeper semantic meaning or relationship.

Slicing

Also called subsetting, this is a method to retrieve a substring from a document by indicating start and stop indices. The stop index is exclusive, so the substring will end one index before the stop value. For example, `‘python’[1:4]` returns `‘yth’`.

Summarization

A technique used to shrink documents or large publications to a much more user-friendly size for easier, faster reading.

spaCy

A Python package that loads an English model and the NLP will create what’s called the sequence of token objects.

Specific rule in contraction expansion

When expanding contractions, a specific rule would be used to expand contractions identified upfront using a key-value map of contracted phrases with their uncontracted forms. Specific rules only expand contractions that the program is told to look for, and so uncommon contractions are often overlooked.

Stem

A partial word with a truncated suffix.

Stopwords

Words that are of little or no significance to a document and that can be removed with no loss of information. For example, “is,” “the,” “when,” “for,” “to,” etc.

TextBlob

A useful Python library for text correction developed by Peter Norvig, the Director of Research at Google.

Token

A contiguous group of characters, such as words, characters, or subwords. Tokens are a fundamental unit in NLP.

Tokenization

The process by which strings are split or parsed into tokens, which are deduplicated to form a vocabulary. Tokenization is a necessary first step in many NLP techniques.

Unicode

A character set that was developed to represent many different characters in many different languages. This is in contrast to the ASCII character set, which had only 128 characters. By default, Python uses Unicode — specifically, UTF-8.

Whitespace characters

Characters in a string indicated by a `/` followed by a letter, such as `/t`. Common whitespace characters are for a tab `/t`, a newline character `/n`, a carriage return character `/r`, or a space character `/s`.