greatlearning
Learning for Life

NLP Basics

Introduction to Natural Language Processing

Agenda

greatlearning
Learning for Life

- Natural Language Processing Overview
- Text pre-processing: removal of html tags, handling accented characters, removal of special characters, tokenization, stop words, stemming and lemmatization
- Text analytics framework– NLTK, TextBlob, Spacy, BeautifulSoup
- Count- Vectorizer, Bag Of Words ,TF-IDF
- Text Classification
- Sentiment Analysis

NLP Basics (TOC)

greatlearning
Learning for Life

S.No	Topic	Scope	Objective
1	Introduction to NLP	Discuss NLP, its application, different tasks in NLP like text classification, segmentation, machine translation, semantic analysis	Understand what NLP stands for and how we can use in industry
2	NLP Terms	Standard NLP Terms	Understand the NLP Jargon
3	Text analytics framework- (NLTK, Spacy, TextBlob),Regex usage, BeautifulSoup	How NLTK can be used for textprocessing, Introduction to Spacy and TextBlob	Understand briefly NLP frameworks and their common usage , learn about text cleansing
4	Text pre-processing	Dealing with text-data, special character, stop words removal, html tags, tokenization, stemming, lemmatization	Understand important steps in text preprocessing

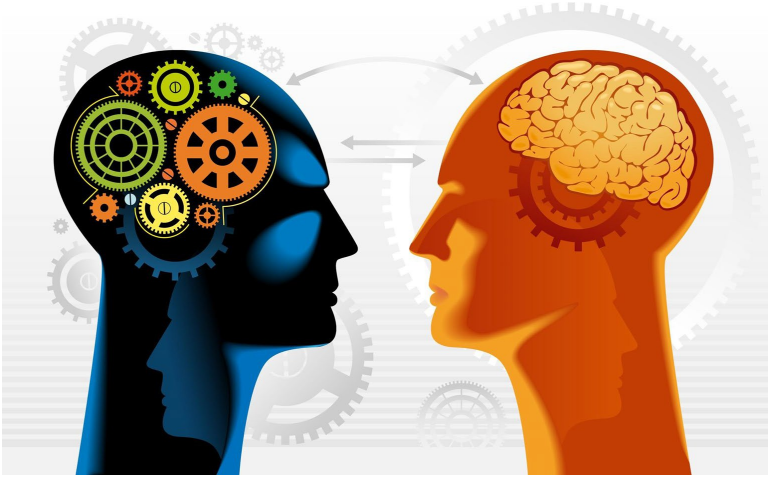
greatlearning
Learning for Life

Natural Language Processing

Natural Language Processing

greatlearning
Learning for Life

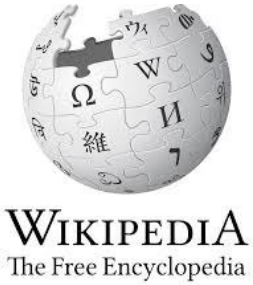
1. Natural Language Processing is a subfield of artificial intelligence concerned with methods of communication between computers and natural languages such as English, Hindi, etc.
2. It is an intersection of fields of Computer Science, linguistics, and AI
3. Objective of Natural Language processing is to perform useful tasks involving human languages like
    - Sentiment Analysis
    - Machine Translation
    - Part of Speech Tags
    - Human-Machine communication(chat-bots)



Why Study NLP?

greatlearning
Learning for Life

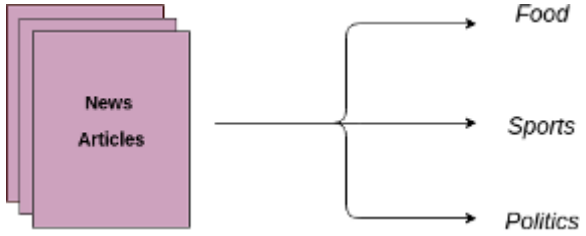
1. Language is involved in most of the activities that involve interaction between humans, e.g. reading, writing, speaking, listening.
2. Voice can be used as an interface for interactions between humans and machines e.g. Cortana, Google Assistant, Siri, Amazon Alexa.
3. There is massive amount of data available in text format which can used to derive insights from using NLP, e.g. blogs, research articles, consumer reviews, literature, discussion forums.



Different Tasks in NLP

greatlearning
Learning for Life

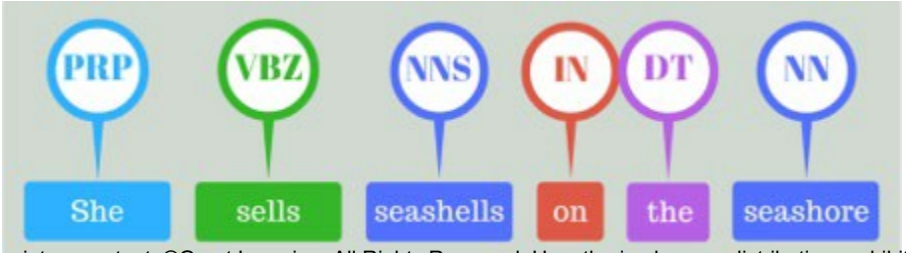
- **Text Classification**
  - Sentiment Analysis: Determining the general context of a review, whether it is positive or negative or neutral.
  - Consumer Complaints Classification: Categorizing complaints on consumer forums to respective departments.
- **Machine Translation**
  - Improving human-human interaction by translating sentences from one language to another.



Different Tasks in NLP

greatlearning
Learning for Life

- **Part of Speech Tagging**
  - In corpus linguistics, part-of-speech tagging (POS tagging or PoS tagging or POST), also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context.
  - A simplified form of this is the identification of words as nouns, verbs, adjectives, adverbs, etc.
  - Tag-set: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html



Different Tasks in NLP

greatlearning
Learning for Life

- **Word Segmentation**
  - In some languages, there is no space between words, or a word may contain smaller syllables. In such languages, word segmentation is the first step of NLP systems.
- **Semantic Analysis**
  - Semantic analysis of a corpus (a large and structured set of texts) is the task of building structures that approximate concepts from a large set of documents.
  - Application of Semantic Analysis :
    - Text Similarity
    - Context Recognition
    - Sentence Parsing
    - Topic Modelling

Why NLP is hard?

greatlearning
Learning for Life

1. Languages are changing everyday, new words, new rules, etc.
2. The number of tokens is not fixed. A natural language can have hundreds of thousands of different words, new words are created on the fly.
3. Words can have different meanings depending on context, and they can acquire new meanings over time (apple(a fruit), Apple(the company)], they can even change their parts of speech(Google--> to google).
4. Every language has its own uniqueness. Like in the case of English we have words, sentences, paragraphs and so on to limit our language. But in Thai, there is no concept of sentences.

Standard NLP Terms

greatlearning
Learning for Life

- Corpus: A body of text samples
- Document: A text sample
- Vocabulary: A list of words used in the corpus
- Language model: How the words are supposed to be organized

greatlearning
Learning for Life

Pre-processing Steps

Why do we need pre-processing

greatlearning
Learning for Life

- Textual data is unstructured and cannot be processed as it is.
- Text data also contains a lot of non- required items such as special characters, punctuations etc.
- We clean up the text corpus to make it processable by ML
- This text clean-up process is called text pre-processing

Text analytics Framework

greatlearning
Learning for Life

- NLTK: The Natural Language Toolkit is a complete platform that contains more than 50 corpora and lexical resources. It also provides the necessary tools, interfaces, and methods to process and analyze text data.

- Beautiful Soup : It can be used to scrape data from web and also for text cleaning with its inbuilt parsers.

Text analytics Framework

greatlearning
Learning for Life

- TextBlob: Provides several capabilities including text processing, phrase extraction, classification, POS tagging, text translation and sentiment analysis

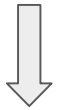
- Spacy : It is a production grade NLP library that offers similar functionality as that of NLTK & TextBlob

Removal of HTML tags

greatlearning
Learning for Life

Often, unstructured text contains a lot of noise, especially if you use techniques like web or screen scraping. HTML tags are typically one of these components which don't add much value towards understanding and analyzing text.

- strip_html_tags('<html><h2>Some important text</h2></html>')



'Some important text'

Handling accented characters

greatlearning
Learning for Life

Usually in any text corpus, you might be dealing with accented characters/letters, especially if you only want to analyze the English language.

Hence, we need to make sure that these characters are converted and standardized into ASCII characters.

A simple example— converting é to e.

- remove_accented_chars('Sómě Áccěntěd těxt')

' Some Accented text'

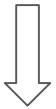
Removal of special characters

greatlearning
Learning for Life

Special characters and symbols are usually non-alphanumeric characters or even occasionally numeric characters (depending on the problem), which add to the extra noise in unstructured text.

Usually, simple regular expressions (regex) can be used to remove them.

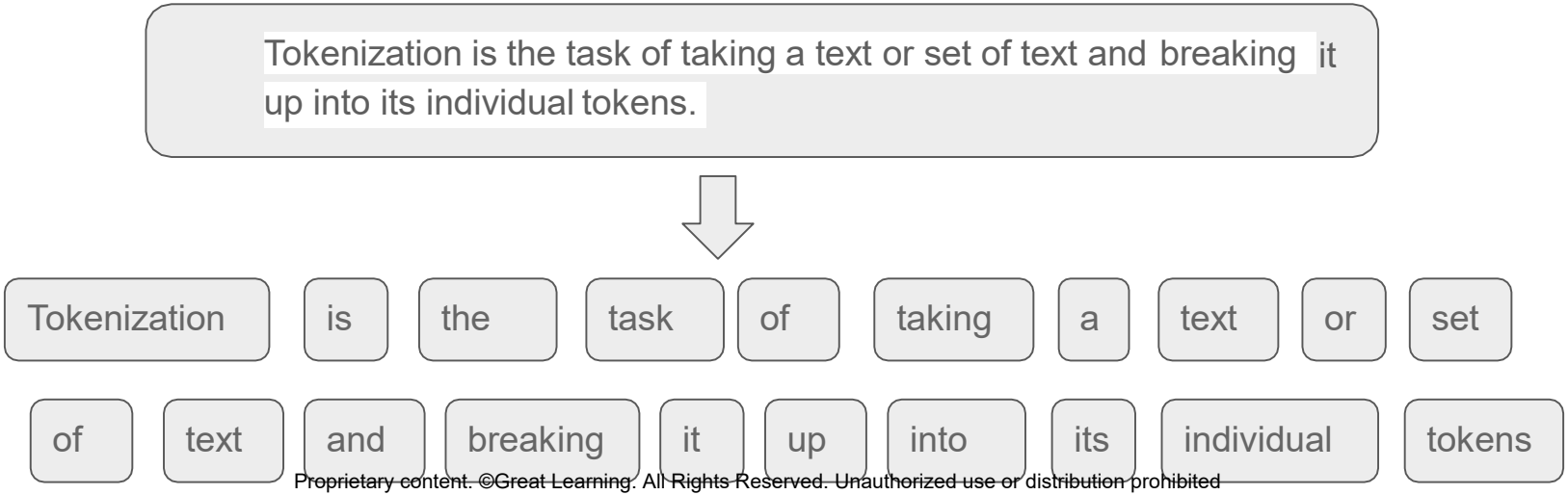
- remove_special_characters("Well this was fun! What do you think? 123#@!", remove_digits=True)



'Well this was fun What do you think '

Tokenization

- Tokenization is the task of taking a text or set of text and breaking it up into its individual tokens.
- Tokens are usually individual words (at least in languages like English).
- Tokenization can be achieved using different methods. Most common method is Whitespace tokenizer and Regexp Tokenizer.



Stop Words Removal

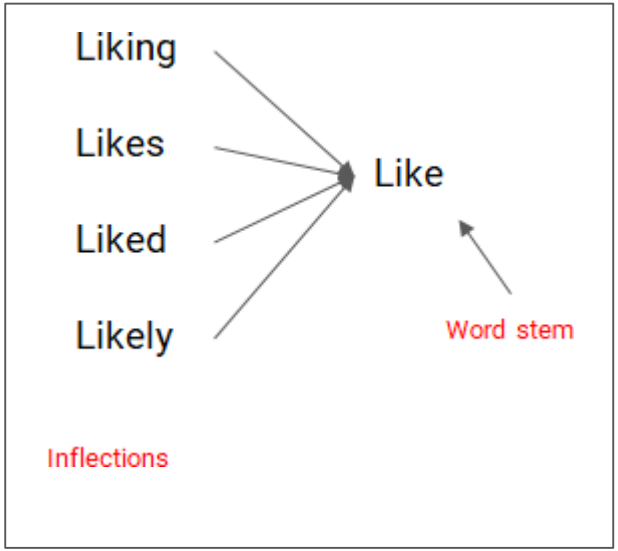
greatlearning
Learning for Life

- Stopwords are common words that carry less important meaning than keyword
- When using some bag of words based methods, i.e, countVectorizer or tf-idf that works on counts and frequency of the words, removing stopwords is great as it lowers the dimensionality
    - remove_stopwords("The, and, if are stopwords, computer is not")
    - Removing stopwords minimizes computation
- Not always a good idea
    - When working on problems where contextual information is important like machine translation, removing stop words is not recommended.

```
> stopwords("english")
  [1] "i"          "me"         "my"         "myself"     "we"
  [6] "our"        "ours"       "ourselves"  "you"        "your"
 [11] "yours"      "yourself"   "yourselves" "he"         "him"
 [16] "his"        "himself"    "she"        "her"        "hers"
 [21] "herself"    "it"         "its"        "itself"     "they"
 [26] "them"       "their"      "theirs"     "themselves" "what"
 [31] "which"      "who"        "whom"       "this"       "that"
 [36] "these"      "those"      "am"         "is"         "are"
 [41] "was"        "were"       "be"         "been"       "being"
 [46] "have"       "has"        "had"        "having"     "do"
 [51] "does"       "did"        "doing"      "would"      "should"
 [56] "could"      "ought"      "i'm"        "you're"     "he's"
 [61] "she's"      "it's"       "we're"      "they're"    "i've"
 [66] "you've"     "we've"      "they've"    "i'd"        "you'd"
 [71] "he'd"       "she'd"      "we'd"       "they'd"     "i'll"
 [76] "you'll"     "he'll"      "she'll"     "we'll"      "they'll"
 [81] "isn't"      "aren't"     "wasn't"     "weren't"    "hasn't"
 [86] "haven't"    "hadn't"     "doesn't"    "don't"      "didn't"
 [91] "won't"      "wouldn't"   "shan't"     "shouldn't"  "can't"
 [96] "cannot"     "couldn't"   "mustn't"    "let's"      "that's"
[101] "who's"      "what's"     "here's"     "there's"    "when's"
[106] "where's"    "why's"      "how's"      "a"          "an"
```

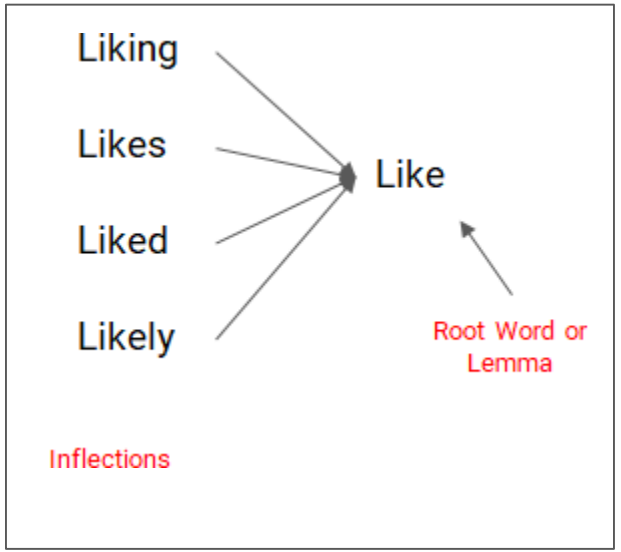
Stemming

- The idea of reducing different forms of a word to a core root.
- Words that are derived from one another can be mapped to a central word or symbol, especially if they have the same core meaning
- Used for dimensionality reduction
- Word stem may not be present in dictionary
- “cook,” “cooking,” and “cooked” all are reduced to same stem of “cook.”



Lemmatization

- Very similar to Stemming
- Converts inflections to root word or Lemma
- Lemmatization involves resolving words to their dictionary form. A lemma of a word is it's dictionary or canonical form



greatlearning
Learning for Life

Thank you!

Happy Learning :)