Natural Language Processing

Introduction

Topics covered

- Introduction to Natural Language Processing
- Pre-processing: Tokenization, Stop words, Normalization, Stemming and Lemmatization
- Bag of Words and Tf-idf features
- Language model
- N-grams

greatlearning
Learning for Life

Session Agenda

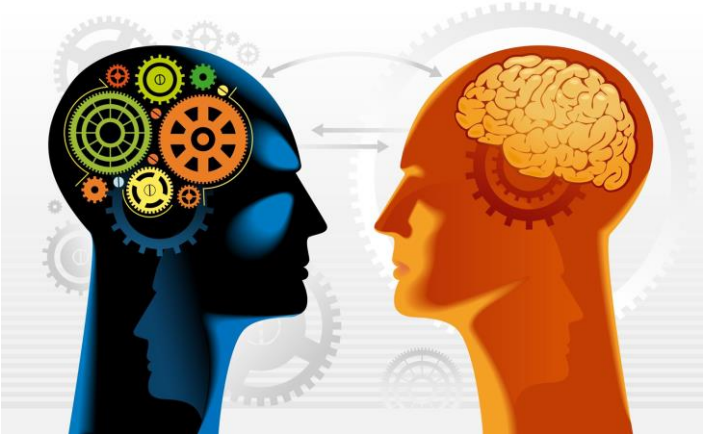
- Basics of NLP
- Pre-processing steps
- Language Models
- Case Study

greatlearning
Learning for Life

Natural Language Processing

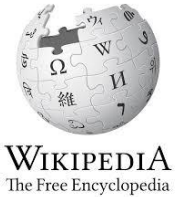
Natural Language Processing

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. 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(chatbots)



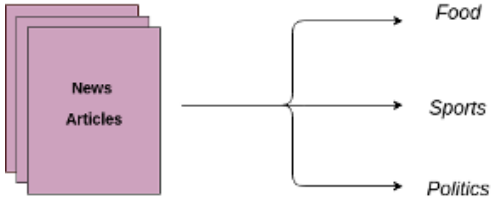
Why study NLP?

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.



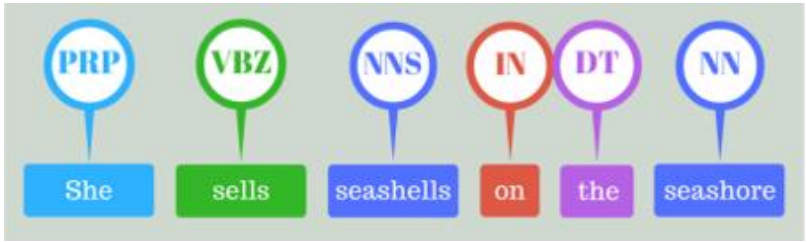
Different Tasks in NLP

- **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

- 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.
  - Tagset: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html



Different Tasks in NLP

- Word Segmentation
  - In some languages, there is no space between words, or a word may contain smaller syllables.n 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?

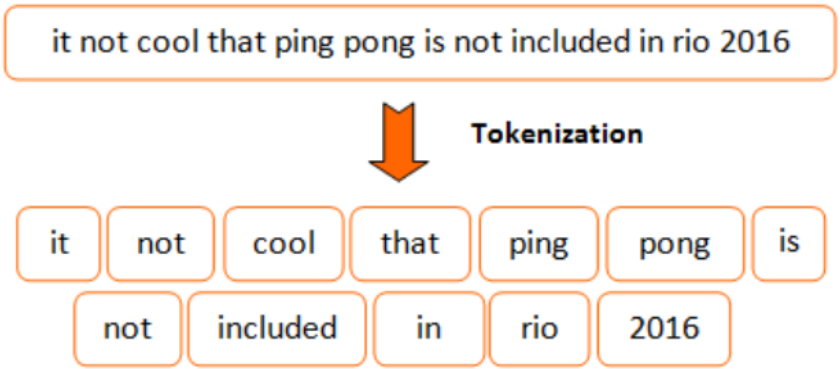
1. Languages are changing everyday, new words, new rules, etc.
2. The number of tokens is not fixed. A natural language can have hundreds of thousands of different words, new words are created on the fly.
3. Words can have different meanings depending on context, and they can acquire new meanings over 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.

greatlearning
Learning for Life

Pre-processing Steps

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. We will use them in our case study.



Stop Words Removal

- Stopwords are common words that carry less important meaning than keywords.
- When using some bag of words based methods, i.e, countVectorizer or tfidf that works on counts and frequency of the words, removing stopwords is great as it lowers the dimensional space.
- 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 and Lemmatization

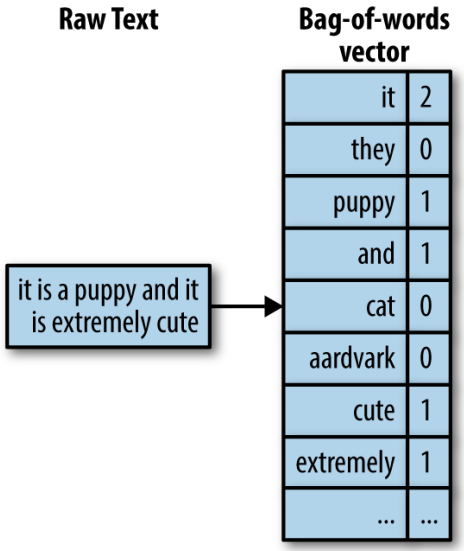
- The idea of reducing different forms of a word to a core root.
- Words that are derived from one another can be mapped to a central word or symbol, especially if they have the same core meaning.
- In stemming, words are reduced to their word stems. A word stem is an equal to or smaller form of the word.
- “cook,” “cooking,” and “cooked” all are reduced to same stem of “cook.”
- Lemmatization involves resolving words to their dictionary form. A lemma of a word is its dictionary or canonical form!

greatlearning
Learning for Life

Word Features

Bag of Words

- In this model, a text (such as a sentence or a document) is represented as the bag of its words, disregarding grammar and even word order but keeping multiplicity.
- We use the tokenized words for each observation and find out the frequency of each token.
- We define the vocabulary of corpus as all the unique words in the corpus above and below some certain threshold of frequency.
- Each sentence or document is defined by a vector of same dimension as vocabulary containing the frequency of each word of the vocabulary in the sentence.
- The bag-of-words model is commonly used in methods of document classification where the (frequency of) occurrence of each word is used as a feature for training a classifier.



Tf-idf Vectors

- Tf-idf (term frequency times inverse document frequency) is a scheme to weight individual tokens.
- One of the advantage of tf-idf is reduce the impact of tokens that occur very frequently, hence offering little to none in terms of information.

$$TFIDF\ score\ for\ term\ i\ in\ document\ j = TF(i,j) * IDF(i)$$

where

$IDF = Inverse\ Document\ Frequency$

$TF = Term\ Frequency$

$$TF(i,j) = \frac{\text{Term i frequency in document j}}{\text{Total words in document j}}$$

$$IDF(i) = \log_2\left(\frac{\text{Total documents}}{\text{documents with term i}}\right)$$

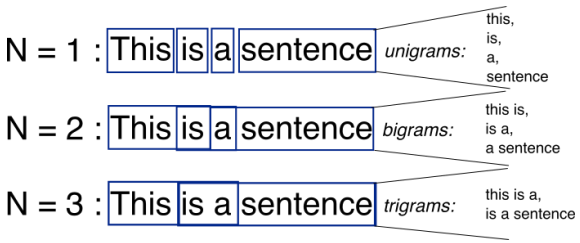
and

$t = Term$

$j = Document$

N-gram and Language Model

- Language models are the type of models that assign probabilities to sequence of words.
- N-grams is the most simplest language model. It's a sequence of N-words.
- Bi-gram is a special case of N-grams where we consider only the sequence of two words (Markovian assumption).
- In N-gram models we calculate the probability of Nth words give the sequence of N-1 words. We do this by calculating the relative frequency of the sequence occurring in the text corpus.



Bigram approximation

$$P(w_1^n) = \prod_{k=1}^{n} P(w_k \mid w_{k-1})$$

N-gram approximation

$$P(w_1^n) = \prod_{k-1}^{n} P(w_k \mid w_{k-N-1}^{k-1})$$

Thank you!

Happy Learning :)