Week 1:Introduction To Natural Language Processing

Topics covered in the week:

- Introduction to Natural Language Processing
- Pre-processing Techniques:
  - HTML tag removal.
  - Tokenization.
  - Stop word removal.
  - Accented characters and special characters removal.
  - Stemming.
  - Lemmatization.

Session Agenda:

- Basics of NLP
- Need of NLP
- Why NLP is hard
- Pre-processing steps
- Case Study

Natural Language Processing:

- Natural Language Processing is a subfield of artificial intelligence concerned with methods of communication between computers and natural languages such as English, Hindi, etc.
- Objective of Natural Language processing is to perform useful tasks involving human languages like:
  - Sentiment Analysis
  - Machine Translation

Why Study NLP?

- Language is involved in most of the activities that involve interaction between humans, e.g. reading, writing, speaking, listening.
- Voice can be used as an interface for interactions between humans and machines e.g. Cortana, google assistant, Siri, Amazon Alexa.
- 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.

Why NLP is hard?

- Languages are changing everyday, new words, new rules, etc.
- 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.
- 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).
- 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.

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.

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

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.

Case Study:

Text Pre-processing of Corporate Messaging Dataset.

Context:

- A data which contains what corporations actually talk about on social media.
- The dataset has statements classified as information (objective statements about the company or it's activities), dialog (replies to users, etc.), or action (messages that ask for votes or ask users to click on links, etc.).
- Our interest is in the text column of dataset, so we can apply pre-processing on it.

Steps:

- Import necessary libraries.
- Get the data.
- Explore the data.
- Do pre-processing:
  - Noise removal (Special character, html tags, accented characters, punctuation removal)
  - Lowercasing (can be task dependent in some cases)
  - Stop-word removal
  - Stemming / lemmatization
- Summary.

Questions?

Thank You.
Happy Learning!