Week 1: Introduction To Natural Language Processing

Topics Covered in the week:

- **Introduction to NLP**
  - Need for NLP
  -New age NLP application

- **Pre-Processing Techniques:**

  - HTML tag removal.
  - Tokenization
  - Stopword Removal
  - Accented Characters and special characters removal
  - Stemming
  - Lemmatization

Session Agenda

- Overview of NLP
- Why do we need NLP
- Why NLP is Complex
- New age NLP Application
- Pre-Processing Step
- Case Study

Overview of Natural Language processing (NLP):

- Unstructured text is the largest human-generated data source, and it grows enormously every day.
- The need to access and process these text data, turns out to be increasingly significant & how it can solve real life/business problems
- Natural language processing holds the potential to improve how we live and work
- New & cutting age NLP application can solve the need to analyze rapidly and vigorously unstructured data.

Natural Language processing

- The term NLP involves the problem area of natural language: as the name suggests, Natural Language Processing refers to the computer processing of natural languages, for whatever purpose, regardless of the processing depth.

- The natural language stands for the languages we use in our daily life, such as English, Russian, Japanese, Chinese. it is synonymous with human language, mainly to be distinguished from formal language, including computer language.

Why do we need NLP ?

- As the amount of data accessible online is expanding day by day, the need to access and process this data, turns out to be increasingly significant & how it can solve real life/business problems.

- The applications of NLP are incredibly diverse and ideal for nearly any situation involving the need to rapidly and vigorously analyze unstructured text.

Why NLP is complex ?

- **Dealing with irony and sarcasm i**n a language:-words and phrases that may be positive or negative according to dictionary, but actually signify the opposite meaning,

- **Domain specific language:-** Different businesses and industries often use very different language for their activities and operations . An NLP processing model needed for healthcare, for example, would be very different than one used to process legal documents.

Why NLP is complex ?

- **Ambiguity** refers to sentences and phrases that are likely to have two or more possible interpretations.
- Languages are changing everyday, new words, new rules, etc.
- 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.

NLP: Numerous Tasks & their objective

TASK	OBJECTIVE
Text Classification	Prediction of tags, Categorization, Sentiment analysis
Word sequence	Language modeling - predict next/prior word(s), text generation
Text Narrative	Representation of meaning of word/sentence/document

New age NLP application

- **Chatbot and virtual assistant:** Recommending a product to getting feedback from the customers & assisting customer for their queries,chatbots does all these task.

- **Machine translation:** automatically converting the text in one language to another language while keeping the meaning intact

- **Auto correct :-** Search engine & email auto correct and auto complete of words & sentences.

New age NLP application

- **Speech recognition & Voice assistant**: The ability of devices to respond to spoken commands. Google Assistant, Apple Siri, Amazon Alexa, ring a bell? Yes, all of these are voice assistants.

- **Sentiment analysis**: The process of detecting positive or negative sentiment in text. It's often used by businesses to detect sentiment in social data, gauge brand reputation, and understand customers sentiments.

- **Text Classification:** It is used to organize, structure, and categorize any kind of text – from documents, medical studies and files,sentiment and all over the web.

Why text pre processing?

- Text preprocessing is an approach for cleaning and preparing text data for use in a certain context.
- To preprocess our text simply means to put forward our text into a form that is predictable and analyzable for our task.
- The ultimate goal of cleaning and preparing text data is to reduce the text to only the words that you need for our NLP goals.

Text pre - processing steps:

- **Removal of HTML Tags and special character:** HTML tags are typically one of these section which don't add much value towards understanding and analyzing text. We should remove it before performing NLP Tasks.

- Beautifulsoup library from BS4, Regex, tools & unicode module are extensively used to find and replace/remove anomaly from text.

- strip_html_tags("<html><h2>Some important Rule</h2></html>")
  → "Some important Rule"

Text pre - processing steps:

- **Removal of special characte**r : Special characters and symbols are usually non-alphanumeric characters which adds to the extra noise in unstructured text.

- Usually, simple regular expressions (regex) is used to remove them.

- remove_special_characters("Well this was fun! 123#@!", remove_digits=True)

  → “Well this was fun!”

Removal of Stopwords

- 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

    - Example: remove_stopwords("a, an, the, and, are, if are stopwords,)
        - Removing stopwords minimizes computation
        - Removing stopwords could increase classification accuracy

Tokenization

- Tokenization is a method of separating a piece of text into smaller units called tokens. Here, tokens can be either words, characters, or subwords.

- Why Tokenization: As tokens are the unit blocks of Natural Language, the most common way of processing the raw text happens at the token level.

- The most common way of constructing tokens is based on space. The Tokenization of the sentence based on space as delimiter.

- Example :- “Consistency is the key to success”
  As “Consistency”-”is”-”the”-”key”-”to”- “success”.

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.
- 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.
- print(SnowballStemmer("English").stem("lovely"))
  Output: love

Lemmatization

- Lemmatization involves resolving words to theirdictionary form. A lemma of a word is its dictionary or canonical form.

- Lemmatization returns an actual word of the language, it is used where it is vital to get valid words.

- walking ---> walk (core-word extraction)
- was ---> be (tense conversion to present tense)
- mice ---> mouse (plural to singular

Word	Stem	Lemma
Studies	Studi	Study

Questions ?

Thank You.
Happy Learning!