Natural Language Processing (NLP)

- technology used to help computers understand, interpret, and generate human language in a way that is meaningful and useful for computers.

ex) <Unstructured data> → NLP (NLU / NLG) → <structured data>

Add Eggs and Milk to the Shopping List.

```
<Shopping List>
  <Item> <Egg>
  <Item> <Milk>
</>
```

- NLU: Natural Language Understanding
- NLG: Natural Language Generation.

NLP usecases:

1.) Machine Translation
2.) Virtual Assistance / ChatBot
3.) Sentiment Analysis
4.) Spam Detection

How does it work?

<Unstructured data> Input → ① Tokenization → ② Stemming / Lemmatization → ④ Part of Speech Tagging → ⑤ Named Entity Recognition → Vectorization. → <structured data>

① Tokenization: cuts the sentences into individual pieces.

ex.) "I love robots" → ["I", "Love", "Robots"]

② Stemming: Finds the stem (root of the word)

ex.) "running", "runs", "ran" → "run" for easier recognition

③ Lemmatization: Finds the Lemma: dictionary, base of a word.

ex.) "Universe", "University" each define completely different words.
∴ Find the dictionary meaning to each word.

④ Part of Speech Tagging: Finds where the token is used within the context in the sentence.

ex.) "make" → "I am going to make dinner"
↑
Used as verb.

→ "what make is your laptop?"
↑
used as noun.

∴ P.S.T: helps us find the meaning used in the content.

⑤ Named entity recognition (N.E.R): For a given token, is there a entity associated with it?

ex.) "Arizona" ← U.S. state

"Rawlph" ← Person's name.

Vectorization

- turn each word into a list of numbers (a vector)

ex.) "Apple", "Pear", "Celery"

Fruit	Sweetness	Crunchiness	Vector (coordinates)
Apple	0.8	0.9	[0.8, 0.9]
Pear	0.7	0.6	[0.7, 0.6]
Celery	0.1	0.9	[0.1, 0.9]

} Computer sees the different levels of sweetness, Crunchiness related and "sees" that they are different.

LLM

LLM: Large Language Model.

- type of AI trained on a massive data scale to understand, generate, and predict human language.