



# Natural Language Processing

## 1.Natural Language Processing

- Overview
  - Understanding and replying the human language
  - NLP is a part of Computer science and Artificial Intelligence which deals Human language
- Branch
  - Natural Language Understanding
    - (NLU) is a branch of artificial intelligence (AI) that uses computer software to understand input made in the form of sentences in text or speech format.
  - Natural Language Generation
    - Natural-language generation is a software process that transforms structured data into natural language. It can be used to produce long form content for organizations to automate custom reports, as well as produce custom content for a web or mobile

## 2. NLP Terminology

- 1
  - Tokenization
    - Breaking strings into small individual Tokens or words

Example

What's the time now?  
to  
What's  
The  
Time  
now- 2
  - Stemming
    - Normalize words into its base form or its root form

Example

Playing , played, plays  
to  
play
- 3
  - Lemmatization
    - Grouping different inflected form of word called Lemma
    - Similar to Stemming, but returns perfect word

Example

Better, super  
to  
Good
- 4
  - POS – parts of speech
    - The most popular POS tagging would be identifying words as nouns, verbs, adjectives, etc.

Example

Google about pantech solutions  
  
Google may be Noun/Verb
- 5
  - Named entity recognition
    - Recognizing the words as movie, monetary values, organization, location, quantity or person.

Example

Google about pantech solutions  
  
Google – verb  
Pantech solutions - organization
- 6
  - Chunking
    - Picking pieces of words and form into phrases

Example

Google  
About  
Pantech  
Solutions  
to  
Google about pantech solutions

## 3. Natural Language Toolkit - NLTK

- This toolkit is one of the most powerful NLP libraries which contains packages to make machines understand human language and reply to it with an appropriate response.
- Tokenization, Stemming, Lemmatization, Punctuation, Character count, word count are some of these packages which will be discussed in this tutorial.

## 4. Install NLTK

```
pip install nltk

Successfully installed nltk-3.5
Installing collected packages: nltk
Successfully built nltk
Successfully installed nltk-3.5
```

## 5.Feature extraction in Text

- CountVectorizer
  - Convert a collection of text documents to a matrix of token counts
- HashingVectorizer
  - Convert a collection of text documents to a matrix of token occurrences
- TfidfVectorizer
  - Overview
    - Term-frequency times inverse document-frequency
    - Convert a collection of raw documents to a matrix of TF-IDF features
    - Equivalent to CountVectorizer followed by TfidfTransformer
  - Example
    - TfidfVectorizer.get\_feature\_names
    - DATA= ['This is the first document.', 'This document is the second document.', 'And this is the third one.', 'Is this the first document?']
    - X = vectorizer.fit\_transform(DATA)
    - TfidfVectorizer.get\_feature\_names
    - ['and', 'document', 'first', 'is', 'one', 'second', 'the', 'third', 'this']