# 0. 시작하기에 앞서

## 가이드라인에 대한 개요

이 학습 가이드라인은 목적에 따라 두 가지 방향으로 작성하고자 함.

### 공통사항

- 자연어 처리는 Text와 음성을 아우르는 개념이나, 이 가이드라인에서는 오직 Text만을 다룸.
- 기본적인 Deep learning의 개념과 방법론을 다룸.
- Deep learning을 적용하기 위한 NLP의 기초적인 지식에 대해서 전반적으로 살펴봄.
- NLP에서 풀고자 하는 다양한 Task를 접하고, 이에 대한 다양한 해결 방법을 실습해봄.
- 실습 및 Code Python, Pytorch, HuggingFace를 기반으로 진행됨.

### “자연어 처리”에 중점을 둔 방향

- 최근 Deep learning의 위상이 매우 높으나, NLP의 문제를 풀기 위한 방법론 중 하나임.
- Deep learning는 풀기 어려우나, 통계적인 방법론으로는 쉽게 해결할 수 있는 NLP Task가 존재함.
- 통계적인 방식, Ranking algorithm 등 다양한 방법론을 살펴봄.
- 업무에서 자연어 관련 research나 ideation에 도움이 될 것.
- 통계적 베이스를 기반으로 하고, 재미가 다소 없을 수 있음.
- refs.
    - https://wikidocs.net/book/2155
    - http://web.stanford.edu/class/cs224n/
    - http://blp.korea.ac.kr/?page_id=3577
- 핵심 내용

    - Statistical Language Model
    - Document-Term Matrix
    - TF-IDF, BM25, Text Similarity
    - …

## “머신러닝”에 중점을 둔 방향

- 전통적인 ML부터, Deep learning으로 대표되는 Nueral Network 까지의 역사를 살펴봄.
- 기본적인 분류나 회귀, CV 등에 대한 내용도 같이 살펴봄.
- 수학적 베이스를 기반으로 하고, 학습 원리 등에 대해 자세히 살펴봄.
- NLP에 사용되는 모델의 자세한 구조와 원리에 대해 자세히 살펴봄.
- 나름 재미가 있으리라 생각하지만, 업무적 성격과는 다소 동떨어질 수 있음.
- NLP의 전통적인 방법론 등에 대해서는 내용이 다소 적을 수 있음.
- refs.
    - https://wikidocs.net/book/2788
    - https://www.inflearn.com/course/기본적인-머신러닝-딥러닝-강좌/
    - https://www.inflearn.com/course/핸즈온-머신러닝
    - http://www.yes24.com/product/goods/89959711
    - http://www.yes24.com/Product/Goods/115633781
- 핵심 내용
    - Regression, Classification
    - Perceptron
    - CNN, RNN
    - …

## 앞으로의 방향

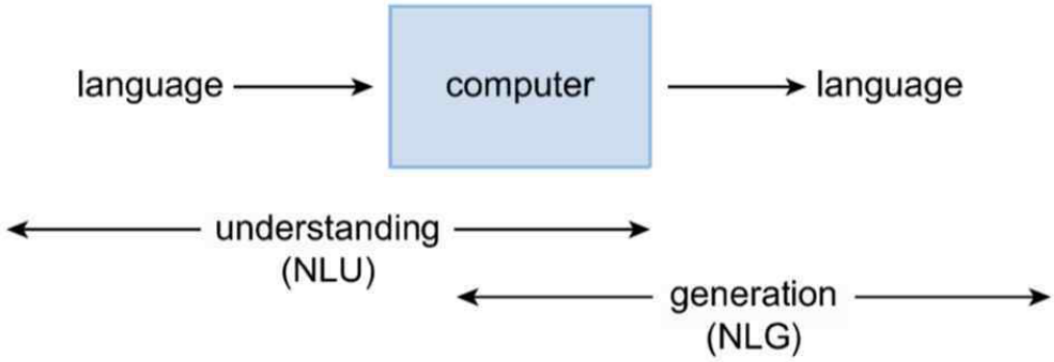
- 기대하는 효과

    - 자연어 처리에 대한 이해를 기반으로 업무 및 프로젝트에서 다양한 아이디어를 제시할 수 있음.
- 학습 목표
    - 전통적인(w/o DL) 방식의 NLP 접근 방식을 이해한다.
        - 형태소 분석 기반, 통계 기반 모델, 랭킹 알고리즘 등
    - ML의 기본적인 아이디어와 학습 방법을 이해한다.
        - Regression, Classification
        - Logit, Loss function, Optimizer, Loss, GD, Parameter
    - NLP를 위한 DL의 발전 과정을 이해한다.
        - Perceptron, RNN, Encoder Decoder, Auto Encoder, Seq2Seq, Attention, Transformer, BERT, GPT

## NLP란?

자연어(특히, Text)라는 비정형 Data를 원하는 목적에 따라 어떻게 다루고 처리하는 지에 대한 분야임.

자연어를 처리하는 단계를 간단하게 도식화 하면 다음과 같음.




자연어 처리에서 대체로 풀고자 하는 문제는 다음과 같음.

- **Tokenization** : 일정한 규칙에 따라 문장을 단어로 Parsing 하는 작업
- **Stemming** : 단어의 의미와 어근을 구별하는 작업
  *(ex. '맑다', '맑은', '맑고' 등)*

- **Named Entity Recognition (NER)** : 고유명사를 인식하는 작업
  *(ex. 'The New York Times' 등)*
- **Part-of-Speech (POS) Tagging** : 단어의 품사나 성분을 인식하는 작업
- **Sentiment Analysis** : 문장이나 글의 감정을 분석하는 작업
- **Machine Translation** : 기계를 통해 문장이나 글을 번역하는 작업
- **Entailment Prediction** : 문장 간 논리 구조를 예측하는 등의 작업
- **Question Answering** : 질문을 이해하고 관련된 정보를 제공하는 작업
  *(ex. '나폴레옹이 죽은날은?' 등)*
- **Dialog Systems** : 챗봇과 같이 대화를 이해하고 생성하는 작업
- **Summarization** : 글이나 다수의 문장을 핵심적인 내용으로 요약하는 작업
- **Extract Useful Information** : 웹과 같은 방대한 정보 속에서 유의미한 정보를 추출하는 작업
  *(ex.  뉴스 분석, 소비자 반응 분석 등)*
- **Document Clustering** : 글을 일정한 Class로 분류하거나 비슷한 글끼리 묶는 작업
- **Recommendation Systems** : 사용자의 정보를 통해 선제적으로 정보를 제공하는 작업
  *(ex. Youtube 영상 추천 기능 등)*

## Trends of NLP

현재 ML은 **Computer Vision** 분야와 **NLP** 분야가 활발히 연구되고 있다. CV는 CNN과 GAN, Defusion 등으로 인해 폭발적인 성장을 이루고 있다. NLP는 다음과 같은 Trend로 발전해나가고 있다.

자연어는 그 자체로는 기계가 이해할 수 없고, 단어의 순서에 따라 의미가 바뀌는 특징을 가진다. 이에 따라, **Embedding**과  **Sequential Data**의 처리가 주된 관건이었다.

최초에 자연어처리는 다양한 규칙을 기반으로 Data를 정제하고 의미를 추출하고자 노력했고, 이후에는 통계적인 방법을 활용해 자연어의 특징을 구축하고자 했다.

Deep learning이 발전하게 되면서, **RNN (LSTM, GRU)**과 같은 모델이 등장했고, Sequential Data를 처리할 수 있게 되었다. 더 나아가 **Transformer** 모델의 등장으로 큰 발전을 이루어 냈다. Transformer이전에는 각 Task에 특화된 모델이 따로 존재했었으나, 근래에는 Transformer의 구조를 이용한 모델이 이를 대체하고 있다.

**Self Attention**구조와 **Self Learning**을 통해 수 많은 데이터를 이미 학습한 **Pre-trained** 모델도 등장했다. **Bert**와 **GPT-X** 모델은 필요한 Task의 데이터를 전이학습하여 어디서든 좋은 성능을 낼 수 있도록 개발되었다.