

# 텍스트 마이닝과 데이터 마이닝

# Part 06. 자연어 처리와 실습

정 정 민

# Chapter 15. NLP와 프레임워크

---

1. NLP, 자연어 처리
2. 주요 프레임워크

# NLP, 자연어 처리

## [RECAP] 자연어 처리 (Natural Language Processing, NLP)

---

- 자연어 처리란,
  - 컴퓨터가 **인간의 언어를 이해하고 해석**하는데 사용되는 분야로
  - 컴퓨터 과학, 인공 지능, 언어학의 개념이 사용됨
- NLP의 목적은 인간 언어의 구조와 의미 이해를 바탕으로
  - **글을 활용한 문제를 해결**하고
  - **향상된 사용자 경험**을 제공하고자 함
    - chatGPT와 같은 사용 경험이 해당하겠죠?
- 대규모 텍스트 데이터 내의 존재하는 패턴, 관계, 정보를 발견하고 분석하는 텍스트 마이닝과 거리가 있음
- 두 개념의 목표 차이는
  - NLP : 언어의 이해
  - TM : 언어 속 내포된 정보 파악

# 자연어 처리의 다양한 문제

---

## 텍스트 이해 (Text Understanding)

- 질의응답 (QA, Question Answering)
- 문장 이해 (Reading Comprehension)
- 정보 검색 (Information Retrieval)

## 텍스트 생성 (Text Generation)

- 문장 생성 (Text Generation)
- 요약 (Text Summarization)
- 번역 (Neural Machine Translation)

## 텍스트 분류 및 태깅 (Text Classification & Tagging)

- 문장 분류 (Text Classification)
- 개체명 인식 (NER, Named Entity Recognition)
- 품사 태깅 (POS tagging, Part of Speech tagging)

## 텍스트 관계 추출 (Text Relation Extraction)

- 문장 관계 추출 (Relation Extraction)

# 주요 프레임워크

# Natural Language Tool Kit (NLTK)

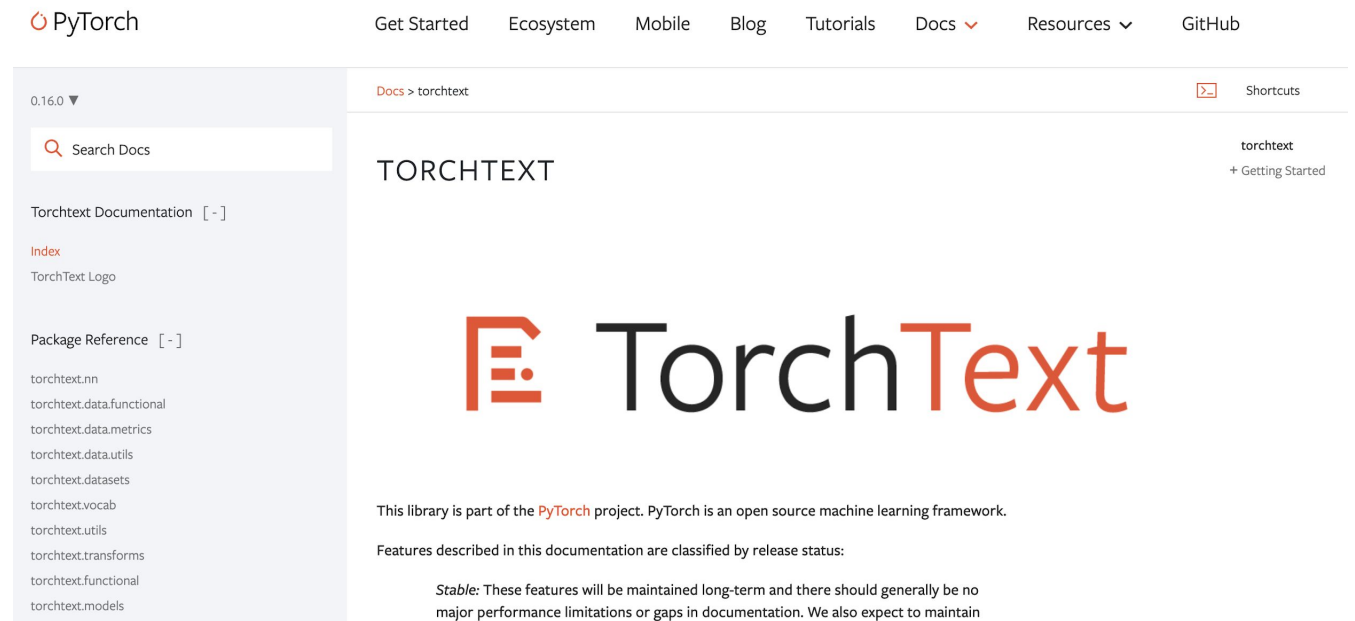
- 전통적인 NLP 기법을 구현한 패키지 모음
- 전처리, 딥러닝 이전의 NLP 방법들이 많이 활용됨
- 공식 페이지 : [링크](#)
- 다양한 텍스트 데이터 제공 : [링크](#)

NLTK	Documentation
Search	Natural Language Toolkit
<div>NLTK Documentation</div> <div>API Reference</div> <div>Example Usage</div> <div>Module Index</div> <div>Wiki</div> <div>FAQ</div> <div>Open Issues</div> <div>NLTK on GitHub</div> <div>Installation</div> <div>Installing NLTK</div> <div>Installing NLTK Data</div> <div>More</div> <div>Release Notes</div> <div>Contributing to NLTK</div> <div>NLTK Team</div>	<p>NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to <b>over 50 corpora and lexical resources</b> such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active <b>discussion forum</b>.</p> <p>Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.</p> <p>NLTK has been called “a wonderful tool for teaching, and working in, computational linguistics using Python,” and “an amazing library to play with natural language.”</p> <p><b>Natural Language Processing with Python</b> provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more. The online version of the book has been updated for Python 3 and NLTK 3. (The original Python 2 version is still available at <a href="https://www.nltk.org/book_1ed">https://www.nltk.org/book_1ed</a>.)</p> <p>Some simple things you can do with NLTK</p>



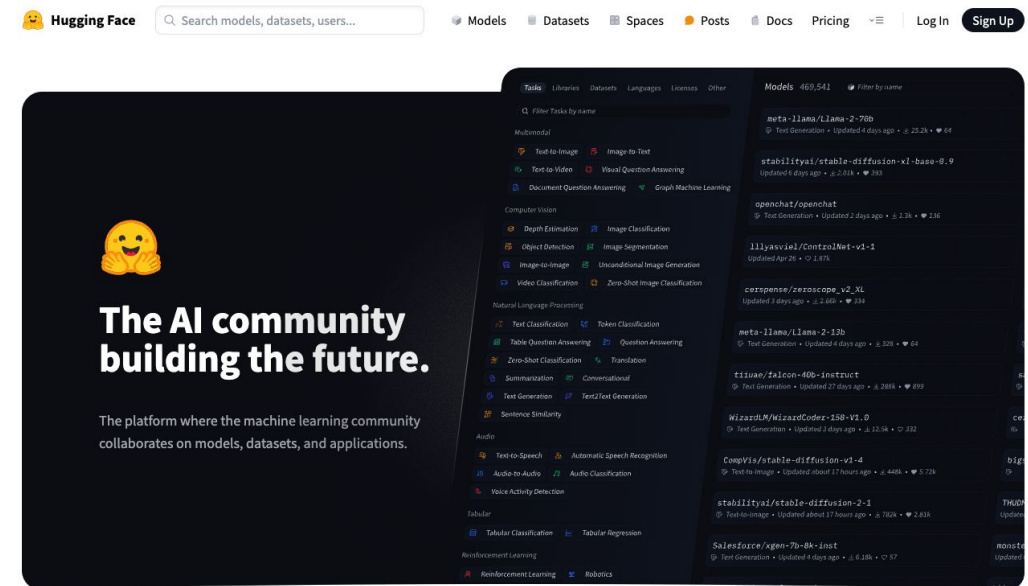
# PyTorch, TorchText

- Facebook에서 개발한 머신 러닝 오픈 소스 라이브러리
- 특히, 딥러닝에 특화
- TorchText는 PyTorch에서 제공하는 NLP에 특화된 내부 라이브러리
- (초기 ~ 최신의) 딥러닝 모델을 쉽게 구현할 수 있는 인터페이스 제공
- 또한, 데이터 전/후처리와 모델 학습에 필요한 여러 요소를 제공
- 공식 페이지 : [링크](#)

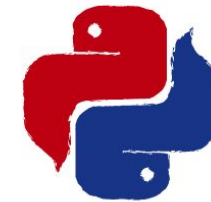


# HuggingFace

- 자연어 처리에 특화된 커뮤니티 기반의 라이브러리
- NLP를 비롯해 다양한 연구 분야(이미지, 음성, 생성 등)의 연구 성과를 공유 & 활용 가능
- 손쉬운 사용과 학습을 위한 유용한 기능을 내포
- 또한, 데모 사이트(Spaces) 사용, 데이터 업로드 & 다운로드(Datasets)가 가능
- 커뮤니티 기반의 플랫폼이라
- 매우 다양한 연구 결과물(학습 결과 모델, 데이터, 데모 등)이 빠르게 업데이트
  - 연구 결과물을 공통된 인터페이스로 강제
  - 코드 진행의 통일성 제공
- 공식 페이지 : [링크](#)



- 자연어 처리(NLP) 중 한글 데이터 처리에 특화된 파이썬 라이브러리
- 한국어에 특화된 전처리 기법을 많이 갖고 있음
  - 형태소 분석
    - 학교에 갑니다 > 학교 / 에 / 가/ㅂ니다
  - 품사 태깅 및 추출
    - 한국어 단어에 특화된 품사 예측
- 공식 페이지 : [링크](#)



## KoNLPy

KoNLPy is a Python package for Korean natural language processing.

### Table of Contents

KoNLPy: Korean NLP in Python

- [Standing on the shoulders of giants](#)
- [License](#)
- [Contribute](#)
- [Getting started](#)
- [User guide](#)
- [API](#)
- [Indices and tables](#)

### Translations

English  
[한국어](#)

### Quick search

## KoNLPy: Korean NLP in Python

Build status [docs](#) [passing](#)

KoNLPy (pronounced "ko en el PIE") is a Python package for natural language processing (NLP) of the Korean language. For installation directions, see [here](#).

For users new to NLP, go to [Getting started](#). For step-by-step instructions, follow the [User guide](#). For specific descriptions of each module, go see the [API](#) documents.

```
>>> from konlpy.tag import Kkma
>>> from konlpy.utils import pprint
>>> kkma = Kkma()
>>> pprint(kkma.sentences(u'네, 안녕하세요. 반갑습니다.'))
[네, 안녕하세요.,
반갑습니다.]
>>> pprint(kkma.nouns(u'질문이나 건의사항은 깃헙 이슈 트래커에 남겨주세요.'))
[질문,
건의,
건의사항,
사항,
깃헙,
이슈,
트래커]
>>> pprint(kkma.pos(u'오류보고는 실행환경, 에러메세지와함께 설명을 최대한상세히!^^'))
[(오류, NNG),
(보고, NNG),
(는, JX),
(실행, NNG),
(환경, NNG),
(,, SP),
(에러, NNG),
(메세지, NNG),
(와, JKM),
(함께, MAG),
(설명, NNG),
(을, JK0),
(최대한, NNG),
(상세히, MAG),
(!, SF),
(^^, EM0)]
```

## 물론 텍스트 마이닝 과정의 패키지도!

---

- 앞선 파트에서 살펴본 텍스트 마이닝 과정의 패키지도 많이 사용
- 특히,
  - 전처리 과정에서 Gensim
  - (딥러닝이 아닌) 머신러닝 모델 학습에서 Sklearn
- 패키지들을 많이 사용!

**E.O.D**