데이터 제작의 A to Z

한지윤
Upstage

1. OT

2. 데이터 제작의 중요성

3. 데이터 구축 과정과 설계 기초

4. 자연어처리 데이터

강연자 소개

한지윤

어떻게 하면 실제 세계를 더 정교하게 표상하는 데이터를 만들 수 있을까 궁리하는 연구자입니다.
자연어 이해 벤치마크인 KLUE의 데이터 제작을 담당했습니다.

현 Upstage 데이터 매니저
전 연세대학교 언어정보연구원

연세대학교
언어정보학 협동과정
전산언어학 박사

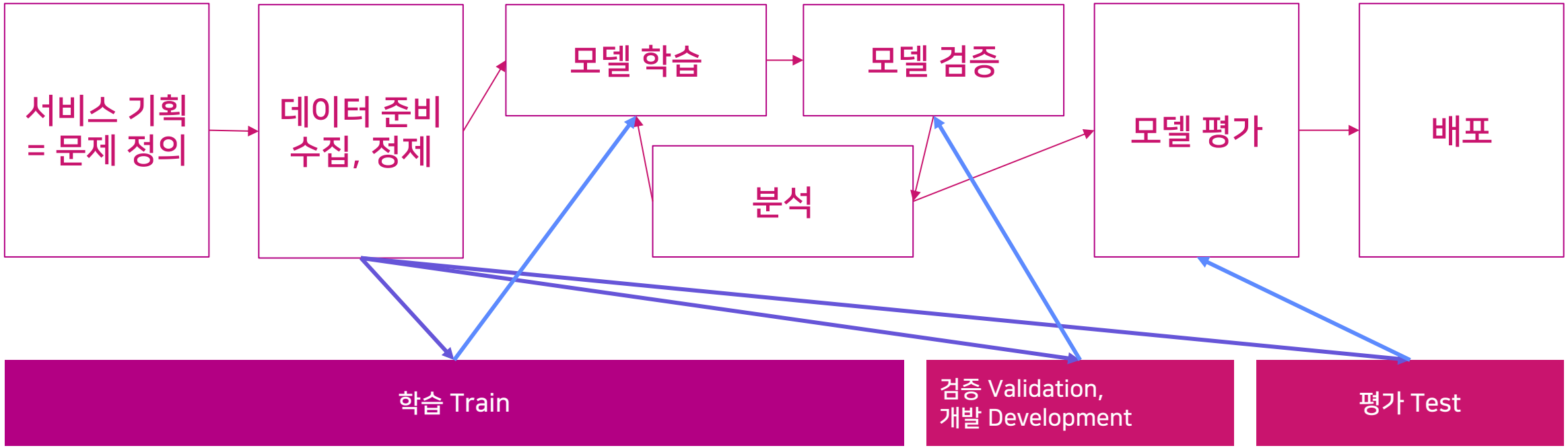
전체 강의 소개 : 강의 목적

인공지능 서비스 개발을 위한 데이터 제작 과정을 이해한다.
자연어처리 과제(Task)별 데이터의 특성을 탐구한다.
실습을 통해 실제 데이터 구축 과정을 체험한다.

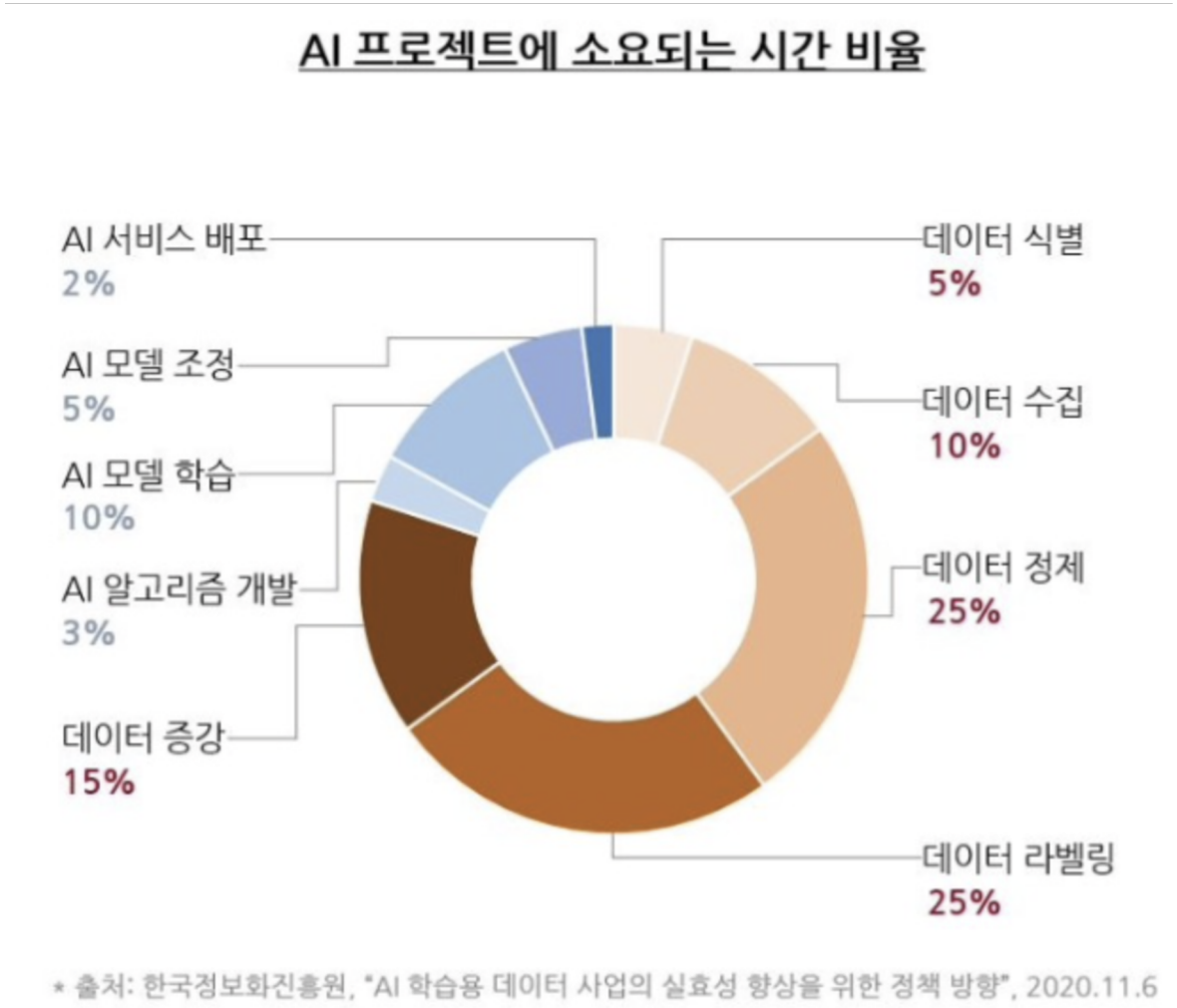
전체 강의 소개 : 강의 목차

데이터 제작의 P;ㅠ

인공지능 서비스 개발 과정과 데이터



데이터 제작의 P;ㅠ



* 출처: 한국정보화진흥원, "AI 학습용 데이터 사업의 실효성 향상을 위한 정책 방향", 2020.11.6

80

전체 프로젝트에서 데이터 관련 작업에 소요되는 시간 비율

데이터 구축 과정

원시 데이터 선정 및 확보

저작권, 태스크 적합성

구축 및 가공 프로세스 확립

구축 및 검수 절차,
작업자 선정

구축 및 가공 지침 작성

플랫폼 소통 및 작업자 교육

데이터 구축 및 가공

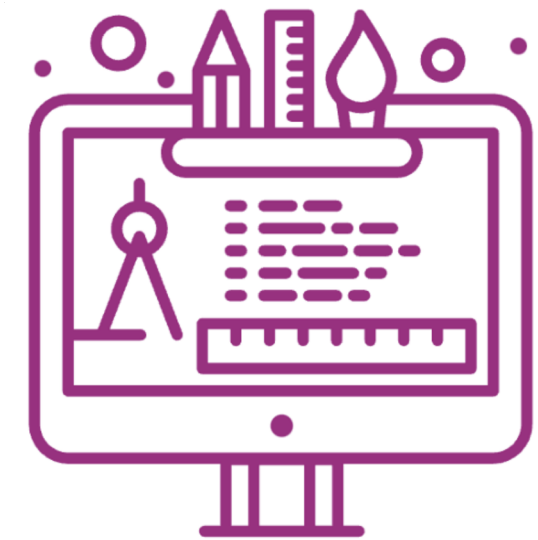
파일럿, 작업자 관리

데이터 검수

품질 평가 기준
데이터 규격, 내용

AI 데이터 설계의 구성 요소

데이터 설계



데이터의 형식
데이터 표상 영역

데이터 수집-가공 설계

원천 데이터 수집 방식:
전산화, 스크래핑, 작업자 작성, 모델 생성

주석 작업 :
전문가 구축, 크라우드 소싱

데이터 설계

데이터의 유형
데이터의 In/Out 형식
데이터(train/dev(validation)/test)별 규모와 구분(split) 방식
데이터의 주석(annotation) 유형

데이터 설계

데이터의 유형

소리

신호처리,
음성인식 등

텍스트

자연어처리

이미지

컴퓨터 비전

영상

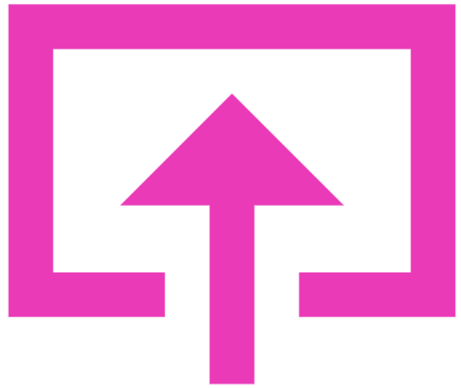
영상처리

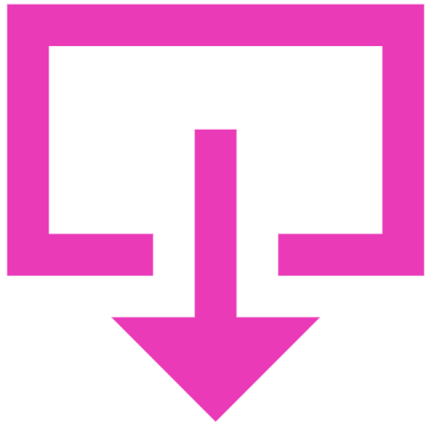
\+ 멀티모달

데이터 설계

데이터의 Input / Output 형식

HTML, XML, CSV, TSV, TXT, JSON, JSONL
JPG, Jpeg, PDF, png, ocr
.wav .mp3 .pcm .script





데이터 설계

데이터(train/dev(validation)/test)별 규모와 구분(split) 방식

규모 선정에 필요한 정보 : 확보 가능한 원시데이터의 규모, 주석 작업 시간

구분 방식 : 데이터별 비율과 기준 정하기
랜덤 vs 특정 조건

데이터 설계

데이터 주석 유형 : 자연어처리

주석 유형(Annotation Type)	주요 활용 용도
클래스 라벨(단일, 다중)	텍스트 분류(Text Classification) *감성, 주제 등
단어(구문) 라벨	명명된 개체명(Entity, 용어 또는 단어) 인식 (Named Entity Recognition)
텍스트 라벨	문장 번역 문장 요약
단어(구문) 라벨링 및 두 단어 사이의 관계	관계-의존성 정의(Relation-Dependencies)
기타	그 밖의 용도

출처: 인공지능 학습용 데이터셋 구축 안내서 https://bit.ly/2Y5e4R5

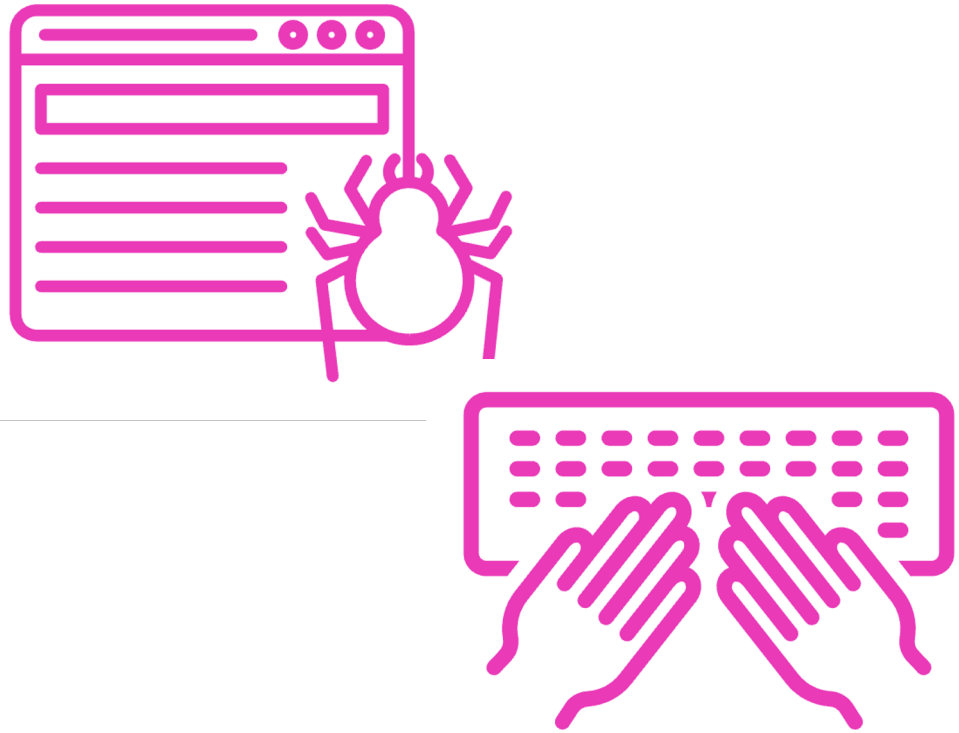
데이터 수집-가공 설계

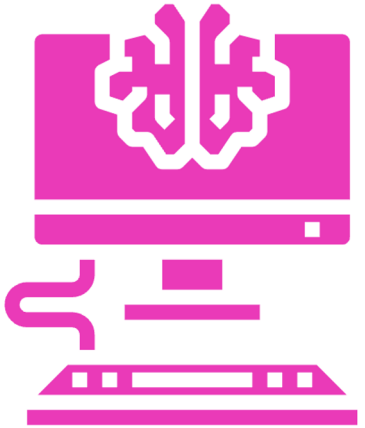
원시 데이터 선정
작업자 선정
구축 및 검수 방법 설계
가이드라인 작성

데이터 수집-가공 설계

원시 데이터 수집 방식

전산화, 스크래핑, 작업자 작성, 모델 생성 : 적합한 데이터란 무엇인지 기준 세우기





데이터 수집-가공 설계

작업자 선정

주석 작업의 난이도와 구축 규모에 맞는 작업자 선정 및 작업 관리

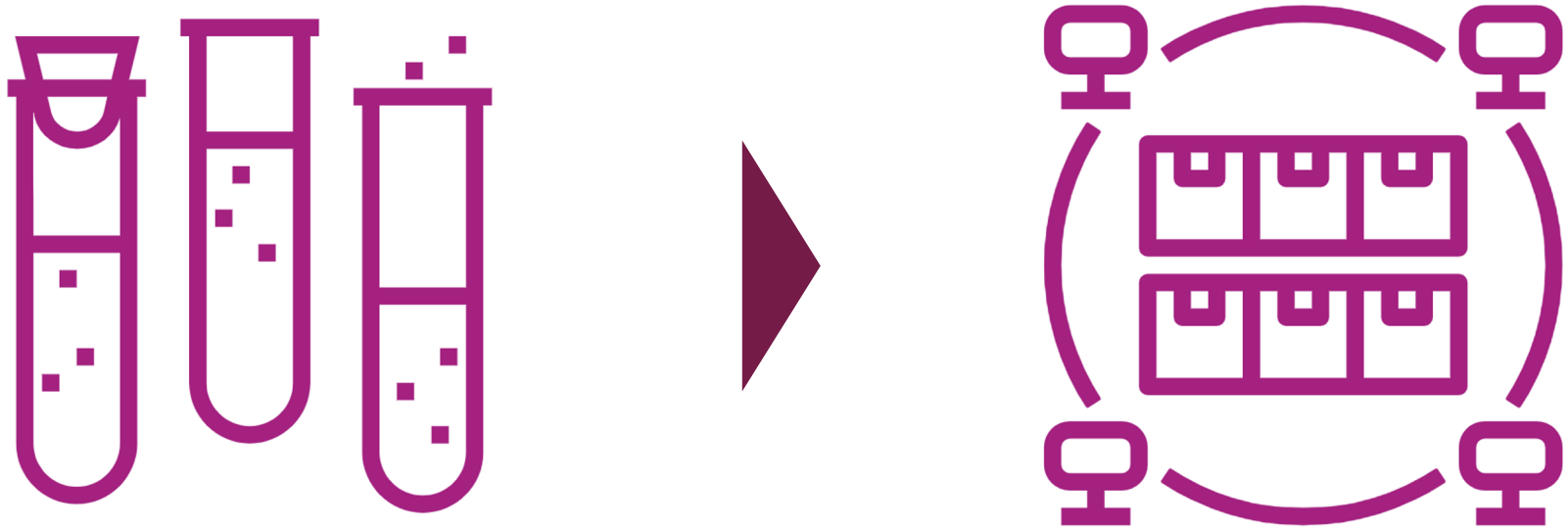
전문가

크라우드 소싱

데이터 수집-가공 설계

구축 및 검수 설계

구축 작업의 난이도와 구축 규모, 태스크 특성에 맞는 구축 및 검수 방식(전문가, IAA) 설계



파일럿 본 구축

데이터 수집-가공 설계

데이터 구축 및 가공

파일럿

설계 시 발견하지 못한 이슈 발굴 및 해결

가이드라인 보완 및 개정

작업자 선정

본 구축

작업 일정 관리

작업자 관리

중간 검수를 통한 데이터 품질 관리

데이터 수집-가공 설계

데이터 검수 및 분석

평가 지표 설정

전문가 평가 및 분석

샘플링 검사

가이드라인 적합도 분석

자동 평가 및 분석

데이터 형식

레이블별 분포 파악

일괄 수정 사항 반영

자연어처리 데이터

자연어란?

자연어(Natural Language) 일상적으로 사용하고 있는 언어 그 자체	인공어 (Artificial Language) 여러 사람의 목적이나 의도에 따라 만든 언어 또는 컴퓨터 언어
한국어 영어 일본어 중국어 이탈리아어 . . .	에스페란토어 or 파이썬 C 언어 자바스크립트 . .

자연어처리 데이터

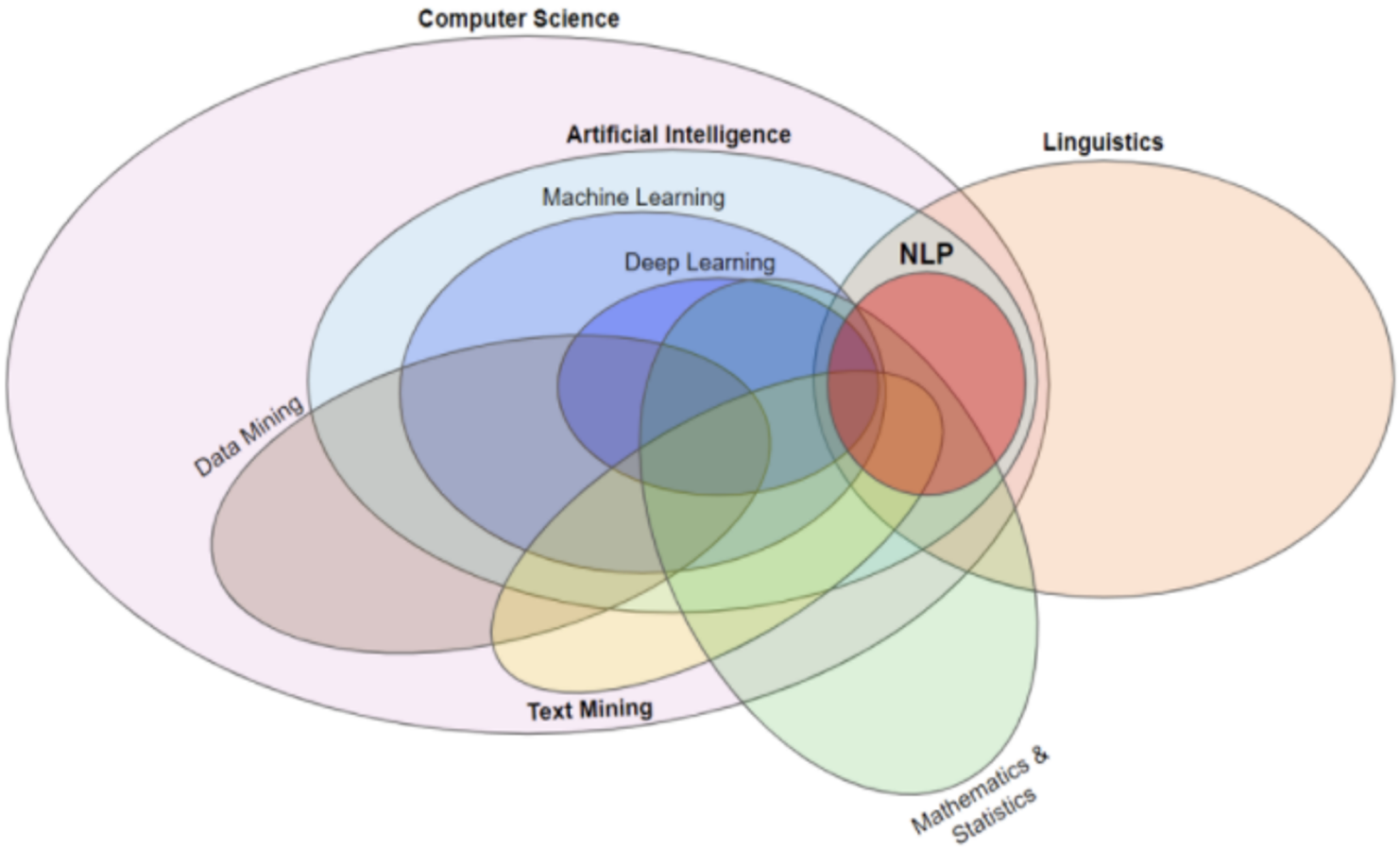
자연어처리(NLP, natural language processing) 란?

인공지능의 한분야, 사람의 언어를 컴퓨터가 알아듣도록 처리하는 인터페이스 역할.
자연어 이해(NLU, natural language understanding)와 자연어 생성(NLG,natural language generation)으로 구성

자연어 처리의 최종 목표 :
컴퓨터가 사람의 언어를 이해하고 여러 가지 문제를 수행할 수 있도록 하는 것

자연어처리 데이터

자연어처리와 관련 연구 분야



https://gritmind.blog/2020/10/09/nlp_overview/

Linguistics

- 음운론(Phonetics and Phonology)
  - Speech recognition/segmentation/synthesis
- 형태론(Morphology)
  - Morphological segmentation
  - Part-of-speech tagging
  - Lemmatization/Stemming
- 통사론(Syntax)
  - Parsing
- 의미론(Semantics)
  - Lexical semantics
    - Distributional semantics
    - Word sense disambiguation
    - Named entity recognition (NER)
    - Sentiment analysis
    - Terminology extraction
  - Relational semantics
    - Relationship extraction
    - Semantic Role Labelling
- 담화론(Discourse)
  - Coreference resolution
  - Anaphora resolution
  - Textual entailment
  - Topic segmentation

AI

- Automatic Summarization
- Paraphrasing
- Natural Language Generation
- Dialogue System
  - (=Conversational Agent)
- Machine Translation
- Question Answering

Text Mining

- Information Retrieval (IR)
  - Search and Ranking
  - Filtering (Recommendation)
  - Categorization
  - Summarization
- Information Extraction (IE)
  - Entity extraction
  - Relationship extraction
  - Event extraction
  - Link analysis
  - Coreference resolution
- Clustering
- Topic Analysis
- Visualization

https://gritmind.blog/2020/10/09/nlp_overview/

자연어처리 데이터

데이터 분류 방식

원천 데이터 장르(도메인) : 문어(기사, 도서 등), 구어(대화 등), 웹(메신저 대화, 게시판 등)

과제의 유형 :

자연어 이해(형태 분석, 구문 분석, 문장 유사도 평가 등)

자연어 생성(기계 번역, 추상 요약 등)

혼합(챗봇 등)

+ 자연어처리 데이터를 만들 때는 복잡한 과제도 단순화하여 단계별로 구축

End of Document
Thank You.