인공지능 학습데이터셋 소개

국민대학교 인공지능학부
강 승 식

인공지능 학습데이터셋 종류

- **텍스트**
  - 신문기사, 문학작품, SNS(트위터, 블로그, 영화평 등)
- **음성**(텍스트 전사)
  - 음성인식 목적: 지역별(방언), 연령별, 재외국인 등
  - 노이즈, 환경 등: 자동차실내 소음, 작업현장/공연장 소음, 새소리 등
- **이미지**: 얼굴/사물(재활용품)/간판 등(개체인식 목적)
- **영상**: CCTV 데이터 등(예측/이상탐지 목적)
- **수치 데이터**: 기상정보, 이동인구 등(예측/이상탐지 목적)

<참고> 원시 데이터, 원천 데이터, labelled 데이터

빅데이터 공개 사이트

- http://kaggle.com/
- https://www.gutenberg.org/ -- 영어 문학작품

빅데이터 수집: 대용량 한글 텍스트 데이터

- https://www.aihub.or.kr/ -- 과기정통부 학습데이터 구축사업
- https://kli.korean.go.kr/corpus/ -- 국립국어원 모두의 말뭉치
- https://cafe.naver.com/nlpk/60 -- 국민대 자연어처리 연구실

빅데이터 경진대회 사이트

- https://dacon.io/