ai tech edu.

Upstage AI Tech Education

Practical training for AI in real business

Data Centric AI

01. 교육과정 개요(Overview)

1. 강좌 소개	Data-Centric AI란 데이터 수집, 전처리, 라벨링, 분석 등과 같이 데이터를 중심으로 AI의 품질을 향상하는 접근 방식을 의미합니다. 본 강좌에서는 이러한 Data-Centric AI가 무엇인지 살펴보고, 데이터 구축에 필요한 다양한 이론들과 방법들에 관해서 설명하고 있습니다. 이렇게 데이터 구축의 시작부터 끝까지를 경험해 보는 것은 흔하지 않은 기회이며, 나중에 실제로 데이터 구축이 필요할 때 큰 힘이 되어 줄 것입니다. 혹여나 데이터를 구축하는 일이 없을지라도 AI와 Data는 떼려야 뗄 수 없는 관계이기 때문에, 데이터 구축에 대한 높은 이해도는 AI의 어느 분야를 가시든 큰 도움이 될 것입니다.
2. 강좌 목표	1) Data-Centric AI가 무엇인지, 왜 중요한지 설명할 수 있다. 2) 데이터 구축에 필요한 전체적인 과정에서 대해서 이해하고, 필요한 지식과 기술들을 습득할 수 있다. 3) 다양한 실습을 통해서 데이터 구축과 Data-Centric AI를 실제로 경험해볼 수 있다.
3. 강사의 메시지	본 강좌는 Data-Centric AI가 무엇인지 살펴보고, 데이터 구축에 필요한 다양한 이론들과 방법들에 관해서 학습하고 경험하는 것이 주된 목표입니다. CV나 NLP에 관계없이 다양한 도메인에 대해서 다룰 예정이기 때문에, 딥러닝과 CV 그리고 NLP의 기초를 배우신 상태에서 수강하시면 더욱 좋습니다. 나중에 여러분들께서 어떤 업무를 하시든지 간에 AI와 Data는 떼려야 뗄 수 없는 관계이기 때문에, 본 강좌에서 배운 다양한 이론들과 방법들을 실제로 활용하게 되실 것으로 기대됩니다. 이렇게 데이터 구축의 시작부터 끝까지를 경험해 보는 것은 흔하지 않은 기회이며, 나중에 실제로 데이터 구축이 필요할 때 큰 힘이 되어 줄 것입니다. Data-Centric AI에 대한 교양 수업을 듣는다는 느낌으로 부담 없이 가볍게 들어주시면 감사하겠습니다!

4. 학습 전 참고사항	1) 필수 선수 지식 - Deep Learning Basic - Pytorch - Machine Learning Basic - Computer Vision Basic - Natural Language Processing Basic

02. 교육과정 상세

1 Data-Centric AI 란?

1/ 챕터 소개:

기존의 AI는 모델 중심적인 AI, 즉 Model-Centric AI로써, 모델을 중심으로 데이터를 수집하고, 모델을 학습시키고, 모델을 배포하는 과정을 거쳐왔습니다. 그러나 이러한 기존의 AI는 데이터의 중요성을 간과하고 있었고, 데이터의 중요성을 인지하고 데이터를 중심으로 AI를 구축하는 Data-Centric AI가 등장하게 되었습니다.

본 챕터에서는 앞으로 Chapter 2부터 7까지 상세히 설명될 데이터 구축이 왜 중요한지를 이해하기 위해 Data-Centric AI에 대해 알아보는 시간을 갖습니다. 첫 번째 강의에서는 Data-Centric AI의 개념 및 특징, Data-Centric AI의 대표적인 사례인 Fine-tuning과 Prompt Engineering에 대해 살펴보고, 그 외의 다양한 Data-Centric AI 관련 연구 분야에 대해 알아봅니다. 두 번째 강의에서는 앞으로의 딥러닝의 발전이 Data-Centric AI에 어떤 영향을 미칠지 근미래의 딥러닝 연구 동향을 살펴봅니다.

2/ 챕터 목표:

Data-Centric AI가 기존의 Model-Centric AI와 어떻게 다른지를 이해하고 그 중요성을 인지하는 것, 나아가 Data-Centric AI와 관련된 연구 분야를 폭넓게 이해하는 것이 본 챕터의 주요한 목표입니다.

Upstage AI Tech Education

Practical training for AI in real business

강의 번호	유형	강의명 / 강의 상세
1	이론	– 강의명: Data-Centric AI란? – 강의 상세: 1/ 강의 개요: 본 강의에서는 Data-Centric AI가 무엇인지, 그리고 데이터가 AI 프로젝트를 진행함에 있어 왜 중요한지 알아보는 시간을 갖습니다. 보다 구체적으로, 우선 Data-Centric AI의 개념 및 특징을 알아보고 기존의 Model-Centric AI와의 차이점을 살펴봅니다. 이후, Data-Centric AI의 대표적인 사례라고 할 수 있는 Fine-tuning과 Prompt Engineering에 대해 알아보면서 양질의 데이터를 구축하는 것이 왜 중요한지를 이해합니다. 마지막으로, Data-Centric AI와 관련된 다양한 학회 및 연구 분야에 대해 알아봅니다. 2/ (선택/권장) Further Readings : – A Chat with Andrew on MLOps: From Model-centric to Data-centric AI 3/ (선택/권장) Further Questions : X
2	이론	– 강의명: Data-Centric AI의 미래 – 강의 상세: 1/ 강의 개요: 본 강의에서는 앞으로의 Data-Centric AI가 어떤 방향으로 나아갈 수 있는지 세 가지 키워드를 통해 살펴봅니다. 이는 앞으로의 딥러닝의 발전 방향에 대해 이해하고, Data-Centric AI의 중요성을 인지하는 데에 도움이 될 것입니다. 단, 본 강의에서 다루는 키워드들이 Data-Centric AI의 전부가 아니라는 점을 알아두시면 좋겠습니다. 본 강의에서 다루는 키워드들은 다음과 같습니다. 1) Multimodal : 다양한 형태의 데이터를 처리하는 미래의 AI 2) Multilingual : 다양한 언어의 데이터를 처리하는 미래의 AI 3) Synthetic Data : 합성 데이터를 실제 데이터처럼 활용하는 미래의 AI 2/ (선택/권장) Further Readings : – On the Opportunities and Risks of Foundation Models 3/ (선택/권장) Further Questions : – 본 강의에서 설명한 헬스케어와 교육 말고 다른 분야에서는 기반 모델이 어떤 식으로 자리잡을 수 있을지 생각해봅니다.

2 데이터 기획

1/ 챕터 소개:

앞으로 다섯 챕터에 걸쳐 설명할 데이터 구축 과정에 대한 전반적인 흐름을 이해하고, 데이터 구축을 시작하기에 앞서 어떤 부분들을 데이터 기획 단계에서 고려해야 하는지를 데이터 구축 기획서를 통해 살펴보는 시간을 갖습니다.

2/ 챕터 목표:

데이터 구축이 어떤 흐름으로 이루어지는지, 그리고 각 단계의 산출물이 무엇인지 설명할 수 있는 것을 목표로 합니다.

강의 번호	유형	강의명 / 강의 상세
1	이론	- 강의명: 데이터 구축 프로세스 소개 - 강의 상세: 1/ 강의 개요: 본 강의에서는 데이터 구축 프로세스에 대한 전반적인 흐름을 이해하고, 각 단계가 어떤 과정으로 이루어지는지를 살펴봅니다. 데이터 구축 프로세스는 아래와 같이 여섯 단계로 설명할 수 있습니다. 1) 데이터 수집 2) 데이터 전처리 3) 데이터 라벨링 4) 데이터 클렌징 5) 데이터 스플릿 6) 데이터 릴리즈 2/ (선택/권장) Further Readings : X 3/ (선택/권장) Further Questions : X

2	이론	– 강의명: 데이터 구축 기획서 작성 – 강의 상세: 1/ 강의 개요: 데이터 구축을 시작하기에 앞서 어떤 부분들을 데이터 기획 단계에서 고려해야 하는지를 데이터 구축 기획서를 통해 살펴봅니다. 2/ (선택/권장) Further Readings : – 2023년도 인공지능 학습용 데이터 구축 지원사업 공고 – [붙임 1] 인공지능 학습용 데이터 구축 계획서 3/ (선택/권장) Further Questions : X

3 데이터 수집

1/ 챕터 소개:

앞선 챕터에서 데이터 구축 파이프라인을 전반적으로 설명했다면, 이번 챕터부터는 본격적인 데이터 구축 단계를 하나씩 알아보게 될 것입니다.

이번 챕터에서는 데이터 구축의 첫 두 단계인 데이터 수집과 데이터 전처리에 대해 알아보고자 합니다. 이 두 단계를 통해 만들어지는 결과물은 원시 데이터와 원천 데이터가 될 것입니다. 이를 위해 크게 세 가지 항목, 데이터 수집 방법, 데이터 수집 시 고려해야 할 사항, 그리고 데이터 전처리 방법에 대해 아홉 강의에 걸쳐 살펴보도록 합니다.

2/ 챕터 목표:

본 챕터의 목표는 크게 세 가지로 나눠볼 수 있습니다.

1) 데이터 수집 방법 네 가지가 무엇인지 각각의 특징을 설명할 수 있다.
2) 데이터 수집 시 유의해야 할 요소들과, 이를 고려하여 데이터를 전처리하는 방법에 대해 설명할 수 있다.
3) 원시 데이터를 원천 데이터로 바꾸어 저장하는 과정에 대해 세 단계로 설명할 수 있다.

강의 번호	유형	강의명 / 강의 상세
1	이론	- 강의명: 데이터 수집 1 : 직접 수집 - 강의 상세: 1/ 강의 개요: 본 강의에서는 앞으로 네 강의에 걸쳐 소개될 데이터 수집 방법들에 대해 간략히 짚어보고, 이렇게 수집된 데이터를 활용하여 새로운 데이터를 만드는 방법들에 대해서도 간단히 살펴봅니다. 나아가 데이터 수집 방법 중 가장 원초적인 방식인 직접 수집하는 방법에 대해 도메인별 예시를 통해 살펴봅니다. 2/ (선택/권장) Further Readings : X 3/ (선택/권장)Further Questions : X

2	이론	- 강의명: 데이터 수집 2 : 크롤링 (이론) - 강의 상세: 1/ 강의 개요: 본 강의에서는 두 번째 데이터 수집 방법으로 웹 크롤링(web crawling)을 통한 데이터 수집 방법을 설명합니다. 이를 위해, 우선 크롤링이 무엇인지 그 개념 및 과정을 알아보고, 도메인별로 크롤링을 얼마나 어떻게 활용하고 있는지 알아봅니다. 나아가 다음 강의인 크롤링 실습에 들어가기에 앞서, 크롤링에 필요한 사전 지식을 아주 가볍게 훑어봅니다. 2/ (선택/권장) Further Readings : - GeeksforGeeks – Python Web Scraping Tutorial 3/ (선택/권장)Further Questions : X
3	실습	- 실습명: 데이터 수집 2 : 크롤링 (실습) - 실습 상세: 1/ 실습 개요: 본 실습은 크리에이티브 커먼즈 라이선스(Creative Commons License, CCL)를 가진 이미지를 필터링하고 크롤링하는 파이프라인을 구현하는 것을 목적으로 합니다. 이를 통해 직전 강의에서 배운 크롤링 방법을 실제로 파이썬을 통해 구현해보고, 나아가 저작권을 확인할 수 있는 이미지만 크롤링함으로써 향후 3-6강에서 배우는 저작권 및 라이선스의 개념을 크롤링에 어떻게 반영할 수 있는지를 미리 실습해볼 수 있습니다. 2/ (선택/권장) Further Readings : X 3/ (선택/권장) Further Questions : - 검색 엔진들의 경우, 크롤링 허가/불허 여부를 기입해놓은 'robots.txt'를 제공하기도 합니다. 이 robots.txt를 크롤링 코드에 어떻게 적용할 수 있을까요? - Microsoft의 Bing 검색 엔진에서는 구글 이미지보다도 더 구체적인 CCL 정보를 필터링하여 검색할 수 있습니다. 이를 크롤링 코드로 구현한다면 어떻게 하면 좋을까요?

4	이론	- 강의명: 데이터 수집 3 : 오픈 소스 - 강의 상세: 1/ 강의 개요: 본 강의에서는 세 번째 데이터 수집 방법으로 오픈 소스(open source) 데이터를 활용하는 방법을 살펴봅니다. 이는 여타 데이터 수집 방법에 비해 가장 접근성이 높고 쉬운 방법입니다. 강의의 내용은 크게 둘로 구성되어 있습니다. 전반부는 오픈 소스 데이터에 대한 정의 및 필요성을 설명하고 있으며, 후반부는 국내외 유명 오픈 소스 데이터 플랫폼에 어떤 것들이 있는지 하나씩 알아보는 내용으로 구성되어 있습니다. 2/ (선택/권장) Further Readings : X 3/ (선택/권장) Further Questions : - 오픈 소스를 활용하는 방식은 다른 수집 방법과 달리 타인이 기수집한 데이터를 활용하는 방식입니다. 해당 오픈 소스 데이터가 믿을만한 데이터인지 판단할 수 있는 기준에는 어떤 것들이 있을까요?
5	이론	- 강의명: 데이터 수집 4 : 크라우드소싱 - 강의 상세: 1/ 강의 개요: 본 강의에서는 마지막 데이터 수집 방법인 크라우드소싱을 설명하고 있습니다. 보다 구체적으로는 크라우드소싱의 개념과 활용 사례, 그리고 국내외 크라우드소싱 업체에 대해 알아봅니다. 2/ (선택/권장) Further Readings : X 3/ (선택/권장) Further Questions : X
6	이론	- 강의명: 데이터 수집시 주의 사항 1 : 라이선스 - 강의 상세: 1/ 강의 개요: 데이터를 외부로부터 수집할 때 반드시 고려해야 하는 요소 중 하나는 바로 저작권 침해 여부입니다. 본 강의에서는 저작권과 라이선스가 무엇인지 그 개념을 알아보고, Data-Centric AI와 관련된 라이선스에는 무엇이 있는지 두 가지를 살펴봅니다. 본 강의에서 구체적으로 다루는 라이선스 두 가지는 아래와 같습니다. 1) 크리에이티브 커먼즈 라이선스(CCL) 2) 오픈 소스 라이선스 2/ (선택/권장) Further Readings : - 오픈소스 소프트웨어 라이선스 가이드 3.0 3/ (선택/권장)Further Questions : X

7	이론	- 강의명: 데이터 수집시 주의 사항 2 : 개인정보보호 - 강의 상세: 1/ 강의 개요: 데이터 수집 시 법에 저촉되기 가장 쉬운 요소는 바로 개인정보 침해 여부일 것입니다. 본 강의에서는 개인정보가 무엇인지 그 범주를 살펴보고, 이를 개인정보 보호법에 위반되지 않게 이용하기 위해 알아두어야 할 지침 및 방침들을 알아봅니다. 나아가, 수집된 개인정보를 식별 불가능하게 전처리하는 '개인정보 비식별화' 방법들에 대해서도 알아봅니다. 2/ (선택/권장) Further Readings : - 개인정보 비식별 조치 가이드라인 3/ (선택/권장)Further Questions : 생각해볼거리 - 본 강의에서는 개인정보 비식별화 모델 중 k-익명성 모델에 대해서만 가볍게 설명하고 있습니다. 이 외에도 l-다양성, t-근접성 모델이 있는데 각각이 어떤 방식으로 개인정보를 비식별화하는지 알아봅시다.
8	이론	- 강의명: 데이터 수집시 주의 사항 3 : 윤리 - 강의 상세: 1/ 강의 개요: 구축한 데이터에 사회윤리적인 문제가 내재되어 있다면 이를 통해 학습한 인공지능 모델 및 서비스에도 사회적인 문제가 발생할 가능성이 매우 높습니다. 본 강의에서는 데이터를 구축함에 있어 사회윤리적 요소를 고려해야 하는 이유를 인공지능 및 데이터 윤리를 통해 알아보고, 사회적으로 가장 문제가 될 수 있는 데이터 요소인 '편향성'과 '노이즈'에 대해 사례와 함께 살펴봅니다. 2/ (선택/권장) Further Readings : X 3/ (선택/권장)Further Questions : X
9	이론	- 강의명: 데이터 전처리 - 강의 상세: 1/ 강의 개요: 앞서 8개의 강의를 통해 배운 내용들은 '원시 데이터'를 수집하기 위한 단계에 대한 설명이었습니다. 이번 강의에서는 이 원시 데이터를 정제된 '원천 데이터'로 만들기 위해 거쳐야 할 과정들에 대해 소개합니다. 여기에는 데이터 전처리, 데이터 스키마 설계, 그리고 데이터 저장이 포함됩니다. 2/ (선택/권장) Further Readings : X 3/ (선택/권장) Further Questions : X

4 데이터 라벨링

1/ 챕터 소개:

전처리 과정을 마친 원천 데이터에 어떤 방식으로 어떻게 라벨을 붙일 것인지 '데이터 라벨링' 단계를 설명하는 챕터입니다. 데이터 라벨링에서 특히나 중요한 데이터 라벨링 가이드라인의 작성 방법 및 예시에 대해 세 강의에 걸쳐 설명한 뒤, 마지막 강의에서는 데이터 라벨링에 사용할 수 있는 도구 및 프로그램들을 도메인별로 설명합니다.

2/ 챕터 목표:

본 챕터의 목표는 크게 두 가지로 나눠볼 수 있습니다.

1) 라벨링 가이드라인에 들어가야 하는 필수 요소에 대해 설명할 수 있다.
2) 도메인별로 적합한 라벨링 도구를 하나 이상 설명할 수 있다.

강의 번호	유형	강의명 / 강의 상세
1	이론	- 강의명: 라벨링 가이드라인 작성 방법 - 강의 상세: 1/ 강의 개요: 본 강의는 데이터 라벨링 챕터의 첫 강의로, 우선 데이터 라벨링이 무엇인지, 그리고 어떠한 순서로 진행되는지를 알아봅니다. 나아가, 라벨러에게 제공되는 라벨링 가이드라인을 작성할 때에 반드시 포함해야 하는 필수 구성 요소들에 대해서도 하나씩 살펴봅니다. 2/ (선택/권장) Further Readings : X 3/ (선택/권장)Further Questions : X
2	이론	- 강의명: 데이터 라벨링 규칙 설정 1 : CV - 강의 상세: 1/ 강의 개요: 본 강의에서는 라벨링 가이드라인에 작성해야 하는 규칙들을 CV 도메인 관점에서 살펴봅니다. 이를 위해, 바운딩 박스나 폴리라인과 같은 이미지 및 영상 데이터를 라벨링하는 대표적인 기법들을 우선 알아본 뒤, 각 방법에 따라 고려할 수 있는 규칙에 무엇이 있는지 간단한 예시와 함께 알아봅니다. 2/ (선택/권장) Further Readings : X 3/ (선택/권장)Further Questions : - 강의에서 예시로 설명한 규칙 외에 어떤 규칙을 추가로 적용할 수 있을지 라벨링 기법별로 하나씩 생각해봅니다.

3	이론	– 강의명: 데이터 라벨링 규칙 설정 2 : NLP – 강의 상세: 1/ 강의 개요: 본 강의에서는 라벨링 가이드라인에 작성해야 하는 규칙들을 NLP 도메인 관점에서 살펴봅니다. 이를 위해, 태깅이나 전사 등 텍스트 및 음성 데이터를 라벨링하는 대표적인 기법들을 우선 알아본 뒤, 각 방법에 따라 고려할 수 있는 규칙에 무엇이 있는지 간단한 예시와 함께 알아봅니다. 2/ (선택/권장) Further Readings : X 3/ (선택/권장)Further Questions : – 강의에서 예시로 설명한 규칙 외에 어떤 규칙을 추가로 적용할 수 있을지 라벨링 기법별로 하나씩 생각해봅니다.
4	이론	– 강의명: 라벨링 툴 소개 – 강의 상세: 1/ 강의 개요: 본 강의는 라벨링 툴에 대해 개략적으로 알아보는 강의입니다. 이를 위해 라벨링 툴이 무엇인지, 어떤 기준으로 선정하는 것이 좋은지 등을 간략히 소개한 뒤, 도메인별로 사용할 수 있는 라벨링 툴 및 서비스를 몇 가지 살펴봅니다. 2/ (선택/권장) Further Readings : – HumanSignal/awesome-data-labeling – jsbroks/awesome-dataset-tools 3/ (선택/권장) Further Questions : – 강의에서는 극히 일부 라벨링 툴에 대해서만 다루고 있습니다. 강의에서 설명하고 있지 않지만 목록에 서술된 다양한 라벨링 툴이 각각 어떤 장점을 가지고 있는지 파악해봅시다.

5 데이터 클렌징

1/ 챕터 소개: 데이터 클렌징과 데이터 평가 방법인 IAA에 대한 챕터입니다. 데이터 클렌징이 필요한 이유와 각 과정에서의 고려 사항을 배워봅니다. IAA를 직접 계산하고 데이터 품질이 좋을 때와 나쁠 때 각각의 경우에 대해 어떻게 접근하면 좋을지 알아봅니다.
2/ 챕터 목표: 데이터 클렌징 과정을 이해하고 IAA를 활용해 데이터 품질을 평가할 수 있습니다.

강의 번호	유형	강의명 / 강의 상세
1	이론	– 강의명: 데이터 클렌징 – 강의 상세: 데이터 클렌징의 정의 및 필요성과 그 방법에 대해 학습합니다. 1/ 강의 개요: – 라벨링 에러가 존재할 경우, 그것을 수정하여 데이터 품질을 향상시키는 과정이 필요합니다. 이번 강의에서는 데이터 클렌징이 필요한 이유와 라벨링 에러를 확인하는 방법, 그리고 마지막으로 각각의 라벨링 에러 케이스에 대한 데이터 클렌징 접근 방식을 다뤄보았니다. 3/ (선택/권장)Further Questions : – 라벨링 규칙을 수정하는 데에는 어떠한 리스크가 있을까요?
2	이론	– 강의명: 데이터 평가 방법 : IAA – 강의 상세: IAA와 그 평가 방법에 대해 학습합니다. 1/ 강의 개요: – IAA가 데이터 품질을 평가하는 데 어떠한 역할을 하는지, 또 어떠한 한계를 가지는지 그리고 IAA는 어떻게 계산할 수 있는지에 대해 다뤄보았습니다. 2/ (선택/권장) Further Readings : – Natural Language Annotation for Machine Learning : 머신 러닝 라벨링 방법 및 과정에 대해 서술된 자료입니다. 3/ (선택/권장)Further Questions : – 상황에 따라 어떠한 IAA 평가 방법을 쓰는 것이 효과적일까요?

3	이론	– 강의명: IAA를 활용한 데이터 클렌징 방법 – 강의 상세: IAA를 활용하여 데이터 클렌징하는 방법에 대해 학습합니다. 1/ 강의 개요: – IAA만으로 데이터 클렌징하는 경우와 IAA와 다른 Metric을 함께 활용하여 데이터 클렌징하는 경우에 대해 다뤄보았습니다. IAA를 기준으로 데이터 품질을 평가하고 향후 방향을 설정할 수 있습니다. 3/ (선택/권장)Further Questions : – 데이터의 품질을 높이기 위해서는 어떻게 해야할까요?

6 데이터 마무리

1/ 챕터 소개: 데이터 클렌징까지 마친 후의 데이터 스플릿, 합성 데이터 생성, 액티브 러닝, 데이터 릴리즈 등의 과정에 대한 챕터입니다. 데이터를 나누고 샘플링하는 과정에 대해 학습하고 합성 데이터를 생성하여 데이터의 부족한 부분을 채우는 방법에 대해 배워봅니다. 라벨링 데이터 또는 자원이 부족한 경우 액티브 러닝을 활용해 효율적으로 라벨링 데이터를 추가할 수 있습니다. 마지막으로 완성한 데이터를 배포할 때의 고려할 점에 대해 배울 수 있습니다.
2/ 챕터 목표: 데이터 스플릿과 샘플링의 개념을 익히고 합성 데이터 생성과 액티브 러닝을 활용하여 데이터의 부족한 부분을 채울 수 있다. 완성한 데이터를 공개할 수 있다.

강의 번호	유형	강의명 / 강의 상세
1	이론	– 강의명: 데이터 스플릿 – 강의 상세: 데이터 스플릿과 데이터 샘플링에 대해 학습합니다. 1/ 강의 개요: – 데이터 스플릿을 하는 이유와 데이터 샘플링 방법론들에 대해 다뤄보았습니다. 데이터의 특성과 분포를 고려하여 데이터를 샘플링할 수 있습니다. 그를 바탕으로 모델 학습 및 통계 자료 제공에 활용할 수 있습니다. 2/ (선택/권장) Further Readings : – TORCH.UTILS.DATA : 아래 생각해볼 거리와 관련된 내용입니다. Pytorch 라이브러리의 DATA 패키지에 대한 설명입니다. 3/ (선택/권장)Further Questions : – Pytorch Dataloader에는 sampler 옵션이 있습니다. 데이터를 어떻게 나누는 것이 학습에 도움이 될까요?

2	이론	– 강의명: 데이터 평가 방법 : 합성 데이터 : CV (이론) – 강의 상세: 컴퓨터 비전 분야의 합성 데이터 생성에 대해 학습합니다. 1/ 강의 개요: – 합성데이터와 그 활용에 대해 다뤄보았습니다. 합성데이터를 사용했을 때의 이점과 한계점을 알 수 있으며 데이터 증강과 합성 데이터의 공통점과 차이점을 구분할 수 있습니다. 합성 데이터 생성 방법을 배우고 이해할 수 있습니다. 3/ (선택/권장)Further Questions : – 합성 데이터 생성으로 해결할 수 있는 문제들에는 어떠한 것들이 있을까요?
3	실습	– 실습명: 합성데이터 : CV (실습) – 실습 상세: GAN, VAE, Diffusion 모델을 사용해 합성 데이터를 생성합니다. 1/ 실습 개요: – GAN, VAE, DIffusion 등 대표적인 이미지 데이터 생성 모델들을 직접 사용해보고 구현해볼 수 있습니다. 네트워크를 직접 설계하고 생성된 이미지를 직접 확인해볼 수 있습니다. 2/ (선택/권장) Further Readings : – Diffusers : 허깅페이스 디퓨저 라이브러리 링크입니다. – Denoising Diffusion Probabilistic Models : Diffusion probabilistic 모델을 소개한 논문입니다. – Score-Based Generative Modeling through Stochastic Differential Equations: 확률적 미분 방정식(SDE)를 활용해 Diffusion process를 모델링하는 방법을 소개한 논문입니다. 3/ (선택/권장)Further Questions : – 어떻게 하면 실습의 예제 모델을 더 발전시킬 수 있을까요?
4	이론	– 강의명: 합성 데이터 : NLP(이론) – 강의 상세: 자연어 처리 분야의 합성 데이터 생성에 대해 학습합니다. 1/ 강의 개요: – 자연어 처리 합성 데이터는 어떠한 특성을 가지는 지 다뤄보았습니다. 자연어 처리 과제(Task)에 맞게 데이터를 생성하고 활용할 수 있습니다. RNN, BART, LLM 등의 모델을 활용해 합성 데이터를 생성할 수 있습니다. 2/ (선택/권장) Further Readings : 학습보충자료, 읽어보면 좋을 추가 논문 등 – Ouyang et al., Training language models to follow instructions with human feedback (2022): InstructGPT 논문 자료 3/ (선택/권장)Further Questions : 생각해볼거리 – 텍스트 데이터 만의 특징에는 어떠한 것들이 있을까요?

5	실습	– 실습명: 합성 데이터 : NLP(실습) – 실습 상세: RNN, BART, LLM 등의 모델을 활용해 합성 데이터를 생성합니다. 1/ 실습 개요: – RNN, BART, LLM 등 대표적인 텍스트 데이터 생성 모델들을 직접 사용해보고 구현해볼 수 있습니다. 네트워크를 직접 설계하고 생성된 데이터를 직접 확인해볼 수 있습니다. 허깅페이스 라이브러리를 적극적으로 활용합니다. 2/ (선택/권장) Further Readings : – HuggingFace Transformers : 허깅 페이스 트랜스포머 라이브러리에 대한 문서 자료입니다. 3/ (선택/권장)Further Questions : – NER 외에 다른 과제의 합성 데이터는 어떻게 만들 수 있을까요?
6	이론	– 강의명: 데이터 평가 방법 : 액티브 러닝(이론) – 강의 상세: 액티브 러닝과 그 방법에 대해 학습합니다. 1/ 강의 개요: – 액티브 러닝과 그 사이클에 대해 다뤄보았습니다. 액티브 러닝을 어떠한 상황에서 사용하는 것이 효과적인지 또 그것의 한계는 무엇인지에 대해 학습할 수 있습니다. 액티브 러닝 샘플링 방법들을 알아보고 각각의 샘플링 방식을 이해할 수 있습니다. 2/ (선택/권장) Further Readings : – B Settles, Active Learning Literature Survey (2010) : 액티브 러닝의 시나리오, 전략, 분석 등 다양한 측면을 다루고 있는 survey 자료입니다 3/ (선택/권장)Further Questions : – 액티브 러닝의 적절한 수준은 어느 정도일까요?
7	실습	– 실습명: 액티브 러닝 (실습) – 실습 상세: Uncertainty Learning, Query-by committee 등의 액티브 러닝을 구현합니다. 1/ 실습 개요: – Uncertainty Learning, Query-by committee을 직접 구현해보고 최소 신뢰 기준, 최소 마진 기준, 엔트로피 기준 등의 접근 방식들을 다뤄봅니다. 다양한 액티브 러닝을 직접 구현할 수 있으며 랜덤 샘플링과의 차이점을 확인해볼 수 있습니다. 2/ (선택/권장) Further Readings : – modAL: A modular active learning framework for Python3 : 액티브 러닝 프레임워크인 modAL입니다. 3/ (선택/권장)Further Questions : – 액티브 러닝의 다른 방식에는 어떠한 것들이 있을까요?

8	이론	– 강의명: 데이터 릴리즈 – 강의 상세: 데이터 릴리즈 방법에 대해 학습합니다. 1/ 강의 개요: – 데이터 릴리즈 시의 고려할 점과 데이터 릴리즈하는 방법들에 대해 다뤄보았습니다. 3/ (선택/권장) Further Questions: – 어떠한 데이터가 좋은 데이터일까요? 사람들이 많이 사용하는 데이터에는 어떠한 특징이 있을까요?

7 데이터 제작 실습

1/ 챕터 소개: CV 파트와 NLP 파트의 데이터 제작에 관한 챕터입니다. 각각 객체 탐지, 개체명 인식 과제를 중심으로 이루어져 있으며 데이터 기획 부터 데이터 구축, 검수 및 평가까지의 과정을 다루고 있습니다. 데이터 제작 시에 도움이 되는 라이브러리와 방법론들을 배울 수 있으며 그를 직접 활용해 데이터의 품질을 높이고 모델 성능을 높일 수 있습니다.
2/ 챕터 목표: 데이터 제작 과정을 직접 체험하고 이해할 수 있다. 데이터 제작 과정의 어려움을 직접 다뤄보고 고민해본다.

강의 번호	유형	강의명 / 강의 상세
1	실습	– 실습명: CV 데이터 제작 실습 – 실습 상세: CV 데이터 제작의 전과정을 다룹니다. 객체 탐지(Object detection) 과제로 구성되어 있으며 데이터 중심 접근 방식을 통해 베이스 라인 모델의 성능을 높이는 것을 목표로 합니다. 1/ 실습 개요: – 데이터 기획, 베이스라인코드, 데이터 구축, 데이터 검수, 데이터 평가의 과정을 따라 데이터 제작에 대해 고민해보고 배워봅니다. 오픈 데이터셋을 찾고 합성 데이터를 생성하여 데이터의 볼륨을 키우고 다양성을 확보할 수 있습니다. 데이터 라벨링을 직접 해보고 라벨링을 검토하고 클렌징합니다. 액티브 러닝 또한 직접 사용해봅니다. 마지막으로 구축한 데이터셋을 활용해 베이스라인 모델의 성능을 향상시켜 봅니다. 2/ (선택/권장) Further Readings : – Hugging Face Stable Diffusino Pipeline : stable diffusion 파이프라인 문서입니다. Image variation, dept-to-image, super-resolution 등의 파이프라인을 활용할 수 있습니다. 3/ (선택/권장)Further Questions : – CV 데이터 제작 과정에서의 어려움은 무엇일까요? – 이미지 데이터를 처리할 때의 주의점은 무엇인가요? – 제작한 이미지 데이터를 기반으로 서비스를 만들 때 어떠한 잠재적 문제가 발생할 수 있을까요?

2	실습	– 실습명: NLP 데이터 제작 실습 – 실습 상세: NLP 데이터 제작의 전과정을 다룹니다. 객체명 인식(Named Entity Recognition) 과제로 구성되어 있으며 데이터 중심 접근 방식을 통해 베이스 라인 모델의 성능을 높이는 것을 목표로 합니다. 1/ 실습 개요: – 데이터 기획, 베이스라인코드, 데이터 구축, 데이터 검수, 데이터 평가의 과정을 따라 데이터 제작에 대해 고민해보고 배워봅니다. 데이터 크롤링을 통해 관련 데이터를 수집하여 데이터의 다양성을 더합니다. 데이터 라벨링을 직접 해보고 또 자연어 처리 라이브러리를 활용해 해봅니다. 이후 라벨링을 검토하고 클렌징합니다. 액티브 러닝 또한 직접 사용해봅니다. 마지막으로 구축한 데이터셋을 활용해 베이스라인 모델의 성능을 향상시켜 봅니다. 2/ (선택/권장) Further Readings : – Spacy : 자연어 처리 라이브러리인 spacy입니다. 자연어 처리 형태소 분석, 태깅 등을 API를 활용해 간단하게 수행할 수 있습니다. 3/ (선택/권장)Further Questions : – NLP 데이터 제작 과정에서의 어려움은 무엇인가요? – 제작한 텍스트 데이터를 기반으로 서비스를 만든다면 어떠한 처리가 필수적일까요?