kakao tech
bootcamp

kakao × goorm

생성형 AI

Day 1

강사소개

이지형 (Jhin) - 구름 AI R&D Engineer

- 구름 IDE Codevisor
  - 코드 어드바이징 서비스
- 구름 DEVTH AIxObserview
  - 온라인 테스트 부정행위 감지 서비스
- AI DigitalTextbook AI 구르미
  - 디지털 교과서 보조교사 및 학습 도우미
- 구름의 Gen AI SQD.
  - AI 서비스 기획, 관리, 개발

목차

일정 안내

- 일정은 오전 9시 시작
- 10분 전후로 출석체크
- AM 9:00 ~ AM 11:00 → 이론
- AM 11:00 ~ PM 12:00 → 실습
- PM 12:00 ~ PM 1:00 → 점심시간
- PM 1:00 ~ PM 2:00 → 실습
- PM 2:00 ~ PM 6:00 → 미션

강의 도구

- 구름 EDU
  - 강의 내용 공유 및 퀴즈
- 구름 EXP
  - 미션 및 학습일기
- ZEP, Discord
  - 커뮤니케이션
- Notion
  - 컨텐츠 공유 및 프로젝트 관리, 기록

강의 시작전 **Q&A**

강의 시작 전 하고싶은 질문!

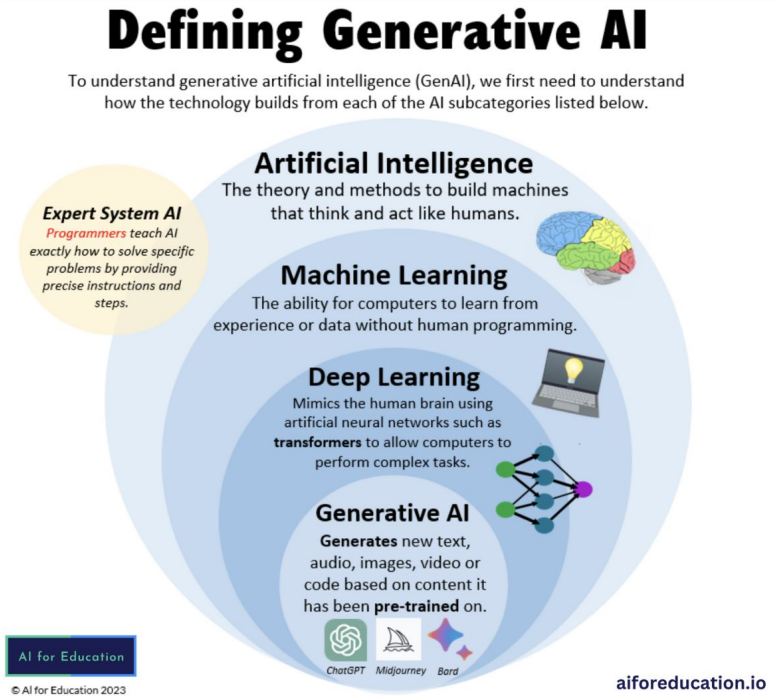
강의에 앞서

우리가 알아야 하는 것

- AI가 뭘까요?
- Generative AI가 뭘까요?

우리가 생각해봐야 하는 것

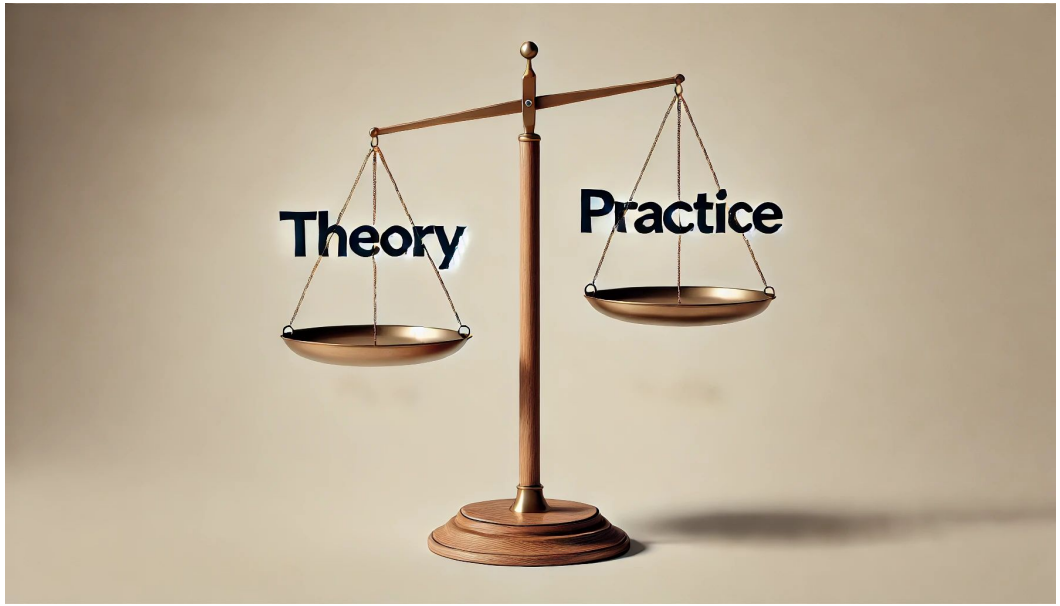
- AI로 무엇이 하고싶나요?
- 어떤 AI Engineer가 되고싶나요?
- 우리는 대체 뭘해야할까요?



https://www.aiforeducation.io/ai-resources/generative-ai-explainer

강의계획

- 목적: 생성형AI 기반 서비스 개발을 위한 기초 학습 (7월)
- 수업 구성: 이론 + 실습



강의계획

데이터의 활용 및 구현

- 데이터 수집
- 데이터 전처리

파이썬 프로그래밍

- 프로그래밍 초급
- 프로그래밍 중급
- 프로그래밍 고급

데이터 분석

- 데이터와 EDA
- 데이터 분석과 문제해결능력

데이터 시각화

- 다양한 데이터와 시각화
- 데이터 시각화와 해석

머신러닝

- 머신러닝 기초
- 머신러닝 초급
- 머신러닝 중급
- 머신러닝 고급

딥러닝

- 딥러닝 기초
- 딥러닝 초급
- 딥러닝 중급
- 딥러닝 고급

생성형 AI

- Gen AI
- 프롬프트
- LLM과 RAG
- 생성형 AI의 미래

개발환경

- 여러분들은 모두 macbook을 지원받으셨습니다.
- Local에서 진행해도 OK
  - python
  - anaconda
  - virtualenv
- Colab에서 진행해도 OK
- 구름 IDE에서 진행해도 OK
- python 3.10
- library ver. stable

Git 세팅

- Github 세팅 필요 - public
- 목적
  - 강의, 공부, 실습
  - 미션, 질문, 아이디어 정리
- 나를 드러내서 팀 구성에 이점
- 백업용이
- E만 가득한 부트캠프안에서 I가 살아남을 수 있는 방법
- Discord 등에 주소 공유
  - 적절한 장소나 방법이 있는지 추후 공지

kakao tech bootcamp

kakao × goorm

생성형 AI

Day 1

데이터의 활용 및 구현 I

데이터의 형태

- 데이터란 무엇인가?
  - 의사결정 및 분석을 위한 원자재

- 데이터의 중요성
  - 의사결정 도구 - 데이터 중심 사고
  - 빅데이터 시대
  - AI

데이터와 AI

IT·과학

생성형AI 시대 … 결국엔 '똘똘한 데이터'가 경쟁력 가른다

“AI의 핵심은 데이터, 데이터를 이해하는 도메인 전문가 역할이 중요” [AI SEOUL 2024]

AI의 보편화와 일상화, 데이터 이해 중요
클라우드 기반 AI 플랫폼, 진입 장벽 낮춰

[AI 데이터 고갈 위기]"2년 후 AI 성장 멈출 수도"…데이터 절벽 '경고'

데이터의 형태

정형 데이터

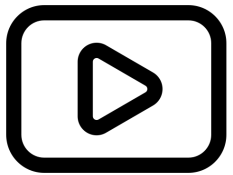
- 구조화된 형식
- 고정된 스키마
- 무결성
- 저장과 관리가 효율적
- 데이터 분석에 용이
- 유연성이 부족
- RDBMS

ID_No	Name	Email	Age
1	James	James@korea.com	21
2	Jack	Jack@korea.com	23
3	Paul	Paul@korea.com	32

데이터의 형태

비정형 데이터

- 텍스트, 이미지, 동영상 다양한 데이터 유형
- 고정된 스키마, 구조가 없음
- 대용량
- 풍부한 정보량
- 저장, 분석이 어려움
- 일관성없는 품질
- NoSQL



데이터의 형태

반정형 데이터

- 일정한 구조를 갖지만 형식이 고정되지 않은 데이터
- 자기 설명적
- 정형 데이터와 비정형 데이터의 장단점 공유
- JSON, XML

```
{
    "ID_No": 1,
    "Name": "James",
    "Email": "James@korea.com",
    "Age": 21
  },
  {
    "ID_No": 2,
    "Name": "Jack",
    "Email": "Jack@korea.com",
    "Age": 23
  }
```

데이터와 인사이트

- 데이터 인사이트?
  - 데이터 분석을 통해 도출된 유용한 통찰이나 정보
  - 단순히 데이터를 해석한 결과값이 아닌 의미있는 정보
- 그러면 데이터 인사이트가 왜 중요한가?
  - 데이터 분석가
    - 데이터 기반 의사결정
  - AI 개발자
    - 모델 성능개선, 데이터 품질
- 데이터 분석과 AI의 궁극적인 목표는 유의미한 인사이트 도출
  - 다양한 이해관계자가 인사이트를 활용하여 의사결정과 혁신을 이끌어내는 것이 중요

데이터의 수집 방법

내부 데이터 **VS** 외부 데이터

VS

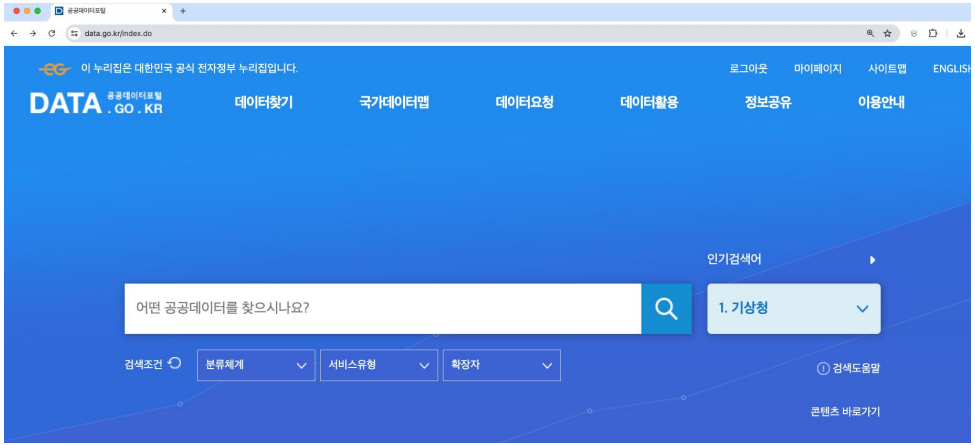
- 내부 시스템에서 수집한 데이터
- 영업, 고객, 트랜잭션 데이터
- 소스코드, 사원정보 등

- 외부 시스템의 데이터
- 공공데이터
- 시장, 조사데이터

데이터의 수집 방법

- OpenAPI
  - 누구나 접근할 수 있도록 개방된 API
  - 구조화된 데이터 제공
  - 인증 필요 (API 키)

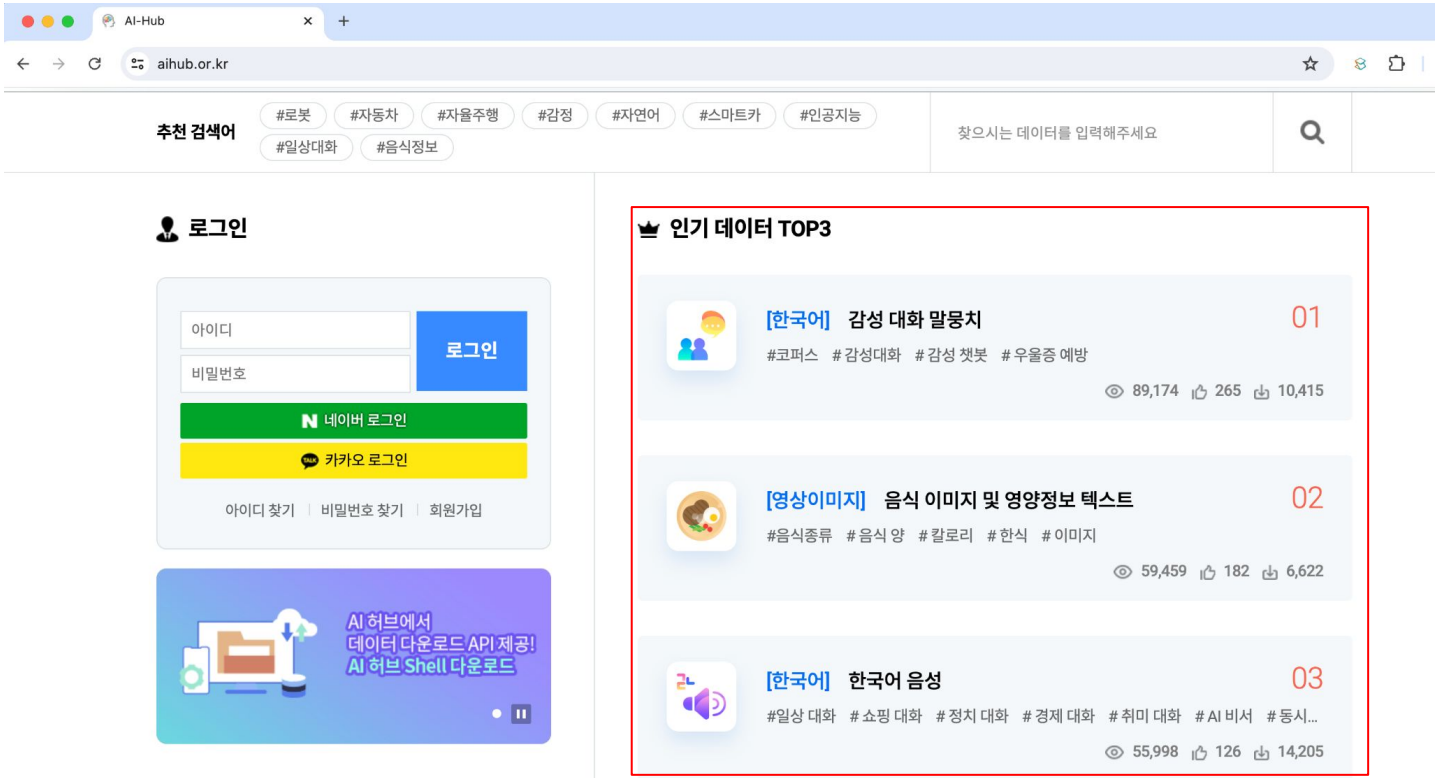
- Open Data
  - 누구나 접근할 수 있도록 공개된 데이터
  - 정부, 기관, 단체 등에서 제공



데이터의 수집 방법

- 크롤링
  - 자동화된 스크립트를 사용하여 웹 페이지를 탐색하고 데이터 수집
  - 전체 사이트 또는 다수의 페이지를 탐색
  - 검색 엔진의 크롤러와 유사
- 스크래핑
  - 특정 웹 페이지에서 필요한 데이터 추출
  - 웹 크롤링의 하위 집합으로서 특정 데이터만 수집

데이터 수집 예시 (스크래핑)



데이터 수집 예시 (스크래핑)

```python
import requests
from bs4 import BeautifulSoup
# AI Hub 페이지 URL
url = 'https://www.aihub.or.kr/'
# 웹 페이지 요청
response = requests.get(url)
response.raise_for_status()  # 요청이 성공했는지 확인
# BeautifulSoup 객체 생성
soup = BeautifulSoup(response.content, 'html.parser')
# 인기 데이터 TOP3 섹션 찾기
top3_section = soup.find('div', class_='secR')
# 각 데이터 항목 추출
data_list = top3_section.find_all('div', class_='list')
```

```python
# 데이터 제목 추출
titles = []
for data in data_list:
    title = data.find('h3').get_text(strip=True)
    clean_title = title.split(']')[-1].strip()
    titles.append(clean_title)

# 추출한 데이터 출력
for idx, title in enumerate(titles, start=1):
    print(f"TOP {idx}: {title}")
```

```
OUTPUT
TOP 1: 감성 대화 말뭉치
TOP 2: 음식 이미지 및 영양정보 텍스트
TOP 3: 한국어 음성
```

크롤링과 스크래핑에 대해 조금 더 알아봅시다

- requests
  - HTTP 요청을 보내고 응답을 받기 위한 라이브러리.
- BeautifulSoup
  - HTML 및 XML 문서를 파싱하여 원하는 데이터를 추출하기 위한 라이브러리.
- Scrapy
  - 크롤링 및 스크래핑을 위한 프레임워크.
  - 정적 웹 크롤링 강점
- Selenium
  - 웹 브라우저 자동화를 위한 라이브러리.
  - 동적 웹 스크래핑 강점

requests

- Colab 참고

BeautifulSoup

- Colab 참고

Selenium

- Colab 참고

실습 과제

- 개발환경 세팅
  - Python 개발환경
  - github
- 7/3 사전평가 진행예정
  - AI 전반에 대한 질문
  - 진로에 관한 간단한 질문
  - 향후 강의 난도 조절
  - 따로 평가에 반영되는 시험은 아니므로 정답, 점수 공개 X
- 크롤링
  - scrapy를 사용해서 한국 위키피디아 또는 원하는 웹사이트 크롤링해보기

kakao tech bootcamp

kakao × goorm

실습 진행