# 윤도연

**Github**: github.com/ydy8989

**E-mail**: ydy89899@gmail.com

안녕하세요. 5년차 데이터 과학자 윤도연입니다.
다양한 도메인의 프로젝트에서 주어진 조건과 트레이드 오프를 고려한 최적의 결과를 도출하는 데 집중합니다.
상황에 맞는 기술을 깊이 있게 다루며, 필요할 땐 빠르게 새로운 스택에도 적응하며 팀의 성장 방안을 모색합니다.
모르는 것은 바로 질문하며, 제 의견을 드러내는 것에 어려움이 없습니다.

## 업무 경험

### 컬리 | Data Scientist | 2025.04 ~ 재직중

- 컬리몰 검색/추천 서비스 내 ML 모델의 개발·운영
- 유저 데이터 분석과 이를 기반으로 하는 프로덕트 가설 수립 및 검증
- OpenSearch 기반 검색 엔진 활용, BigQuery 데이터 핸들링 및 Airflow 배치 생성 업무
- 이에 기반한 시멘틱 검색 적용을 위한 PoC 업무 진행

**Data Scientist**

2025.05 ~ 2025.08

#### 검색 엔진 Semantic Search 적용을 위한 PoC 진행 → LINK

휴먼 리소스 기반의 비효율적 Lexical 검색 유지 구조에서 벗어나기 위해 Semantic 검색 도입의 필요성 확인

- Semantic 검색 적용을 위한 Fine-tuned 모델의 도입 가능성 검토 및 실험 설계
- 검색량 상위 키워드(Head query) 기준, 상품 유사도 기반 recall이 기존 대비 9% 향상됨을 검증
- PoC 결과를 바탕으로 Semantic 검색 도입 결정의 핵심 근거 제공, 조직 내 도입 확정 견인

### 라이앤캐처스 | Data Scientist & ML Engineer | 2021.11 ~ 2025.03

- 다양한 산업군의 프로젝트에서 데이터 분석 및 머신러닝 모델 개발 주도
- Python, Pandas, Pytorch, Scikit-learn, 등을 활용한 분석 및 모델링 파이프라인 다수 구축
- 도메인별 요구사항에 따라 데이터 수집 → 전처리 → 모델 학습/평가 → 배포까지 전 단계 경험

**ML Engineer**

2024.09 ~ 2025.02

#### 한림대학교 RAG 기반 AI 조교 구축 → LINK

강의계획서, 논문 초안, 문제 생성 및 수정 등의 기능을 지닌, 교수님들을 위한 AI 조교 챗봇 개발 (1/4년차) 진행

- LangChain을 활용한 Advanced RAG 파이프라인 설계 구현 주도

- Hybrid search 적용 & IVF 인덱싱 도입을 통한 Retriever 속도 및 정확도 향상
- 문항 생성 니즈에 맞는 retriever 구현 및 프롬프트 엔지니어링으로 베타 테스터(교수진) 20명의 생성 요청 후 생성물 저장 비율 20% 향상

**Data Scientist**

2023.06 ~ 2024.04

### AI 채용서류 평가 자동화 모델 구현 → LINK

채용 서류 내 14개 위반사항(표절, 블라인드 위반, 복붙 등) 검출 시스템 개발

- 토큰 역인덱스 + 해시값 비교를 통한 표절 검출 스크립트의 완료 속도 2일에서 2시간으로 단축
- SimCSE 기반 문장 임베딩과 leave-one-out 평균 비교 방식 도입하여, 문항에 상관없는 답변 검출시 **휴먼 리소스 90% 이상 감소**
- 전체 14개 위반 사항 검출 정확도 **recall > 0.95, f1-score 기존 대비 0.2 향상**

**ML Engineer**

2023.01 ~ 2023 03

### 사내 지식거래 플랫폼 유사 문서 랭킹 → LINK

자사 신규 서비스에 들어갈 "문서 클러스터링" 기능 구현을 위한 랭킹 시스템 구현

- 작성 시점 이전 문서에 한정된 유사도 탐색 로직을 설계하여 지식재산권 보호 기준 반영 및 검색
- KoELECTRA 기반 문서 임베딩과 SimCSE 기반 리랭커 조합으로 50개 쿼리 문단에 대한 정량평가 **Hit@3 = 0.92**
- 기존 MongoDB 활용 벡터 저장 → FAISS 벡터 검색 적용으로 **평균 검색 소요 시간 1.2에서 0.4초로 약 66% 단축 개선**

**Data Scientist**

2022.09 ~ 2022.12

### "대스타 해결사 플랫폼" 데이터 증강 대회 → LINK

텍스트 분류 모델에 대한 정보 없이, 데이터 증강 전략만으로 성능을 극대화하는 데이터 증강 대회에서 준우승 수상(상금 9천만원)

- 모델 정보 없이 주어진 원시 텍스트만으로 텍스트 분류 성능을 향상시켜야 하는 조건 하에서, 전략 중심의 데이터 증강 기법을 설계 및 적용
- 문체 변환 모델, Back Translation, TEM 등 다양한 증강 기법을 앙상블 적용하고, 사전 구축한 자동화 파이프라인을 통해 증강 효율 및 다양성 확보
- 실험 환경의 빠른 셋업 및 증강 전략 최적화를 통해 제한된 시간 내 높은 품질의 증강 데이터 생성, *준우승 달성*

---

## ScatterX | Data Scientist | 2019.6 ~ 2020.3

- 반도체 공정 내 이상 탐지 자동 분류 모델 개발 → 삼성전자 전 공정 적용
- 센서 기반 Interlock 분류를 통한 휴먼 리소스 절감

**Data Scientist**

### 삼성전자 반도체 센서데이터 이상탐지 모델 개발 → Link

일 발생 수천만건의 반도체 센서 데이터의 공정 기준을 만족하지 못할 시 정지시키는 인터락 상황에 대해, 진짜 고장 상황인지 여부를 분류하는 모델 개발

2019.06 ~
2020.03

- 인터락 발생 시점 기준 과거 30일 시계열 센서 데이터 기반 7종 이상 케이스 분류를 위한 데이터 처리 및 모델링 파이프라인을 주도적으로 설계 및 구현
- 단일 모델의 한계를 보완하기 위해, 7가지 케이스별 특성에 따라 Autoencoder, SGAN, 회귀 모델을 병렬 적용하는 구조를 설계하여 분류 정확도와 유연성을 향상
- Precision 0.99, F1 score 0.83 달성하고, 전체 공정에 적용하여 휴먼 리소스 80% 절감에 기여

---

### 포항공대 인공지능연구원 | Intern Data Scientist | 2018.11 ~ 2018.12

- 경북 중소기업 기술 애로사항 극복을 위한 AI 기술지원
- 연구부 연구 보조 - 베이스라인 구현, 논문 서베이 보조
- 아산병원 "HeLP Chalenge 2018 Contest" 참가. 최종 5위

## 그 외 프로젝트

### Chat PDF Clone coding

2023.12 ~
2024.01

- Langchain 숙련도를 올리기 위한 클론코딩
- FAISS를 활용하여 PDF 파일의 문서 청크 인덱싱 후 간단한 RAG chain으로 구현
- Streamlit을 활용하여 배포

---

### 비-글(Bigle) 문체 변환 서비스

2022.01 ~
2022.10

- 글 또는 대화를 입력했을 때 성경(Bible) 말투로 변환해주는 서비스. 이후 급식체, 할아버지체, 등등 다양한 문체로 말투를 바꿔주는 모델을 학습하였음.
- KoBart 모델을 Fine tuning 후 BentoML을 이용하여 서빙 진행
- 전사 개발 조직을 위한 사이드 프로젝트로써 추후 “대스타 해결사 플랫폼” 대회 진행시 만들어 놓은 문체들을 활용하여 데이터 증강을 빠르게 진행하는데 도움이 됨

---

### Dialogue State Tracking

2021.05 ~
2021.06

- DST 분야 대회 참여. Public 1위,  Private 1위 달성
- SOTA 모델인 SOMDST 모델 구현 및 최적화로 다른 팀원의 모델과 앙상블 성능 개선에 기여
- BLEU 스코어 후처리를 통해 토큰 생성 후보정으로 점수 향상에 기여

---

### Senticle: 뉴스 기사를 활용한 주가 상하락 보조 앱

2018.09 ~
2018.10

- 뉴스기사를 활용한 주가 상하락 예측 모델 구현
- 투자 보조를 위해 예측의 신뢰구간을 수치적으로 명확히 제시
- LIME을 활용하여 모델 예측 결과에 대한 근거를 시각적으로 제시
- 앱 데모 : https://www.youtube.com/watch?v=syQfQGFAAZ0

# AI Training

### Naver Boostcamp AI Tech 1기

2021.01 ~ 2021.06

- 약 6개월간 네이버 현업자분들과 함께하는 교육 이수 (NLP, CV, GNN 등)
  - 전반기 3개월 : 온라인 강의 수료
  - 후반기 3개월 : 대회 형식으로 리더보드 순위를 가르는 프로젝트 진행

### 포항공대 PIRL(현 PIAI) - AI/Big Data Advanced Course

2018.09 ~ 2018.10

- 2개월간의 교육 (Tensorflow, Pytorch, 인공지능 개론 등)과 함께 팀 프로젝트를 진행

# Skills

### 숙련

실무에서 주도적으로 사용

- Python (Pandas, NumPy, Scikit-learn, PyTorch)
- sLM Fine tuning(HuggingFace)
- LLM 오케스트레이션(Langchain)
- Search Engine(OpenSearch)
- Serving(FastAPI)
- NoSQL(MongoDB, Redis)
- SQL(BigQuery)
- VectorDB(Faiss, chromaDB, Milvus)
- Docker

### 사용 경험 있음

사이드 프로젝트 및 실험 수준에서 사용 혹은 예제 테스트

- LLM Fine tuning(LoRa, Deepspeed, peft)
- Serving(Langserve, BentoML, Cog)
- No Code Tool(Dify)
- MLflow, Kubeflow, Airflow

# Education

- 2008.03 ~ 2013.03 - 서울과학기술대학교 기계시스템・디자인공학과
- 2013.03 ~ 2015.02 - 광운대학교 수학과
- 2015.03 ~ 2017.02 - 광운대학교 일반대학원 수학과