데이터 전처리 및 텍스트마이닝 특강

목차

- 강사 소개
- 텍스트 데이터 수집 (Youtube API 활용)
- 텍스트 데이터 전처리
- 텍스트 데이터 시각화
- 데이터 직군 소개
- Q&A

강사 소개

사수 없는 스타트업에서
유일한 데이터 분석가로 커리어 시작

2022

- 분석용 이벤트 로그 설계
- 데이터 인프라 도입 및 설계 (Google BigQuery, Firebase, Airflow, …)
- Data Warehouse, Data Mart 운영
- 주요 지표 모니터링 대시보드 제작
- 유저 구매 예측 모델링, 모델서빙
- A/B 테스트 (세미나, 인터뷰 진행)
- …

2023 ~ 현재

- Data Team Lead
- 클라우드 데이터 플랫폼 구축,고도화
- 전사 데이터 리터러시 & SQL 교육
- 데이터 ETL 파이프라인 개발
- …

Data Analysis

Data Delivery

Data Guide

텍스트 데이터 수집

Youtube API

Google Colab Link

텍스트 데이터 전처리

정규표현식, 형태소 분석기

Google Colab Link

데이터 시각화

워드클라우드, 시각화 BI tool

Google Colab Link

데이터 직군 소개

AI 에서 조금 더 넓혀보기


Big Data
Data Scientist
Engineering
Python
SQL
Artificial Intelligence
Machine Learning
LLM
Generative AI
AI Engineer
Dacon
Data Analyst
Deep Learning
PyTorch
Data
Tensorflow
Kaggle
Data Engineer
ChatGPT
AB Test
Classification
Transformer
Pandas
ML Engineer
Analysis
Prediction
MLOps

ML code 의 비중

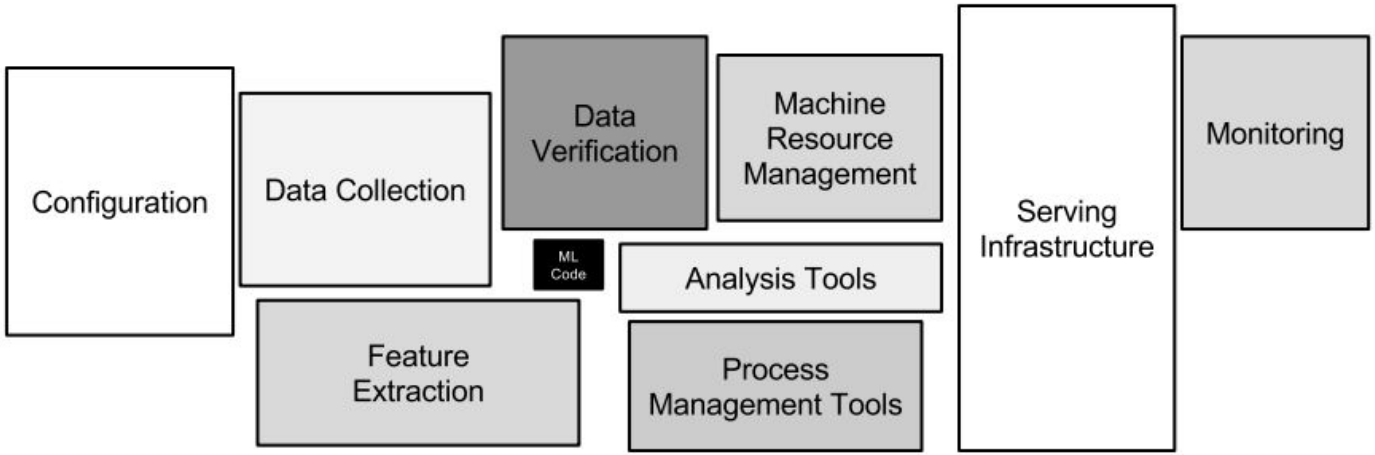
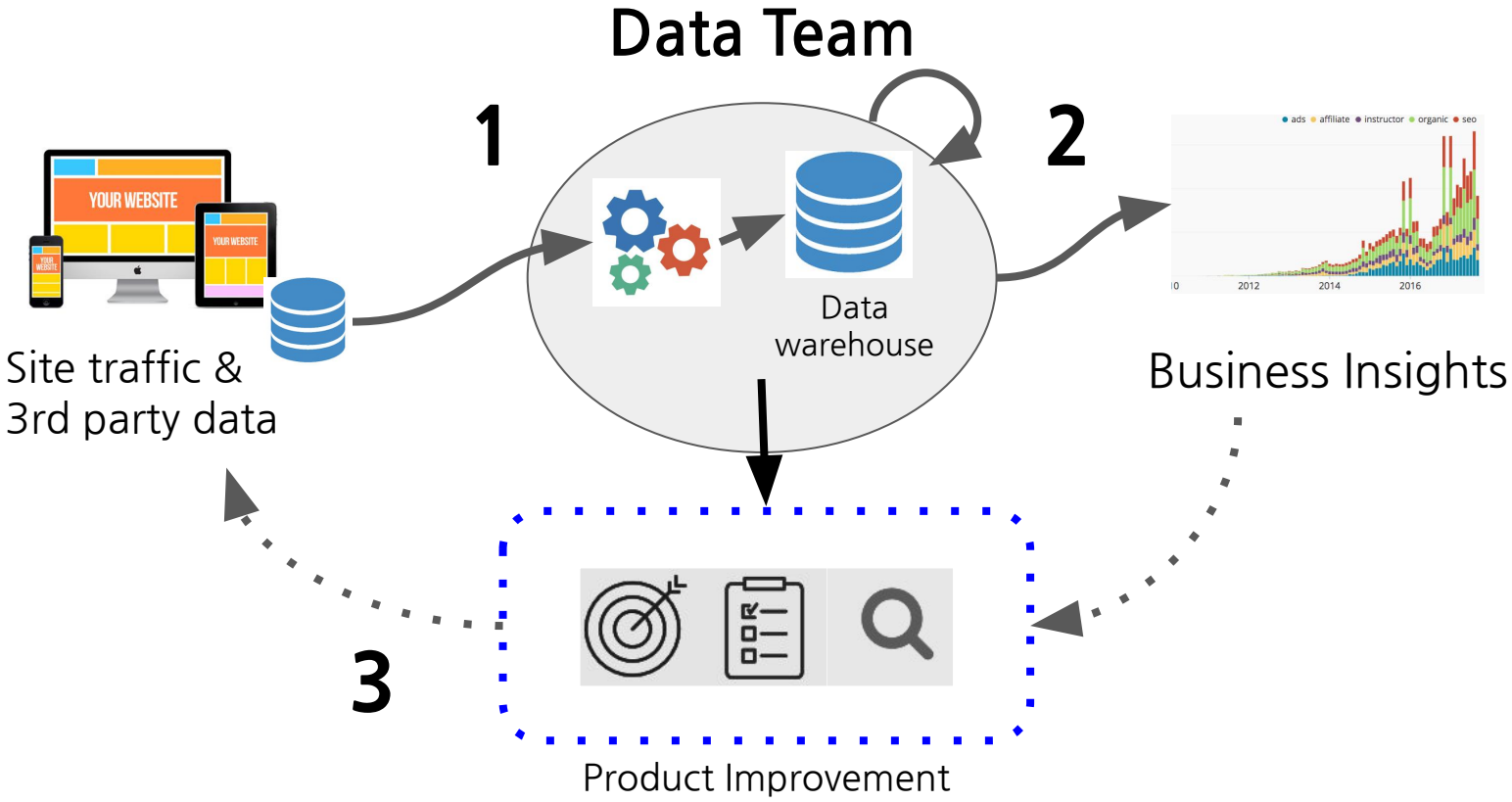


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

[출처] Hidden Technical Debt in Machine Learning Systems, NEURIPS, 2015

Data Flow


Data Team
1
2
3
YOUR WEBSITE
YOUR WEBSITE
Site traffic &
3rd party data
Data
warehouse
ads
affiliate
instructor
organic
seo
2012
2014
2016
Business Insights
Product Improvement

Data Team 삼대장

Engineer Analyst Scientist

Data Engineer (데이터 엔지니어)

- 데이터 인프라의 첫 단추
- 서비스 구석구석 흩어져 있는 데이터를 모으고 사용할 수 있는 형태로 정제해서, 데이터를 필요로 하는 사내 여러 곳에 공급하는 역할
- 데이터 파이프라인 관리, 데이터 웨어하우스 관리, 사내 데이터 툴 개발 (야근이 많은 편...)
- 필수 역량
  - Python/Spark/Java/Scala (Coding)
  - SQL/Hive
  - 클라우드 플랫폼
  - CS, 백엔드 지식

Data Analyst (데이터 분석가)

Analyst

- BI(Business Intelligence)를 책임지는 사람
- 회사의 규모에 따라 분석의 영역이 다름
  - 제품분석, 비즈니스분석, 마케팅분석 모두 하는 경우도 있고, 일부만 하는 경우도 존재
- 분석 도구(Google Analytics, Amplitude, ...)도 활용하고, SQL을 활용한 분석을 많이 하는 편
- 필수 역량
  - SQL/Python/Spark/
  - 통계, 수학 지식
  - 시각화 도구 사용
  - 도메인 지식
  - 약간의 Engineering 지식

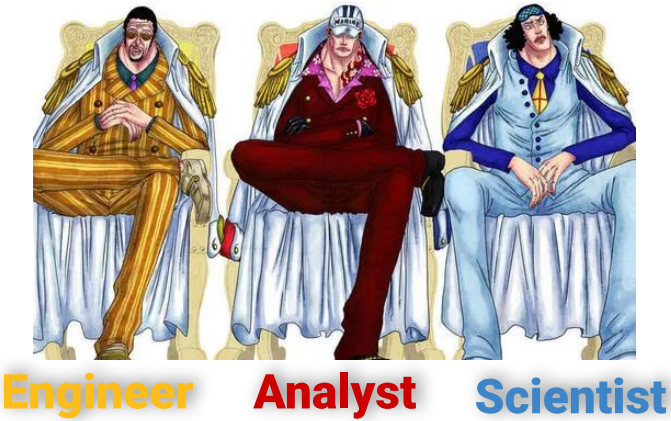
Data Scientist (데이터 사이언티스트)

Scientist

- ML, 알고리즘 모델링으로 서비스를 개선하는 역할
- 한국에선 Analyst와 Scientist 는 비슷한 느낌
  - Analyst 업무를 맡기는데 Scientist 라는 타이틀을 주는 경우도 존재
  - 나뉘는 경우, Scientist 는 주로 머신러닝 모델, 딥러닝 모델, 연구, 통계, 모델링을 조금 더 하는 편
- 비즈니스 임팩트를 갖는데 오랜 시간이 걸리기 때문에 끈기와 노력이 필요 (석사, 박사 학위를 선호하는 이유(?))
- 필수 역량
  - ML, DL 지식
  - 통계, 수학 지식
  - Python/Spark (Coding)
  - SQL/Hive

데이터 관련 직군

- 직군간의 교집합이 있음
  - 데이터 분석(Analyst)을 통해 모델을 만들고 예측가능(Scientist)
  - 예측에 필요한 데이터(Scientist)를 수집하기 위해 데이터 파이프라인 개발(Engineer)
- 공통적으로 요구하는 역량은 SQL, Python, 논리적 사고, 문제정의 능력
- 가장 중요한 건, 회사마다 정의하는게 다르기 때문에 채용공고 꼼꼼히 읽기



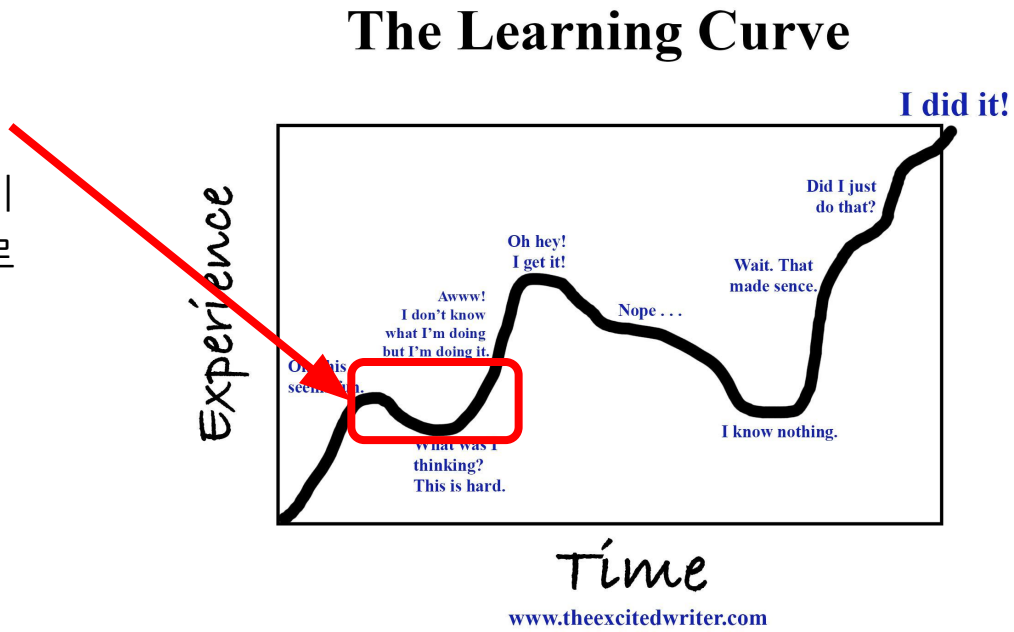
So What?

그래서 무엇을 할까?

어느 시기에 있나요?

여기서 어떻게 하는지가 중요

- 무엇을 모르는지 생각해보기
- 어디서 막혔는지 구체적으로 답변 가능해야함



회사에서 필요한 역량

- 기술적인 Python, SQL, ML활용 역량은 기본!
- 소프트 스킬, 그 중에서도 문제정의 능력이 중요함
  - 모델이 항상 좋은 것은 아니고, 서비스의 상황에 따라 룰베이스가 더 좋을 수 있음
  - A,B,C라는 문제의 해결 방식이 있다면, 각각의 장단점이 무엇이고 현재는 어떤 결정을 할지 고민해보기

=> 개인 프로젝트나 포트폴리오에 잘 녹여낸다면 Best

회사에서 필요한 역량

- 기술적인 Python, SQL, ML활용 역량은 기본!
- 소프트 스킬, 그 중에서도 문제정의 능력이 중요함
  - 모델이 항상 좋은 것은 아니고, 서비스의 상황에 따라 룰베이스가 더 좋을 수 있음
  - A,B,C라는 문제의 해결 방식이 있다면, 각각의 장단점이 무엇이고 현재는 어떤 결정을 할지 고민해보기

  => 개인 프로젝트나 포트폴리오에 잘 녹여낸다면 Best

- 어떤 공부를 해야할지 고민된다면
  - 채용공고 읽어보고 비교해보기 (공통적으로 요구하는 역량이나 기술스택이 있음)
  - 기술블로그 읽어보기

Data Analyst (검색/추천)

팀 소개

오늘의집 검색팀은 탐색 경험을 극대화 할 수 있도록 검색과 추천 서비스를 만드는 역할을 하고 있습니다.
검색 품질과 전환을 높이는 것을 목표로 하며, 이를 위해 데이터 기반의 논리적 사고를 바탕으로 업무를 수행합니다.
더 나은 품질과 기능을 위해 집요하게 한 걸음 더 파고들어 고민하며 탄탄한 서비스를 만들어 나가고 있습니다.
검색팀과 함께 No.1 Lifestyle Tech Company를 만들어 갈 수 있는 동료를 기다립니다.

📄 주요목표

- 컨텐츠와 상품을 분석하고, 사용자의 피드백을 찾아 연구하여 검색 결과 품질을 향상시킵니다.
- 검색어 교정, 필터/옵션 고도화를 통해 검색의 경로를 단순화하고 사용자의 고민을 줄입니다.
- 커머스와 컨텐츠 간의 정보 소비 탐색 경험을 자연스럽게 연결합니다.
- 사용자를 이해하고 분석하여 추천을 통한 새로운 발견 경험을 제공합니다.

📄 주요업무

- Data-driven 의사결정을 지원하기 위한 정량적 수치와 분석 결과 제공
- 프로덕트 핵심 지표 정의, 기획, 설계, 시각화
- 서비스 내/외부에 대한 종합적인 조사/분석을 통해 새로운 비즈니스 기회 발굴 및 전략 수립에 기여
- 데이터 기반 개선 과제 도출 및 실행
- 로깅, 추출, 분석 등 데이터 로그 설계와 관리
- 실험 설계 및 분석

💡 자격요건

- 가설을 세우고 통계적 근거를 기반으로 실무 데이터를 분석한 경험이 있으신 분
- 각 통계 검정 방법에 대한 기본 이해와 이를 바탕으로 한 가설 검증 경험이 있으신 분
- 로그 데이터를 분석에 용이한 형태로 추출 및 재가공 가능한 수준의 SQL, Python 스킬을 갖추신 분
- 검색, 추천 서비스 경험 및 품질 관련 업무와 관련이 있으신 분

👍 우대사항

- 내 분야와 담당이 아니더라도 같이 논의하고 협업하는 것을 좋아하시는 분
- 서비스 개선을 위한 실험 설계 및 분석 경험(A/B Test 설계 분석)이 있으신 분
- 데이터 시각화 툴 사용 경험(Tableau, Looker, Redash 등)이 있으신 분
- 검색/추천 관련 도메인 지식 또는 업무 경험이 있으신 분
- 인테리어 산업에 대한 관심과 이해도가 높으신 분

지원 및 진행 절차

- **지원서류 : 자유양식의 이력서(필수), 포트폴리오 및 커버레터(선택)** / PDF 형식 권장
  - 연봉, 신체 정보, 가족 사항 및 주민번호 등의 민감한 개인정보는 제외 부탁드립니다.
- **진행절차 : 서류전형 > SQL Test > 직무 인터뷰 > 조직문화 인터뷰 > 처우협의 > 최종합격**
  - 지원자의 이력 및 경력 사항에 따라 일부 면접 과정이 생략되거나 추가될 수 있습니다.
- 진심과 정성을 담는 오늘의집 합류 여정 자세히 알아보기 🔍

Data Scientist, Decision

당근 · 정규직 · 경력

영입정보 지원하기

데이터 가치화 팀을 소개해요

당근 팀은 동네 안에서 연결되지 못한 가치있는 정보를 발견하고, 지역 생활 속의 불편함을 해결하기 위해 모였어요. 이러한 사용자 가치를 만들어내기 위해서는 사용자들에 대한 믿을 수 있는 정보를 손쉽게 접근해서 의사결정에 반영할 수 있어야 해요. 당근은 의사결정에 수많은 데이터를 이미 활용하고 있지만, 당근의 데이터의 가치를 극대화하기 위해서는 많은 변화가 필요해요. 데이터 가치화 팀은 "매일 데이터를 통해 사용자를 위한 의사결정을 해요."라는 비전을 달성하기 위해 모였어요. 이 가슴 뛰는 비전을 달성하기 위해 Data Scientist, Decision과 Data Engineer, Server Engineer로 이루어진 목적 조직을 만들었고, 데이터 가치화의 문제를 주도적으로 해결하기 위해서 고민하면서 일해요.

🥕 데이터 가치화 팀이 어떻게 일하는지 구경하기

🥕 당근에서 실험을 어떻게 하는지 구경하기

Data Scientist, Decision을 소개해요

1800만의 사용자가 방문하는 서비스인 당근의 데이터는 방대한 규모이면서 정형화되어 있지 않아요. 사용자를 위한 의사결정을 누구나 할 수 있으려면 이 데이터에서 사용자에 대한 의미있는 정보를 얻어내야 해요. 따라서 Data Scientist, Decision은 어떻게 하면 다양한 직군의 팀원들이 사용자에 대한 유용한 정보를 스스로 손쉽게 얻어낼 수 있을까 고민해요. 또한 어떻게 하면 의사결정을 할 때 발생할 수 있는 편향을 체계적으로 없앨 수 있을까 고민해요. Data Scientist, Decision은 이러한 고민에 대해 주도적으로 문제를 정의하고 통계, 머신러닝, 엔지니어링, 문제 정의와 논리적 사고라는 도구를 활용해요. 궁극적으로는 누구나 데이터 사이언티스트처럼 과학적 의사결정을 일상적으로 할 수 있게 하기 위해 매일 노력하고 있어요.

이런 일을 해요

- 전사적으로 중요한 질문을 데이터 관점에서 정의하고 의사결정의 방향성을 보여줘요
- 의사결정과 제품의 개선이 사용자에게 어떠한 영향을 주는지 누구나 파악할 수 있도록 해요
- 전사의 서로 다른 팀이 사용자라는 같은 방향을 바라볼 수 있도록 데이터로 기여해요
- 지표, 분석, 실험을 쌓일 수 있는 형태로 만들어서 전사의 누구나 보고 자신의 업무나 의사결정에 활용할 수 있도록 해요.
- A/B 테스트와 같은 방법을 통해 의사결정 과정에서 발생하는 편향을 체계적으로 제거해요
- 누구나 의사결정을 내리기 위해 필요한 데이터를 스스로 볼 수 있도록 해요

이런 분을 찾고 있어요

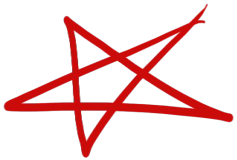
- 3년 이상의 데이터 도메인에서 임팩트를 낸 경험이 있으신 분
- 복잡한 문제에 대해서 문제정의를 잘 할 수 있는 분
- 통계를 활용해서 데이터나 의사결정 과정에서 발생할 수 있는 편향을 제거할 수 있는 분
- 데이터를 통해 얻은 정보를 누구나 이해할 수 있는 형태로 설명할 수 있는 분
- A/B 테스트와 인과 추론에 대한 깊은 통계적 이해를 가지신 분

이런 분이면 더 좋아요!

- 비정형 데이터를 다양한 기술을 활용해서 유의미한 정보로 만들 수 있는 분
- 정말 큰 양의 데이터를 Airflow나 DBT와 같은 도구를 활용해 효율적으로 분석에 활용할 수 있는 분
- 코드를 통해 사람들이 사용할만한 가치있는 것을 만들어본 경험이 있으신 분
- 프로덕트를 만드는 과정에 직접적으로 참여해서 어떻게 프로덕트를 만들어가는지 잘 이해하시는 분

우선 해보는 것!

+



꾸준히 하는 것!

감사합니다