AI기반 데이터 분석 및 AI Agent 개발 과정

『1과목 : 』
AI기반 데이터 분석

2025.09.22-10.02(9일, 62시간)

Prepared by DaeKyeong K

Ph.D.

삼성전자 구미사업부 G.제조팀

HS Global
HARDWARE SOFTWARE

목차

『1-2』
조사 및 데이터 수집 방법

조사란?

학습 개요

학습목표

- 이 워크샵에서는 조사에 대해 알 수 있습니다.

눈높이 체크

- 조사가 필요했던 상황은?
- 어떤 방법을 사용하셨나요?
- 어떤 어려움이 있었나요?
- 결과를 어떻게 활용하셨나요?

1. 조사의 정의

조사란?

- 특정한 문제나 질문에 대한 답을 찾기 위해
- 체계적이고 과학적인 방법을 사용하여
- 정보를 수집하고 분석하는 과정
- 단순한 정보 수집 ≠ 조사
- 명확한 목적과 방법론 + 신뢰할 수 있는 결과 = 조사

1. 조사의 정의

조사의 핵심 특징

- 4가지 핵심 원칙

특징	설명
체계성	논리적이고 순서적인 절차를 따름
객관성	편견을 배제하고 중립적인 관점 유지
과학성	검증 가능하고 재현 가능한 방법 사용
목적성	명확한 목표와 가설을 설정

1. 조사의 정의

조사의 중요성

- **효과적인 의사결정**
  - 추측이나 직감이 아닌 데이터 기반 결정
  - 위험 요소 사전 파악 및 대비

- **문제 해결의 정확성**
  - 문제의 원인과 해결책을 객관적으로 파악
  - 효과적인 대안 제시

- **자원의 효율적 활용**
  - 시간, 비용, 인력의 최적화
  - 투자 대비 효과 극대화

2. 조사 프로세스

조사 프로세스 (6단계)

- **체계적인 조사 진행 과정**

1단계: 문제 정의
2단계: 조사 설계
3단계: 자료 수집
4단계: 자료 분석
5단계: 해석 및 결론
6단계: 보고서 작성

- 각 단계는 순차적이면서도 상호 연결되어 있음

2. 조사 프로세스

1단계: 문제 정의 및 목적 설정

- **조사의 출발점**
  - **핵심 활동**
    - 해결하고자 하는 문제를 명확히 정의
    - 조사의 목적과 범위 설정
    - 연구 질문(Research Question) 수립
    - 가설 설정 (필요시)

  - **성공의 열쇠**
    - 문제가 명확해야 해답도 명확해진다

2. 조사 프로세스

2단계: 조사 설계

- **조사의 청사진 작성**
  - **주요 결정사항**
    - **조사 방법론 선택 (정량적/정성적)**
    - **표본 설계 및 대상 선정**
    - **자료 수집 방법 결정**
    - **조사 도구 개발**
  - **설계 품질 = 결과 품질**

2. 조사 프로세스

3-4단계: 자료 수집 및 분석

- **3단계: 자료 수집**
  - 설계된 방법에 따라 데이터 수집
  - 품질 관리 및 검증
  - 윤리적 고려사항 준수

- **4단계: 자료 분석**
  - 수집된 데이터 정리 및 검토
  - 통계적 분석 또는 질적 분석 수행
  - 패턴 및 트렌드 파악

2. 조사 프로세스

5-6단계: 해석 및 활용

- **5단계: 해석 및 결론 도출**
  - 분석 결과 해석
  - 가설 검증
  - 결론 및 시사점 도출

- **6단계: 보고서 작성 및 활용**
  - 조사 결과 보고서 작성
  - 의사결정자에게 결과 전달
  - 후속 조치 및 개선방안 제시

3. 조사 분류

조사 목적에 따른 분류

- 3가지 주요 유형

유형	목적	특징	주요 방법
탐색적 조사	기초적 이해, 가설 생성	새로운 현상 탐구	질적 방법론
기술적 조사	현황 파악, 실태 기술	체계적 현상 기술	설문조사, 관찰법
설명적 조사	인과관계 규명	가설 검증 중심	실험, 종단연구

3. 조사 분류

정량적 vs 정성적 조사

- 두 가지 접근 방식

구분	정량적 조사	정성적 조사
목적	객관적 측정, 일반화	심층적 이해, 맥락 파악
데이터	수치화 가능한 자료	언어적, 행동적 자료
표본 크기	대규모	소규모
분석 방법	통계적 분석	내용 분석, 해석
결과	일반화 가능	깊이 있는 통찰

3. 조사 분류

정량적 조사 방법의 주요 방법론

- 설문조사 (Survey)
  - 표준화된 질문을 통한 대규모 데이터 수집
  - 높은 일반화 가능성

- 실험 (Experiment)
  - 변수 조작을 통한 인과관계 규명
  - 높은 내적 타당성

- 2차 자료 분석
  - 기존 데이터 활용
  - 시간과 비용 절약

3. 조사 분류

정성적 조사 방법의 주요 방법론

- 심층 인터뷰 (In-depth Interview)
  - 개별 면담을 통한 깊이 있는 정보 수집
  - 포커스 그룹 (Focus Group)
  - 집단 토론을 통한 다양한 관점 파악

- 참여관찰 (Participant Observation)
  - 현장에서 직접 관찰하며 자료 수집

- 사례연구 (Case Study)
  - 특정 사례에 대한 심층적 분석

4. 조사 설계 고려사항

타당성 (Validity)

- **내적 타당성**
  - 측정하고자 하는 것을 정확히 측정하는가?
  - 조사 도구의 정확성

- **외적 타당성**
  - 결과를 다른 상황에 일반화할 수 있는가?
  - 연구 결과의 적용 범위

4. 조사 설계 고려사항

신뢰성 (Reliability)

- **일관성의 원리**
  - 동일한 조건에서 반복 측정시 일관된 결과
  - 측정 도구의 안정성

- **표본 설계**

- **대표성 확보**
  - 모집단 정의 및 표본 추출 방법
  - 적절한 표본 크기 결정

4. 조사 설계 고려사항

윤리적 고려사항

- **참여자 권리 보호**
  - 자발적 참여 보장
  - 충분한 정보 제공

- **개인정보 보호**
  - 익명성과 기밀성 보장
  - 데이터 보안 관리

- **연구윤리 준수**
  - 허위 정보 방지
  - 이해관계 충돌 회피

5. 성공적인 조사를 위한 체크리스트

체크리스트

- **준비 단계**
  - □ 문제 정의가 명확한가?
  - □ 조사 목적이 구체적인가?
  - □ 적절한 방법론을 선택했는가?

- **실행 단계**
  - □ 표본이 대표성을 가지는가?
  - □ 자료 수집이 체계적으로 이루어지는가?
  - □ 윤리적 기준을 준수하고 있는가?

- **마무리 단계**
  - □ 분석이 객관적으로 수행되었는가?
  - □ 결론이 데이터에 근거하는가?
  - □ 실행 가능한 제안을 제시했는가?

6. 실습

실습1 : ChatGPT를 이용한 시장조사

문제 :
우리 회사는 식품회사다, 건강기능식품에 대한 신제품을 출시하기 위해 소비자 요구, 경쟁 상황, 시장 트렌드 등을 파악하고 싶다.

『1-3』
데이터 전처리

데이터

데이터 전처리

데이터 수집, 인제스트

분석 주제 탐색 및 문제해결 단계별 접근

데이터 확인 및 검증

결측값/데이터 분포/이상치

학습 개요

학습목표

- 이 워크샵에서는 데이터에 대해 알 수 있습니다.

눈높이 체크

- DIKW 피라미드를 알고 계신가요?

1. 데이터

데이터의 정의

- 라틴어인 dare(주다)의 과거분사형으로 '주어진 것 ' 이라는 의미.
  - 데이터(data)라는 용어는 1646년 영국 문헌에 처음 등장.
  - 1940년 이후 컴퓨터 시대 시작과 함께 자연과학뿐만 아니라 경영학, 통계학 등 다양한 사회과학이 진일보하며, 데이터의 의미는 과거의 관념적이고 추상적인 개념에서 기술적이고 사실적인 의미로 변화
  - 데이터는 추론과 추정의 근거를 이루는 사실-옥스퍼드 대사전
  - 데이터는 단순한 객체로서의 가치뿐만 아니라 다른 객체와의 상호관계 속에서 가치를 갖는 것

1. 데이터

데이터 특성

구분	특성
존재적 특성	객관적 사실
당위적 특성	추론, 예측, 전망, 추정을 위한 근거

데이터의 유형

구분	유형
정성적 데이터	언어, 문자 형태의 데이터 (회사 매출의 증가 등)
정량적 데이터	수치, 도형, 기호 형태의 데이터 (나이, 몸무게, 주가 등)

1. 데이터

지식경영의 핵심 이슈

구분	내용
암묵지	학습과 경험을 통해 개인에 체화된 지식, 공유와 전달의 어려움 (내면화->공통화 필요)
형식지	문서나 메뉴얼처럼 형식화된 지식, 공유와 전달이 용이 (표출화->연결화 필요)

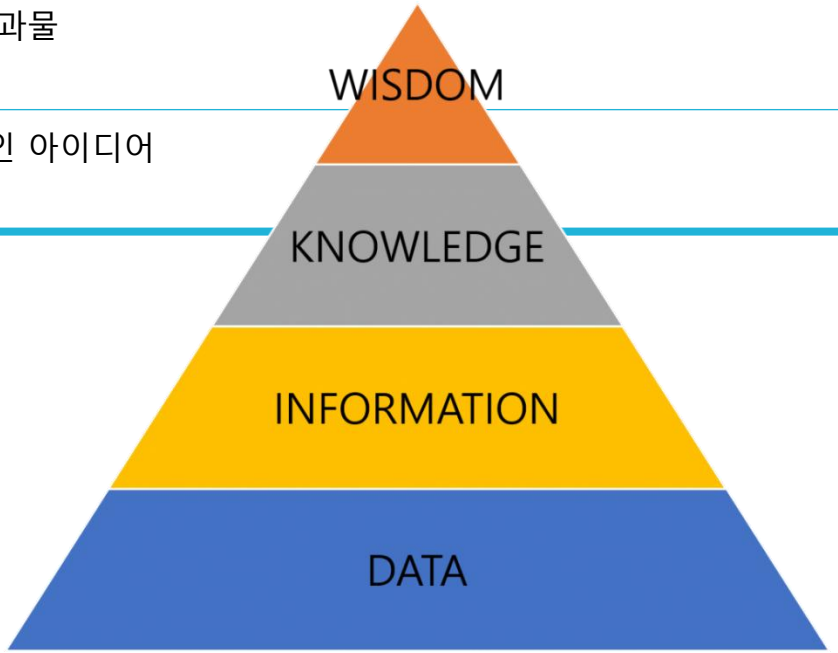
形式知 例 ：Coding Convention

1. 데이터

데이터와 정보의 관계

구분	내용	
데이터	가공하기 전의 순수한 수치나 기호 관찰이나 측정을 통해 수집된 사실이나 값	
정보	데이터의 가공 및 상관관계간 이해를 통해 패턴을 인식하고 그 의미를 부여 데이터 상호간 관계-의미 개념화(체계화)	집계: 데이터를 있는 그대로 정리하는 일. 사실의 정리 분석: 그대로의 사실과 그 사실로부터 도출될 가치를 발견해 가는 일
지식	상호 연결된 정보 패턴을 이해하여 이를 토대로 예측한 결과물 정보의 이용에 대한 노하우	
지혜	근본 원리에 대한 깊은 이해를 바탕으로 도출되는 창의적인 아이디어 지식을 활용하는 창의적 아이디어-의사결정	

DIKW 피라미드



1. 데이터

자료의 종류 LOTS(Lifetime, Observational, Test , Self-Reported)

- **L 자료: 생애 데이터**
  - 한 대상의 통사적 정보를 알 수 있는 자료
  - 특히 특정 개인을 대상으로 한 임상 장면에서 많
    이 사용
  - 생활기록부, 범죄이력, 신용정보, 졸업증명, 병력 조회 등이 이에 해당
  - 객관화된 자료이지만, 이용에 한계가 존재

- **T 자료: 검사 데이터**
  - 실험적 절차를 거치거나 표준화된 검사를 통해 얻어진 데이터
  - 대중매체에서 과학자 인물들이 손에 들고 있는 도표들도 대부분 T-자료
  - 가장 객관적이고 질 좋은 자료이지만, 현실적으 로 접해보기는 그다지 쉽지 않음
  - 자료를 확보하는 과정에서의 연구윤리 문제도 개
    입

- **O 자료: 관찰 데이터**
  - 숙련된 관찰자 혹은 대상을 잘 아는 관계자, 친지 등이 제공하는 자료
  - 면접법, 참여관찰법 등을 통해 확보 가능
  - 주변 사람들의 증언이나 CCTV 영상 자료 역시 O-자료에 속함

- **S 자료: 자기보고 데이터**
  - 어떤 대상에 대한 정보를 얻을 때 그 대상에게 직 접 물어보아 얻은 자료
  - 당연히 사람을 대상으로 하므로, 그 분야는 심리
    학이나 사회학 등에 한정될 수밖에 없음
  - 매우 흔하게 접할 수 있는 자료로, 흔한 설문조사 나 여론조사 등을 통해 얻어짐
  - "사람은 자신이 자신을 제일 잘 안다" 는 전제에 기초해 있으며, 사회적 선망에 의해 답변이 왜곡

1. 데이터

Data set

- 데이터 모음
  - 하나의 데이터베이스 테이블의 내용이나 하나의 통계적 자료 행렬과 일치
- 컬럼column: 특정한 변수를 대표
- 로우row: 주어진 멤버와 일치
  - 변수 개개의 값들을 나열하고, 각각의 값은 데이터라고 부름
  - 하나 이상의 멤버에 대한 데이터를 이루며, 로우의 수와 일치
  - 웹에서 접근하고 다운로드 할 수 있는 다양한 형태의 데이터 세트가 존재

Id	Duration(hrs)	# Packets	#NetFlows	Size	Bot	#Bots
1	6.15	71,971,482	2,824,637	52GB	Neris	1
2	4.21	71,851,300	1,808,123	60GB	Neris	1
3	66.85	167,730,395	4,710,639	121GB	Rbot	1
4	4.21	62,089,135	1,121,077	53GB	Rbot	1
5	11.63	4,481,167	129,833	37.6GB	Virut	1
6	2.18	38,764,357	558,920	30GB	Menti	1
7	0.38	7,467,139	114,078	5.8GB	Sogou	1
8	19.5	155,207,799	2,954,231	123GB	Murlo	1
9	5.18	115,415,321	2,753,885	94GB	Neris	10
10	4.75	90,389,782	1,309,792	73GB	Rbot	10
11	0.26	6,337,202	107,252	5.2GB	Rbot	3
12	1.21	13,212,268	325,472	8.3GB	NSIS.ay	3
13	16.36	50,888,256	1,925,150	34GB	Virut	1

출처 : Google Dataset: https://toolbox.google.com/datasetsearch Google AI Dataset: /tools/hniruja/datasets/https://ai.google

1. 데이터

데이터 형태

- 질적자료(정성적자료, Qualitative or Categorical): 범주 또는 순서 형태의 속성을 가지는 자료
  - 범주형(명목형, nominal) 자료: 사람의 피부색, 성별
  - 순서형(서수형, ordinal) 자료: 제품의 품질, 등급, 순위
- 양적자료(정량적자료, Quantitative or Numeric): 관측된 값이 수치 형태의 속성을 가지는 자료
  - 범위형interval 자료: 화씨, 섭씨와 같이 수치 간에 차이가 의미를 가지는 자료.
  - 비율ratio 자료: 무게와 같이 수치의 차이 뿐만 아니라 비율 또한 의미를 가지는 자료

1. 데이터

데이터 마트 개발

- 데이터 전처리(Data Preprocessing)와 데이터 웨어하우스의 관계:

  - 데이터 전처리는 데이터 분석을 위해 데이터를 준비하고 정리하는 과정으로, 데이터 웨어하우스에 저장되는 데이터의 일관성과 품질을 유지하는 데 중요한 역할을 함.
  - 데이터 전처리는 데이터의 클린징, 변환, 통합, 축소 등의 과정을 포함하여 데이터를 분석 가능한 형태로 만듬.
  - 데이터 전처리는 데이터 웨어하우스에 저장된 데이터를 정제하고 가공하여, 비즈니스 분석 및 의사 결정에 활용할 수 있는 형태로 만듬.

- 따라서 데이터 웨이 하우스에 저장된 데이터는 데이터 전처리를 거치고 정제되어 분석 가능한 상태로 유지.

1. 데이터

데이터 마트 개발

- **데이터마트**
  - 데이터 웨어하우스와 사용자의 중간층에 위치한 것. 하나의 주제 또는 하나의 부서중심의 데이터 웨이 하우스 라고 할 수 있음.
  - 데이터마트 내 대부분 데이터는 데이터 웨이 하우스로부터 복제되지만, 자체적으로 수집될 수 있으며, 관계형 데이터베이스나 다차원 데이터베이스를 이용하여 구축.

2. 데이터베이스

데이터베이스 용어의 연혁-해외

출처	내용
1950년대	미국 정부의 자국 군대의 군비 상황을 집중 관리하기 위해서 컴퓨터를 활용한 도서관의 개념으로 등장, 데이터의 기지(Base)라는 뜻의 의미
1963년 6월	미국 SDC(System Development Corporation)에서 개최한 "컴퓨터 중심의 데이터베이스 개발과 관리(Development and Management of a Computer-centered Data Base)라는 주제의 심포지엄(symposium)이었으며, 발표는 "대량의 데이터를 축적하는 기지" 단계에 머물렀다.
1963년	GE(General Electronic)의 C. 바크만(Charles Bachman)이 최초의 데이터베이스 관리 시스템인 IDS(Integrated Data Store)를 개발
1965년	2차 심포지엄에서는 "체계적 관리와 저장"등의 의미가 포함된 데이터베이스 시스템(Database System) 용어가 등장하기 시작했다. 이는 현대의 데이터베이스 관리(DBMS)와 유사한 개념
1970년대 초반	유럽에서 데이터베이스라는 단일어가 일반화
1970년 후반	미국 주요 신문에서 흔히 사용

2. 데이터베이스

데이터베이스 용어의 연혁-국내

출처	내용
1975년	미국의 CAC(Chemical Abstracts Condensates)가 KORSTIC(韓國科學機術情報研究院, Korea Institute of Science and Technology, 한국과학기술정보센터)를 통해 서비스되며 데이터베이스 이용을 시작한다. 다만 온라인 형태가 아닌 자기 테이프 형태로 배치 방식으로 제공
1980년	INSPEC이나 COMPENDEX와 같은 해외 데이터베이스를 확충하여 TECHNOLINE이라는 온라인 정보검색 서비스 개시
1980년 중반	국내의 데이터베이스 연구 개발 진행

2. 데이터베이스

데이터베이스의 정의

출처		내용
1차개념확대, 정형데이터 관리	EU의 데이터베이스의 법적 보호에 관한 지침	체계적이거나 조직적이로 정리되고 전자식 또는 기타 수단으로 개별적으로 접근할 수 있는 독립된 저작물, 데이터 또는 기타 소재의 수집물
	국내 저작권법	소재를 체계적으로 배열 또는 구성한 편집물로서 개별적으로 그 소재에 접근하거나 그 소재를 검색할 수 있도록 한 것
2차개념확대, 비정형 데이터 포함	국내 컴퓨터 용어사전	동시에 복수의 적용 업무를 지원할 수 있도록 복수 이용자의 요구에 대응해서 데이터를 받아들이고 저장, 공급하기 위하여 일정한 구조에 따라서 편성된 데이터의 집합
	국내 Wikipedia	관련된 레코드의 집합, 소프트웨어로는 데이터베이스관리시스템(DBMS)
	데이터분석 전문가 가이드	문자, 기호, 음성, 화상, 영상 등 상호 관련된 다수의콘텐츠를 정보 처리 및 정보통신 기기에서 의하여 체계적으로 수집/축척하여 다양한 용도와 방법으로 이용할 수 있도록 정리한 정보의 집합체

2. 데이터베이스

데이터베이스의 일반적인 특징

출처	내용
통합된 데이터(Integrated Data)	동일한 내용의 데이터가 중복되어 있지 않다는 것을 의미 데이터의 중복은 관리상의 복잡한 부작용을 초래
저장된 데이터(Stored Data)	컴퓨터가 접근할 수 있는 저장 매체에 저장되는 것을 의미 데이터베이스는 기본적으로 컴퓨터 기술을 바탕으로 함
공용 데이터(Shared Data)	여러 사용자가 서로 다른 목적으로 데이터를 공동으로 이용한다는 의미 대용량화되고 구조가 복잡한 것이 보통
변화되는 데이터(Changable Data)	데이터데이스에 저장된 내용은 곧 데이터베이스의 현 상태를 나타내는 것을 의미. 다만 이 상태는 데이터의 삽입, 기존 데이터의 삭제, 갱신으로 항상 변화하면서도 항상 현재의 정확한 데이터를 유지해야 함.

2. 데이터베이스

데이터베이스의 다양한 측면에서의 특징

출처	내용
정보 축적 및 전달 측면	- 기계가독성 : 정보처리기기가 읽고 쓸 수 있음 - 검색가독성 : 다양한 방법으로 정보 검색 - 원격조작성 : 정보통신망 이용 온라인 이용
정보 이용 측면	- 이용자의 요구에 따른 정보 신속 획득 - 원하는 정보를 정확, 경제적 찾아냄
정보 관리 측면	- 질서와 구조에 따라 정리, 저장, 검색, 관리하며 방대한 양의 정보 축적 - 새로운 내용의 추가, 갱신 용이
정보기술 발전 측면	- 정보처리, 검색/관리 소프트웨어 및 하드웨어, 전송을 위한 네트워크 기술 발전 견인
경제 산업 측면	- 인프라 특성으로 경제, 산업, 사회 활동의 효율성 제고와 편의증진 수단

2. 데이터베이스

기업내부 데이터베이스

- 정보통신망 구축 가속화로 1990년대부터 기업내부 데이터베이스(인하우스 DB)에 기업의 모든 자료를 연계하며 경영 활동의 기반이 되는 전사 시스템으로 확대하였다.
- 1990년 중반 이전에는 정보의 수집과 공유를 위한 MIS(경영정보시스템), 생산자동화, 통합자동화 등 즉 OLTP(Online Transaction Processing)의 주축이 되었다면, 이후에는 수집에서 벗어나 분석이 중심이 되는 OLAP(Online Analytical Processing) 시스템 구축을 하게 되었다.

OLTP

- 실시간성과 트랜잭션 처리, 데이터 하나하나가 중요한 핵심 시스템

OLAP

- 통계와 분석 등을 위해서 만들어진 DB아키텍쳐

2. 데이터베이스

기업내부 데이터베이스

- DB를 설계할 때 테이블에 설정을 하게 되는데 해당 테이블이 트랜잭션이 중심이 되는 보편적인 데이터들이 쌓이는 곳이라면, OLTP성으로 구현되며 분석과 관련된 서비스라면 OLAP을 설정할 수 있는 기능들이 DBMS에서 제공이 된다.
- 2000년에 들어오면서 기업 DB 구축의 화두는 CRM(Costomer Relationship Management, 고객관계관리)과 SCM(Supply Chain Management, 공급망관리)였다.CRM은 고객을 새로 얻는 것보다, 이탈을 하지 않는 것이 중요하다 여겨지면서 생긴 시스템으로 2000년대에 갑자기 붐처럼 여기저기 CRM 시스템을 구축하기 시작했었고, SCM은 최적의 공급과 유통들을 위해서(정확히 말해서 최적의 시간 단축을 위해서) 구축되어진 시스템이다.

2. 데이터베이스

각 분야별 내부 데이터베이스

- 제조부분
  - 2000년 기점으로 부품이나 재고관리의 DB 활용에서 설계, 제조, 유통 전 공정을 포함하는 범위로 확대
  - 초기 기업별 고유 시스템 -> 솔루션 유형으로 발전
  - ERP(Enterprise Resource Planning) -> SCM으로 기능 확장. 이 변화는 대기업 중심으로 이루어졌는데 SCM이 기업의 협력사를 관리하는 것도 포함이 되서이다. 즉, 생산이 제대로 이루어지는지 유통이 제대로 이루어지는지 제대로 감시 및 관리하기 위해서 만들어졌다라고 생각하면 된다.
  - 2000년 중반 이후에는 실시간 기업(RTE, Real-Time Enterprise)이 화두였으며 대기업-중소기업 간의 협업적 IT화의 비중이 확대되었다.
  - 최근에는 DW(Data Warehouse), CRM, BI(Business Intelligence) 등의 DB 구축이 주류를 이루고 있음

2. 데이터베이스

각 분야별 내부 데이터베이스

- 금융부분
  - 1998년 IMF 이후, 부실 타파 위해 통합 시스템 구축 등이 크게 확산
  - 2000년 초반, EAI(Enterprise Application Integration), ERP, e-CRM 등 시스템 통합, 정보 공유, 고객 정보의 전략적 활용이 주된 테마
  - 2000년 중반, DW를 적극적 도입, 인터넷뱅킹 정착, 방카슈랑스 도입
  - 최근, 차세대 프로젝트, 다운 사이징, 바젤 II 등 대형 프로젝트 마무리, 향후 EDW(Enterprise Data Warehouse) 확장 예상
  - 참고로 다운 사이징이라는 의미는 점포를 줄이는 다운 사이징이라는 말도 있지만, 기존의 금융권들은 IBM 메인 프레임이라는 시스템을 중심으로 업무를 운영해왔었는데 이 시스템을 유닉스 등으로 대체해서 비용을 줄이는 것을 다운 사이징이라고 한다.(정확히는 메인프레임 다운 사이징)

2. 데이터베이스

각 분야별 내부 데이터베이스

- 유통부분- 2000년 이후, 유통 채널이 늘면서 CRM의 중요성과 유통과 가장 밀접한 SCM 구축이 주를 이루었음
  - 이외에 상거래를 위해 각종 인프라, KMS(Knowledge Management System)을 위한 백업 시스템 구축
  - 2000년 중반, BSC(Balanced ScoreCard, 균형성과표), KPI(Key Performance Indicator, 핵심성과지표), 웹리포팅 등 툴을 기존 DB와 연계
  - 최근 RFID(Radio-Frequency Identification, 전자태그) 등장으로 사물인터넷(IoT), 유비쿼터스의 핵심기술로 주목되고 있으며 이를 지원하는 대용량 데이터베이스 플랫폼 요구

2. 데이터베이스

분야별 사회기반 구조의 데이터베이스

- 물류부문
  - 물류부문은 '실시간 차량추적'이라고 할 수 있는 종합물류망이 대표적종합물류망은 CVO(Commercial Vehical Operation System, 화물운송정보), EDI 서비스, 데이터베이스 서비스(물류정보), 부가서비스로 구성종합물류망에 해양수산부의 PORT-MIS(항만운영 정보시스템), 철도청의 KROIS(철도운영정보 시스템), 복합화물터미널망, 항공정보망, 민간기업물류 VAN(Value Added Network)를 연결

2. 데이터베이스

분야별 사회기반 구조의 데이터베이스

- 지리/교통부문
  - 1995년 시작된 국가지리정보체계(NGIS) 구축은 국가지형도와 공통주제도, 지하매설물도를 전산화하여 기본 공간정보 데이터베이스를 구축하고 관련 기술 개발과 함께 범국가적인 활용을 위한 국가 표준설정과 활용체계를 개발하는 사업
  - 교통정보는 동적(실시간) 교통정보와 정적(비실시간) 교통정보로 나뉘며, 실시간 교통정보는 ITS, 비실시간 교통정보는 교통정책 및 계획 수립 등에 필요한 교통 분야별 기초자료 및 통계를 제공하는 DB를 말함
    - GIS(Geographic Information System, 지리정보시스템)
    - RS(Remote Sensing, 원격탐사)
    - GPS(Global Positioning System, 글로벌 포지셔닝 시스템 혹은 범지구위치결정 시스템)
    - ITS(Intelligent Transport System, 지능형 교통시스템)
    - LBS(Location Based Service, 위치기반 서비스)
    - SIM(Spatial Information Management, 공간정보관리)

2. 데이터베이스

분야별 사회기반 구조의 데이터베이스

- 의료부문
  - 1996년부터 53개 기관을 대상으로 의료EDI 상용서비스가 제공
    - PACS(Picture Archiving and Communications System, 영상처리 시스템)
    - u헬스(ubiquitous-Health)
    - ABC, BSC, 6시그마 등의 경영기법이 주요 병원을 필두로 도입

- 교육부문
  - 1단계 교육정보화종합계획(1997~2000)은 정보 소양 교육을 위주로 진행
  - 2단계 사업(2001 ~ 2005)은 첨단 정보통신기술을 활용한 각종 교육정보의 개발 및 보급, 정보 활용 교육, 대학 정보화 및 교육 행정정보화 사업 추진
    - NEIS(Nation Education Information System, 교육행정정보시스템)

『1-3』
데이터 전처리

데이터

데이터 전처리

데이터 수집, 인제스트

분석 주제 탐색 및 문제해결 단계별 접근

데이터 확인 및 검증

결측값/데이터 분포/이상치

Syllabus

학습목표

- 이 워크샵에서는 데이터 분석을 실행하기 전 데이터의 표현과 Data pre-processing 을 가능케 할 것입니다. Data Scientist의 기본역량인 Data 처리 및 AI 활용 역량을 가질 수 있습니다.

눈높이 체크

- Data pre-processing 을 알고 계신가요?

1. Data pre-processing

Data pre-processing

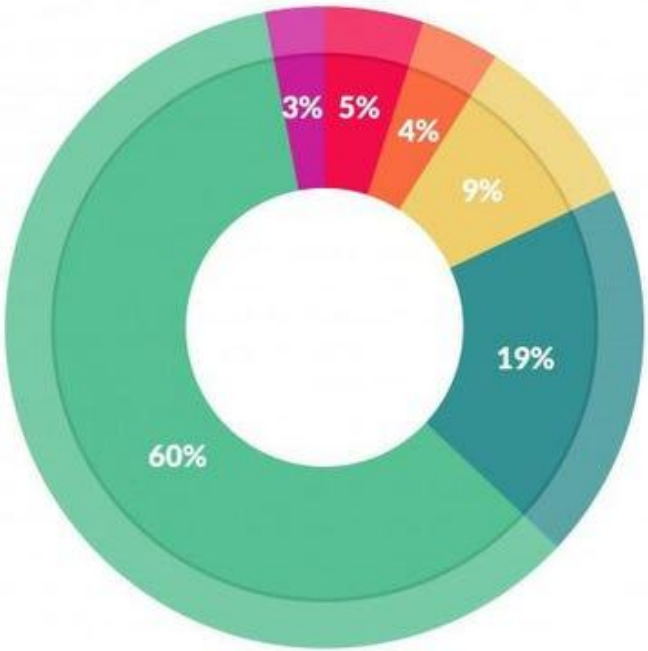
- Data pre-processing?
  - Data pre-processing에 대한 논의는 플랫폼 제공업체인 CrowdFlower의 약 80명의 데이터 과학자를 대상으로 한 설문 조사에 대한 Gil Press의 "Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says: Mar 23, 2016 "를 통해서이다.
  - "데이터 py3_10_basic는 데이터 과학자 작업의 약 80%를 차지합니다. 데이터 과학자는 데이터를 정리하고 구성하는 데 시간의 60%를 사용합니다. 데이터 세트 수집은 시간의 19%로 두 번째입니다. 즉, 데이터 과학자는 분석을 위한 데이터 및 관리에 시간의 약 80%를 소비합니다."

출처 : https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#16ece2467f75

1. Data pre-processing

Data pre-processing

- Data pre-processing?



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

출처 : https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#16ece2467f75

1. Data pre-processing

Data pre-processing

- Gil Press는 또한 같은 글에서 “In 2009, data scientist Mike Driscoll popularized the term “data munging,” describing the “painful process of cleaning, parsing, and proofing one’s data” as one of the three Gendery skills of data geeks.” 라고 말하며, Data pre-processing에 대한 통찰을 보여주고 있는 데, 그것을 정리하면 “data munging”이라 하며 그 내용은 다음과 같다.

  - Data cleaning and organizing
  - Data cleaning, parsing, and proofing

출처 : https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#16ece2467f75

1. Data pre-processing

Data pre-processing

- 그렇다면, “data munging”은 무엇인지 계속 살펴보면,

“

Data wrangling, sometimes referred to as data munging, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics…
The goal of data wrangling is to assure quality and useful data…Data wrangling typically follows a set of general steps which begin with extracting the data in a raw form from the data source, "munging" the raw data (e.g. sorting) or parsing the data into predefined data structures, and finally depositing the resulting content into a data sink for storage and future use.

”

출처 : https://en.wikipedia.org/wiki/Data_wrangling

1. Data pre-processing

Data pre-processing

- 데이터 랭글링?
  - 데이터 랭글링 data wrangling은 원본 데이터를 정제하고 사용 가능한 형태로 구성하기 위한 변환 과정을 광범위하게 의미하는 비공식적인 용어. 데이터 랭글링은 데이터 전처리의 한 단계에 불과하지만 중요한 단계
  - Gather
    데이터를 얻는 방법으로는 다운로드(Download), 웹 스크레이핑(Web scrapping), API(Application Programming Interface)가 있다.
  - Assess
    얻은 데이터를 읽고 데이터가 깨끗한지 아닌지 판단하는 단계.
- Clean
  데이터를 정제하는 방법으로는 Define, Code, Test가 있습니다. 2단계(Assess)에서 발견된 데이터의 문제점을 보고 어떤 부분을 정제할지 정의하고(Define), 정제하기 위한 코드를 짜고(Code), 잘 정제가 되었는지 테스트를 해보는(Test) 것.
- Reassess and Iterate
  다시 2단계로 돌아가 데이터가 잘 정제되었는지 판단을 합니다. 추가로 정제해야 할 부분이 있다면 다시 2-3-4 단계를 반복.
- Store (optional)
  나중에 다시 사용하기 위해 저장하는 단계.

1. Data pre-processing

Data pre-processing

- 그렇다면, Data pre-processing 에 대한 정의를 좀 더 살펴보자.

"… manipulation or dropping of data… Examples of data preprocessing include cleaning, instance selection, normalization, one hot encoding, transformation, feature extraction and selection, etc. The product of data preprocessing is the final training set.

…

Tasks of data preprocessing

- Data cleansing
- Data editing
- Data reduction
- Data wrangling"

출처 : https://en.wikipedia.org/wiki/Data_pre-processing

1. Data pre-processing

Data pre-processing

- 다음 예를 보자,

예	설명

		Gender	Pregnant
Adult	1	Male	No
	2	Female	Yes
	3	~~Male~~	~~Yes~~
	4	Female	No
	5	~~Male~~	~~Yes~~

- 성별이 남성 또는 여성이고 임신 여부를 가진 5명의 불가능한 데이터 조합의 성인 데이터 셋

출처 : https://en.wikipedia.org/wiki/Data_pre-processing

1. Data pre-processing

Data pre-processing

- 다음 예를 보자,

Data cleansing			설명
		Gender	Pregnant
Adult	1	Male	No
	2	Female	Yes
	4	Female	No

~~Male~~	~~Yes~~

- 데이터세트에 존재하는 이러한 데이터가 사용자 입력 오류 또는 데이터 손상으로 인해 발생한 것으로 판단할 수 있기 때문에 이러한 데이터를 제거.
- 이러한 데이터를 삭제해야 하는 이유는 불가능한 데이터가 데이터 마이닝 프로세스의 후반 단계에서 계산 또는 데이터 조작 프로세스에 영향을 미치기 때문.

출처 : https://en.wikipedia.org/wiki/Data_pre-processing

1. Data pre-processing

Data pre-processing

- 다음 예를 보자,

Data editing	설명

		Gender	Pregnant
Adult	1	Male	No
	2	Female	Yes
	3	**Female**	**Yes**
	4	Female	No
	5	**Female**	**Yes**

~~**Male**~~	~~**Yes**~~

- 성인이 여성이라고 가정하고 그에 따라 변경할 수 있다.
- 데이터 마이닝 프로세스의 후반 단계에서 데이터 조작을 수행할 때 데이터를 보다 명확하게 분석할 수 있도록 데이터 세트를 편집

출처 : https://en.wikipedia.org/wiki/Data_pre-processing

1. Data pre-processing

Data pre-processing

- 다음 예를 보자,

Data reduction			설명
		Gender	Pregnant
Adult	2	Female	Yes
	4	Female	No
	1	Male	No
	3	Male	Yes
	5	Male	Yes

- 데이터를 성별로 정렬.
- 이를 통해 데이터 세트를 단순화하고 더 집중하고 싶은 성별을 선택할 수 있다.

출처 : https://en.wikipedia.org/wiki/Data_pre-processing

1. Data pre-processing

Data pre-processing

- Data pre-processing 에 대한 또 다른 예를 보자,

"

- 데이터 결합 (예: 행 결합, 열 결합, JOIN)
- 데이터 분할, 필터링, 샘플링
- 파생변수 생성 (예: 날짜 → 주말/평일 구분, 점수 → 등급)
- 더미변수 생성 (원-핫 인코딩, 예: 성별 → 0/1)
- 결측치 처리 (제거·보간)
- 이상치 처리 (제거·보간)
- 스케일 조정 (예: MixMax → 0~1, 표준점수, 로그스케일)
- 자료형 변경 (예: String → Datetime, String → Integer)
- 기타 데이터 수정·보정

"

출처 : https://zetawiki.com/wiki/데이터_전처리

1. Data pre-processing

Data pre-processing

- 마지막으로 NAVER 지식백과를 통해, Data pre-processing 에 대한 정의를 살펴보면

"

원자료(raw data)를 데이터 분석 목적과 방법에 맞는 형태로 처리하기 위하여 불필요한 정보를 분리 제거하고 가공하기 위한 예비적인 조작…데이터의 측정 오류를 줄이고 잡음(noise), 왜곡, 편차를 최소화한다. 정밀도, 정확도, 이상 값(outlier), 결측 값(missing value), 모순, 불일치, 중복 등의 문제를 해결하기 위한 방법으로 다음 전처리 방법들을 사용한다.

- 데이터 정제(cleansing): 결측 값(missing value; 빠진 데이터)들을 채워 넣고, 이상치를 식별 또는 제거하고, 잡음 섞인 데이터를 평활화(smoothing)하여 데이터의 불일치성을 교정하는 기술
- 데이터 변환(transformation): 데이터 유형 변환 등 데이터 분석이 쉬운 형태로 변환하는 기술. 정규화(normalization), 집합화(aggregation), 요약(summarization), 계층 생성 등의 방법을 활용한다.
- 데이터 필터링(filtering): 오류 발견, 보정, 삭제 및 중복성 확인 등의 과정을 통해 데이터의 품질을 향상하는 기술
- 데이터 통합(integration): 데이터 분석이 용이하도록 유사 데이터 및 연계가 필요한 데이터들을 통합하는 기술
- 데이터 축소(reduction): 분석 시간을 단축할 수 있도록 데이터 분석에 활용되지 않는 항목 등을 제거하는 기술

"

출처 : https://terms.naver.com/entry.naver?docId=850102&cid=50333&categoryId=50333

1. Data pre-processing

Data pre-processing

- **이상의 내용을 통해, Data pre-processing 작업을 요약해 보면 다음과 같다.**

① Data cleaning and organizing
② Data cleaning, parsing, and proofing
③ the raw data (e.g. sorting) or parsing the data into predefined data structures, and finally depositing the resulting content into a data sink for storage and future use
④ Data cleansing
⑤ Data editing
⑥ Data reduction
⑦ Data wrangling
⑧ 데이터 정제(cleansing)
⑨ 데이터 변환(transformation)
⑩ 데이터 필터링(filtering)
⑪ 데이터 통합(integration)
⑫ 데이터 축소(reduction)

① 데이터 결합 (예: 행 결합, 열 결합, JOIN)
② 데이터 분할, 필터링, 샘플링
③ 파생변수 생성 (예: 날짜 → 주말/평일 구분, 점수 → 등급)
④ 더미변수 생성 (원-핫 인코딩, 예: 성별 → 0/1)
⑤ 결측치 처리 (제거·보간)
⑥ 이상치 처리 (제거·보간)
⑦ 스케일 조정 (예: MixMax → 0~1, 표준점수, 로그스케일)
⑧ 자료형 변경 (예: String → Datetime, String → Integer)
⑨ 기타 데이터 수정·보정

1. Data Preprocessing

데이터 마트와의 관계

- 데이터 전처리(Data Preprocessing)와 데이터 웨어하우스의 관계:

  - 데이터 전처리는 데이터 분석을 위해 데이터를 정리하는 과정으로, 데이터 웨어하우스에 저장되는 데이터의 일관성과 품질을 유지하는 데 중요한 역할을 합니다.
  - 데이터 전처리는 데이터의 클린징, 변환, 통합, 축소 등의 과정을 포함하여 데이터를 분석 가능한 형태로 만듭니다.
  - 데이터 전처리는 데이터 웨어하우스에 저장된 데이터를 정제하고 가공하여, 비즈니스 분석 및 의사 결정에 활용할 수 있는 형태로 만듭니다.

- 따라서 데이터 웨어하우스에 저장된 데이터는 데이터 전처리를 거치고 정제되어 분석가능한 상태로 유지됩니다.

3 데이터 전처리

데이터 마트 및 데이터 전처리 단계 Data Mart & Data Preprocessing Pipeline

1. Data Structures
   - 데이터의 형태(정형, 반정형, 비정형), 변수의 타입(수치형, 범주형 등)
2. Data Gathering(Collect, Acquisition), Data Ingestion
   - 데이터 수집(Collect) / 데이터 획득(Acquisition) / 데이터 유입(Ingestion)
3. Data Invest & Exploratory Data Analysis, Data Visualization
   - 통계 요약, 시각화, 패턴 탐색
   - ※ "Data Invest" → 일반적으로는 EDA 또는 Data Understanding으로 표현
4. Data Cleansing (정제)
   - 결측치 처리, 이상치 제거, 오타 수정, 잡음 제거 등
5. Data Integration (통합)
   - 여러 소스/테이블/파일 병합, 공통 키 기반 조인 등

3 데이터 전처리

데이터 마트 및 데이터 전처리 단계 (Data Preprocessing Pipeline)

6. Data Reduction (축소)
   - 특성 선택, 차원 축소(PCA 등), 샘플링, 중복 제거
7. Data Transformation (변환)
   - 정규화, 표준화, 로그 변환, 이산화 등
8. Feature Engineering & Data Encoding
   - 새로운 변수 생성, 파생변수, 원-핫 인코딩, 라벨 인코딩 등
9. Cross Validation & Data Splitting
   - 학습 / 검증 / 테스트 세트로 나누기
10. Data Quality Assessment and Model Performance Evaluation
    - 전처리 결과 점검, 데이터 분포 확인, 모델 입력 적합성 평가 등
    - ※ "데이터 성능 측정" → 일반적으로는 전처리 품질 확인 또는 학습 후 모델 성능 평가 단계에서 수행

『1-3』
데이터 전처리

데이터

데이터 전처리

데이터 수집, 인제스트

분석 주제 탐색 및 문제해결 단계별 접근

데이터 확인 및 검증

결측값/데이터 분포/이상치

Syllabus

학습목표

- 이 워크샵에서는 Data Gathering(Collect, Acquisition), Data Ingestion에 대해 알 수 있다.

눈높이 체크

- Data Gathering(Collect, Acquisition), Data Ingestion을 알고 계신가요?

1.데이터 수집 (Data gathering)

Data gathering 정의

- "데이터 수집 (Data gathering)"은 분석, 모델링, 의사결정 등에 활용하기 위해 필요한 데이터를 체계적으로 모으는 과정을 의미합니다.
  - 데이터 획득 과정은 웹사이트 크롤링, 센서로부터의 데이터 수집, 데이터베이스 쿼리, 외부 API 호출, 파일 로드 등과 같이 다양한 방법으로 이루어질 수 있습니다. 이러한 데이터 획득 단계는 데이터 분석 및 의사 결정에 중요한 부분이며, 품질이 높고 신뢰할 수 있는 데이터를 확보하는 것이 매우 중요합니다.
  - 어떤 서비스를 할 것인지 결정했으면 먼저 수집할 원천 데이터 탐색 필요합니다.
  - 서비스 활용에 대해 수집 대상 데이터의 위치・주기・수집방법이 결정됐으면, 일반적으로 분산 파일 시스템에 저장하겠지만, 수집된 데이터를 가공・처리하기 위해서 DBMS가 사용될 수도 있고 서비스를 DBMS를 통해 제공할 수도 있으므로 서비스 환경에 맞는 아키텍처를 설계해야 합니다.

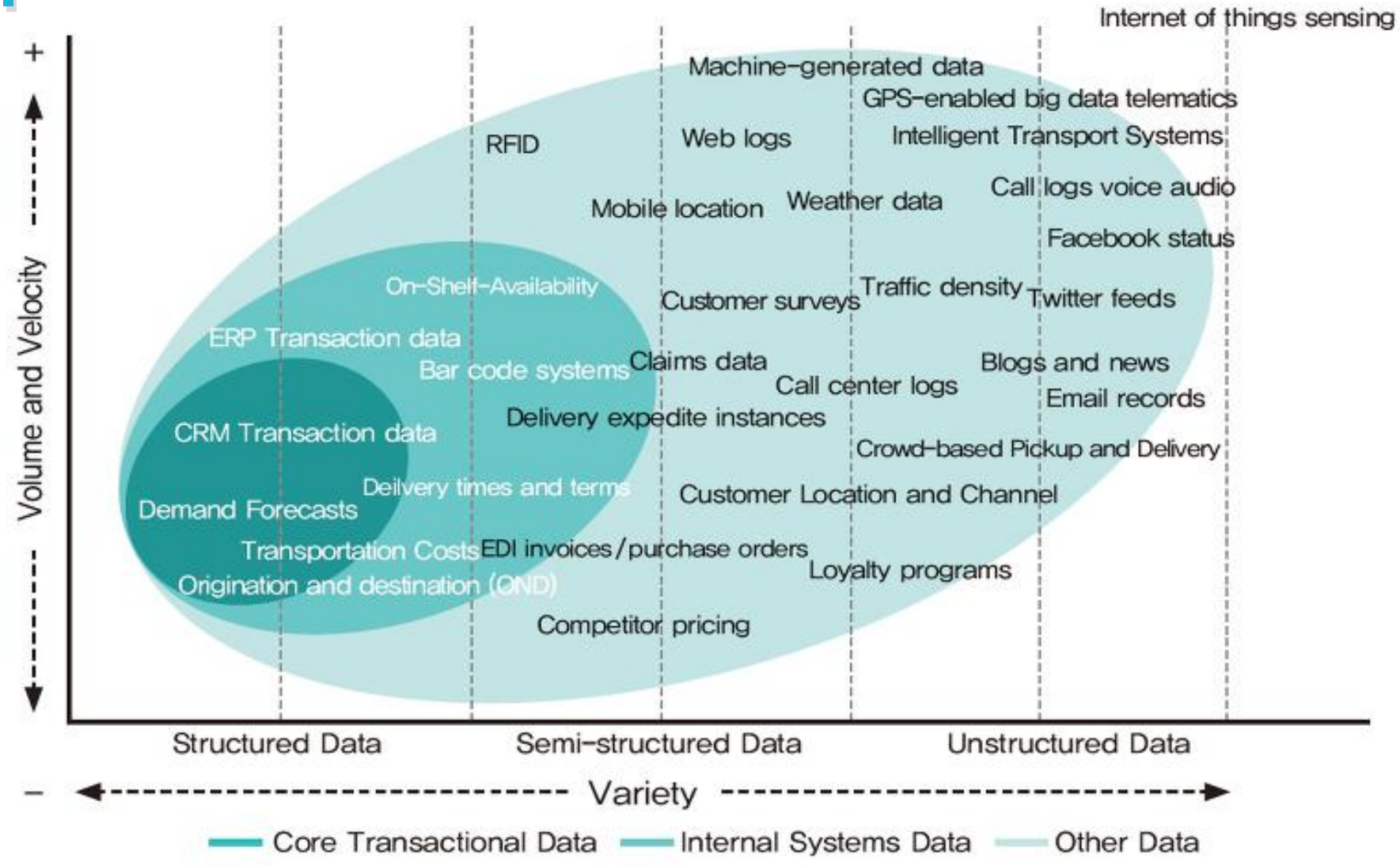
1.데이터 수집 (Data gathering)

Data 양 단위

바이트 크기 v • d • e • h					
SI 접두어		전통적 용법		이진 접두어	
기호(이름)	값	기호	값	기호(이름)	V값
kB (킬로바이트)	$1000^1 = 10^3$	KB	$1024^1 = 2^{10}$	KiB (키비바이트)	2^{10}
MB (메가바이트)	$1000^2 = 10^6$	MB	$1024^2 = 2^{20}$	MiB (메비바이트)	2^{20}
GB (기가바이트)	$1000^3 = 10^9$	GB	$1024^3 = 2^{30}$	GiB (기비바이트)	2^{30}
TB (테라바이트)	$1000^4 = 10^{12}$	TB	$1024^4 = 2^{40}$	TiB (테비바이트)	2^{40}
PB (페타바이트)	$1000^5 = 10^{15}$	PB	$1024^5 = 2^{50}$	PiB (페비바이트)	2^{50}
EB (엑사바이트)	$1000^6 = 10^{18}$	EB	$1024^6 = 2^{60}$	EiB (엑스비바이트)	2^{60}
ZB (제타바이트)	$1000^7 = 10^{21}$	ZB	$1024^7 = 2^{70}$	ZiB (제비바이트)	2^{70}
YB (요타바이트)	$1000^8 = 10^{24}$	YB	$1024^8 = 2^{80}$	YiB (요비바이트)	2^{80}

1.데이터 수집 (Data gathering)

다양한 데이터 유형



1.데이터 수집 (Data gathering)

Data 유형

- 정형데이터(RDBMS, CSV)
  - 형태가 있으며 연산 가능, 주로 관계형데이터베이스(RDBMS)에 저장
  - 데이터 수집 난이도가 낮고 형식이 정해져 있어 처리가 쉬움

- 반정형데이터(XML, HTML, JSON, 로그데이터)
  - 형태(스키마, 메타데이터)가 있으며 파일로 저장됨
  - 데이터 수집 난이도가 중간, 보통 API 형태로 제공됨

- 비정형데이터(소셜데이터, 영상, 이미지, 음성, 텍스트 등)
  - 형태가 없으며 연산 불가, 주로 NoSQL에 저장
  - 데이터 수집 난이도가 높고, 텍스트마이닝/파일일 경우 파일을 데이터 형태로 파싱해야 하므로 수집 데이터 처리가 힘들다.

1.데이터 수집 (Data gathering)

빅데이터 분석 활용을 위한 데이터 유형

- 빅데이터의 머신 러닝 또는 통계적 기법 적용을 위해 후보 데이터를 수치, 범주, 구조, 형 태 등에 따라 상세 유형으로 분류한다.

기준	분류	설명	사례
수치	이산	연속적이지 않은 수치형 데이터	나이
	연속	연속적인 수치형 데이터	온도, 키, 몸무게
범주	명사형	순서를 정할 수 없는 범주형 데이터	차량(택시, 버스)
	순서형	순서가 있는 범주형 데이터	1순위, 2순위, 3순위
구조	정형	형식이 정해져 있고 구조화된 데이터	DB
	비정형	형식이 정해지지 않은 데이터	사진, 동영상
	반정형	정형과 비정형의 중간 형태	신문 기사

출처: www.ncs.go.kr

1.데이터 수집 (Data gathering)

빅데이터 분석 활용을 위한 데이터 유형

- 빅데이터의 머신 러닝 또는 통계적 기법 적용을 위해 후보 데이터를 수치, 범주, 구조, 형 태 등에 따라 상세 유형으로 분류한다.

기준	분류	설명	사례
형태	문자형	문자로 구성된 데이터	대학교
	수치형	숫자로 구성된 데이터	12345, 19.27
	날짜/시간	날짜 또는 시간으로 구성된 데이터	2019-06-28
	불린	관계를 나타내는 데이터	참, 거짓
	이미지	바이너리(Binary)로 구성된 데이터	지도, 그림, 동영상
출처	내부	조직 내부에서 생성된 데이터	내부
	외부	조직 외부에서 생성/수집된 데이터	외부

출처: www.ncs.go.kr

1.데이터 수집 (Data gathering)

빅데이터 분석 활용을 위한 데이터 유형

- 정형 데이터의 개념
  - 정형 데이터는 관계형 데이터베이스 시스템의 테이블과 같이 고정된 필드에 저장되어 활용되는 구조화된(Structured) 데이터이다.

(가) 정형 데이터의 특징

– 정형 데이터는 데이터의 스키마를 가지고 있으며, RDB/스프레드시트 등에 저장되 어 활용된다. 정형 데이터는 데이터베이스를 설계자가 정의한 제한적인 구조로 정 보가 저장된다.

(나) 정형 데이터의 구조

– 정형 데이터는 컬럼(Column)과 로우(Row) 구조를 가지며, 설계된 구조 기반 목적에 맞는 정보들(예: 구매, 판매 및 사용자의 정보, 인기 품목 등)을 저장하고 분석하는 데 사용할 수 있다.

1.데이터 수집 (Data gathering)

빅데이터 분석 활용을 위한 데이터 유형

- **반정형 데이터의 개념**
  - 반정형 데이터는 고정된 필드에 저장된 정보는 아니지만, 데이터 내부에 정형 데이터 의 스키마에 해당하는 메타데이터(Metamata)를 갖고 있는 반구조적(semi-structed) 형 태를 가지는 데이터이다.

(가) 반정형 데이터의 특징과 사례

– 인터넷의 발달과 비즈니스 대 비즈니스 또는 프로세스 간 상호 정보 교환이 증가하여 일정 규약을 가지는 XML 또는 HTML 형태의 반정형 데이터가 방대하게 존재 하며, 이를 통한 정보 분석 요구 사항이 빅데이터 분석의 중요한 요건이다.

구분	내용
특징	• 데이터 내부에 데이터 구조에 해당하는 메타데이터를 가진다. • 관계형 데이터처럼 데이터 또는 스키마 간 엄격한 제약 관계를 갖지 만, 일반적으로 Well-Formed 디자인을 따른다.
유형	• Key : Value 구조를 기반으로 데이터를 구성하며, 파일 형태를 가진다. • HTML(Hypertext Markup Language), XML(eXtensible, Markup Language), JSON(Javascript Object Notation), NoSQL 데이터 등 • 웹로그, IoT에서 제공하는 센서 데이터

1.데이터 수집 (Data gathering)

빅데이터 분석 활용을 위한 데이터 유형

● 반정형 데이터의 개념

(나) 반정형 데이터의 구조

– 반정형 데이터는 데이터 내부의 메타 정보에 대해 어떤 형태로 구성되어 있는 데이터인지 파악 후 규칙에 따라 데이터를 추출할 수 있는 파싱 규칙을 적용한다.

구분	내용
Node형	XML, HTML과 같은 웹 데이터가 Node 형태의 구조를 가진다. (예) <Parent> <Child Node> Value </Child Node> <Child Node> Value </Child Node> </Parent>
Key-Value형	• 최근 웹 간 통신에 주로 사용되는 Json 데이터가 Key-Value 형태의 구조를 가진다. • NoSQL 데이터도 Key-Value 형태를 기반으로 Directory 구조를 가 진다. (예) Directory [{Key:Value, Key-Value, ..., {Key:Value}, {Key:Value, Key-Value, ..., {Key:Value}]

1.데이터 수집 (Data gathering)

빅데이터 분석 활용을 위한 데이터 유형

- **비정형 데이터의 개념**
  - **비정형 데이터는 고정된 필드가 아닌 구조화되지 않은(Unstructured) 데이터로, Data-Set 가 아닌 하나의 데이터가 수집 데이터로 객체화되어 있다.**

구분	내용
이진 파일	• 동영상, 문서, 음원, 이미지 파일 등
스크립트 파일	• SNS(Social Network Service) 또는 포털 사이트에 등록된 텍스트 데이터 • SNS, 포털 등 웹에 존재하는 데이터는 HTML 또는 XML 형태로 구성되어 있어 반정형 데이터로 분류할 수 있다.

1.데이터 수집 (Data gathering)

빅데이터 분석 활용을 위한 데이터 유형

- 서비스 모델과 관련한 데이터 후보 수집 가능 영역(외부 Site) 수집 사례

데이터 후보	서비스 모델 내 항목	추출·수집 가능 장 소 (제공 방법)	내부 데이터로 저장 필요성	비고
기상 수치 예보	• 지역별 일사량 • 지역별 기온, 습도 • 지역별 풍속	기상청 (API)	내·외부 데이터 분석 성능 향상	무료
태양광 자재	• 자재 타입 • 자재별 발전량 실적	H 자재 주식회사 (Web Service)	서비스 확대	유료

1.데이터 수집 (Data gathering)

식별된 후보 데이터의 유형과 크기 파악

- 서비스 모델과 관련한 후보 데이터의 유형을 파악한 사례

후보 데이터	유형	비고
기상 사진 정보	비정형, 이미지	
기상 수치 정보(기온, 풍속)	정형, 연속형	수치형
미세 먼지 정보	비정형, 이미지	
기상 특보 수치 정보	정형, 명사형	범주형
발전 용량	정형, 이산형	수치형
발전 타입	정형, 순서형	범주형
발전 위치	정형, 이산형	수치형

1.데이터 수집 (Data gathering)

식별된 후보 데이터의 유형과 크기 파악

● 식별된 후보 데이터의 최대 크기, 평균 크기를 파악한 사례

후보 데이터	유형	보관 형태	크기	비고
기상 사진 정보	비정형, 이미지	파일	최대 1.5MB 평균 0.7MB	파일 시스템
지역별 기상 수치 정보 (기온, 풍속)	정형, 연속형	DB	최대 1.5MB 평균 0.7MB	O사 제품
100m 고해상도 미세 먼지 정보	비정형, 이미지	NoSQL	최대 2.5MB 평균 1.7MB	M DB
기상 특보 수치 정보 (강수량, 일조 시간)	정형, 범주(명사형)	DB	최대 0.5MB 평균 0.3MB	O사 제품
발전 용량	정형, 이산형	DB	최대 0.1MB 평균 0.1MB	M사 제품
발전 타입	정형, 범주(명사형)	DB	최대 0.2MB 평균 0.1MB	O사 제품
발전 위치	정형, 이산형	DB	최대 0.2MB 평균 0.1MB	O사 제품

1.데이터 수집 (Data gathering)

식별된 후보 데이터의 생산 주체 및 생성 주기 파악

- **후보 데이터의 생산 조직 및 생산 시스템 파악 사례**

후보 데이터	유형	보관 형태	크기	생산 조직	생산 시스템
100m 고해상도 미세 먼지 정보	정형, 이산형	NoSQL	최대 2.5MB 평균 1.7MB	기상부	미세 먼지 관리 시스템
발전 용량	정형, 이산형	DB	최대 0.1MB 평균 0.1MB	발전부	발전 관리 시스템
발전 타입	정형, 범주 (명사형)	DB	최대 0.2MB 평균 0.1MB	발전부	발전 관리 시스템
발전 위치	정형, 범주 (명사형)	DB	최대 0.2MB 평균 0.1MB	발전부	발전 위치 관리 시스템

1.데이터 수집 (Data gathering)

식별된 후보 데이터의 생산 주체 및 생성 주기 파악

- 후보 데이터의 생산 조직 및 생산 시스템 파악 사례

후보 데이터	유형	보관 형태	크기	생산 조직	생산 시스템	생산 주기	데이터 건수
미세 먼지 고해상도 정보	비정형, 이미지	NoSQL	최대 2.5MB 평균 1.7MB	기상부	미세 먼지 관리 시스템	1분	30만
발전 용량	정형, 이산형	DB	최대 0.1MB 평균 0.1MB	발전부	발전 관리 시스템	1일	12만
발전 타입	정형, 범주 (명사형)	DB	최대 0.2MB 평균 0.1MB	발전부	발전 관리 시스템	1분	30만
발전 위치	정형, 범주 (명사형)	DB	최대 0.2MB 평균 0.1MB	발전부	발전 위치 관리 시스템	1일	1만

1.데이터 수집 (Data gathering)

Data 형태

- 질적 자료(정성적 자료, Qualitative or Categorical): 범주 또는 순서 형태의 속성을 가지는 자료
  - 범주형(명목형, nominal) 자료: 사람의 피부색, 성별
  - 순서형(서수형, ordinal) 자료: 제품의 품질, 등급, 순위
- 양적 자료(정량적 자료, Quantitative or Numeric): 관측된 값이 수치 형태의 속성을 가지는 자료
  - 범위형interval 자료: 화씨, 섭씨와 같이 수치간에 차이가 의미를 가지는 자료.
  - 비율ratio 자료: 무게와 같이 수치의 차이 뿐만 아니라 비율 또한 의미를 가지는 자료

2. Data gathering 절차

Data Collection or Data acquisition 절차

- 데이터 수집 목표 확인
- 데이터 선정
- 선정된 데이터 위치 파악
- 데이터 유형 파악
- 데이터 수집 시 필요한 기술 및 보안사항
- 데이터 수집 계획서 작성

2. Data gathering 절차

[1단계] 연구 목표 설정

- 프로젝트와 관련된 모든 참여자가 연구 목표를 함께 정의하고 산출물과 일정 등의 계획에 합의한 뒤 프로젝트 헌장을 작성합니다.

● 분석 목표 정의서 예시 (특허정보 경우)

분석 기본 정의	분석 명칭	특허 전략에 대한 텍스트 마이닝과 특허와 기업 정보 그룹의 연관성 분석	분석목표 확정일	2022-10-13
	분석 목적	특허 전략에 대한 텍스트 마이닝 후 특허와 기업 정보 매칭	분석 목표 워크숍	2022-10-13
	분석 우선순위	상	담당 조직명	-
	분석 접근 방안	각 팀원이 특허 전략에 대한 텍스트 마이닝과 특허와 기업 정보 그룹의 연관성 분석을 한 후 개인별 결과를 취합하여 발표자료 작성		
성과 측정	정성적 기준	신규 기법/기술 : 다양한 EDA 및 예측 기법 및 머신러닝/ 딥러닝활용으로Best Model 선정 기존 데이터: 한국 특허 데이터(2021).csv 신규 데이터 : ...		
	정량적 기준	Doc2Vec으로 특허 유사도 계산해 추천하기를 통해 특허코드 관련 기업 추천 목표 달성		
데이터 정보	내부 데이터	한국 특허 데이터(2021).csv	데이터 입수 난이도	중
	외부 데이터		데이터 입수 난이도	

2. Data gathering 절차

[1단계] 연구 목표 설정

- 프로젝트 현장(Project Charter) 예시 (특허정보 경우)

프로젝트 현장(Project Charter)	
프로젝트 명 (Project Name)	특허 전략에 대한 텍스트 마이닝과 특허와 기업 정보 그룹의 연관성 분석
프로젝트 설명 (Project Description)	특허 전략에 대한 텍스트 마이닝 후 특허와 기업 정보 매칭

프로젝트 매니저 (Project Manager, PM)	홍길동	승인 날짜 (Date Approved)	2022-10-20
프로젝트 스폰서 (Project Sponsor)	Ki4C	서명 (Signature)	홍길동

비즈니스 케이스(Business Case)		목표(Goals) / 산출물(Deliverables)
불량률에 대한 추세분석을 통해 장비 점검·교체		• 기존의 앨러스틱 분석과통계적 정보에 대한 실시간 조회 대국민 서비스의 한계를 극복하고 보다 진일보한 서비스를 모색하기 위해 텍스트 마이닝을 통한 특허 전략 분석 후 키워드 검색 서비스와 특허와 기업 정보 매칭 등의 연관성 분석을 하고자 한다. 본 사업을 통해 특허 전략에 대한 텍스트 마이닝 후 특허와 기업 정보 매칭을 할 수 있는 것을 기대한다.
팀 구성원(Team Member)		
이름(Name)	역할(Role)	
홍길동	PM	
박문수	엔지니어	

2. Data gathering 절차

[1단계] 연구 목표 설정

- 프로젝트 현장(Project Charter) 예시 (특허정보 경우)

위험과 제약사항(Risk and Constraints)		주요 일정(Milestones)	
			1.문제 정의(Problem Definition)와 question, 알고리즘 선정
			2.헌장(Charter)
			3.작업환경 만들기
			4.데이터 획득(Data Acquisition)
			5.EDA : Exploratory Data Analysis : 탐색적 데이터 분석(Data Exploration and Analysis)과 데이터 시각화(Data Visualization)를 통한 insight 얻기
			6.Machine Learning Algorithms을 위한 데이터 준비
			7.통계적 모델링 혹은 모형화(Statistical Modeling) 혹은 확증적 데이터 분석(Confirmatory Data Analysis, CDA)
			8.효과 검증
			9.Data 도전과 배움
			10.서비스 구현, 활용 계획

2. Data gathering 절차

[1단계] 연구 목표 설정

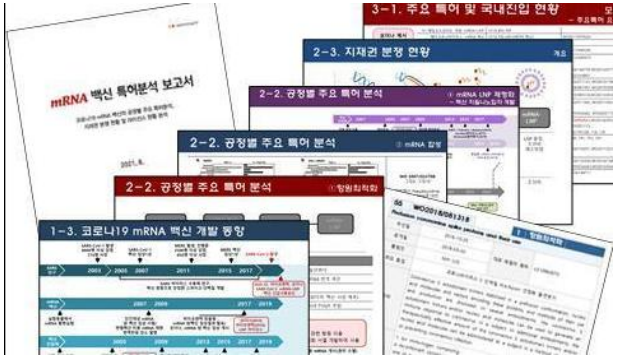
- 분석목표 정의 예시 (특허정보 경우)

특허 전략에 대한 텍스트 마이닝과 특허와 기업 정보 그룹의 연관성 분석	
목표	특허 전략에 대한 텍스트 마이닝 후 특허와 기업 정보 매칭
핵심개념	1) 특허 전략에 대한 텍스트 마이닝과 특허와 기업 정보 그룹의 연관성 분석을 위한 텍스트 데이터 변환하기 2) 텍스트 데이터 분석 수행 방법 계획하기 3) 텍스트 데이터 분류 결과 분석하기 4) 정형 데이터 결합 분석 수행하기 5) 자율학습 모델 적용하기 연관성 및 패턴화 관련 머신러닝 기법은 다음 알고리즘을 포함한다. - Apriori, FP-Growth, Eclat, Collaborative Filtering 등 6) 분석 및 모델 적용 결과 및 자료 취합
데이터 수집	○○○ 데이터 셋 : 한국 특허 데이터(2021).csv
데이터 준비	수집한 데이터 파일 병합
데이터 탐색	1. 정보 확인 : info() 2. 기술 통계 확인 : describe(), unique(), value_counts()

2. Data gathering 절차

[1단계] 연구 목표 설정

- 분석 시나리오 예시 (특허정보 경우)

결과 시각화	
특허정보 분석	연관성 및 패턴화 관련 머신러닝
	Doc2Vec으로 특허 유사도 계산해 추천하기 ▾ 형태소 분석기 Mecab 설치 [ ] %env JAVA_HOME "/usr/lib/jvm/java-8-openjdk-amd64" env: JAVA_HOME="/usr/lib/jvm/java-8-openjdk-amd64" [ ] %%bash bash <(curl -s https://raw.githubusercontent.com/konlpy/konlpy/master/scripts/mecab.sh) pip3 install /tmp/mecab-python-0.996

2. Data gathering 절차

2단계 > 데이터 수집

- 프로젝트에 필요한 데이터의 위치와 형태를 확인하고 원시 데이터를 수집.
  - 필요한 데이터를 수집할 때는 이미 가지고 있는 내부 데이터베이스나 데이터 저장소를 이용
  - 외부에서 수집하는 경우 다양한 수집 기술을 활용할 수 있음
  - 수집할 데이터의 유형과 종류를 파악한 뒤 그에 맞는 수집 기술을 선택해서 사용
  - 데이터 유형과 종류에 따라 사용할 수 있는 수집 기술 예

유형	종류	수집 기술
정형 데이터	RDB, 스프레드시트	ETL, FTP, Open API
반정형 데이터	HTML, XML, JSON, 웹 문서, 웹 로그, 센서 데이터	크롤링, RSS, Open API, FTP
비정형 데이터	소셜 데이터, 문서(워드, 한글), 이미지, 오디오, 비디오, IoT	크롤링, RSS, Open API, 스트리밍, FTP

2. Data gathering 절차

2단계 > 데이터 수집

- 데이터 수집이 데이터 처리 분석 및 모델 생성의 첫 과정
  - 목적과 목표가 되는 정보를 수집하고 측 정하기 위해 정의가 필요
  - 문제의 정의와 문제해결을 위한 데이터 분석 기획과 시나리오가 중요
  - 문제를 식별하고 탐색함으로써 정보 수집 시기 및 방법을 결정
  - 데이터 종류에 따라서 내부 또는 외부, 질적 또는 양적 데이터 수집
- 서비스 활용에 필요한 데이터를 시스템의 내부 혹은 외부에서 주기성을 갖고 필요한 형 태로 수집하는 활동
- 데이터 수집 정의 요소
  - 서비스 활용
  - 데이터 위치
  - 주기성
  - 수집 데이터의 저장 형태

2. Data gathering 절차

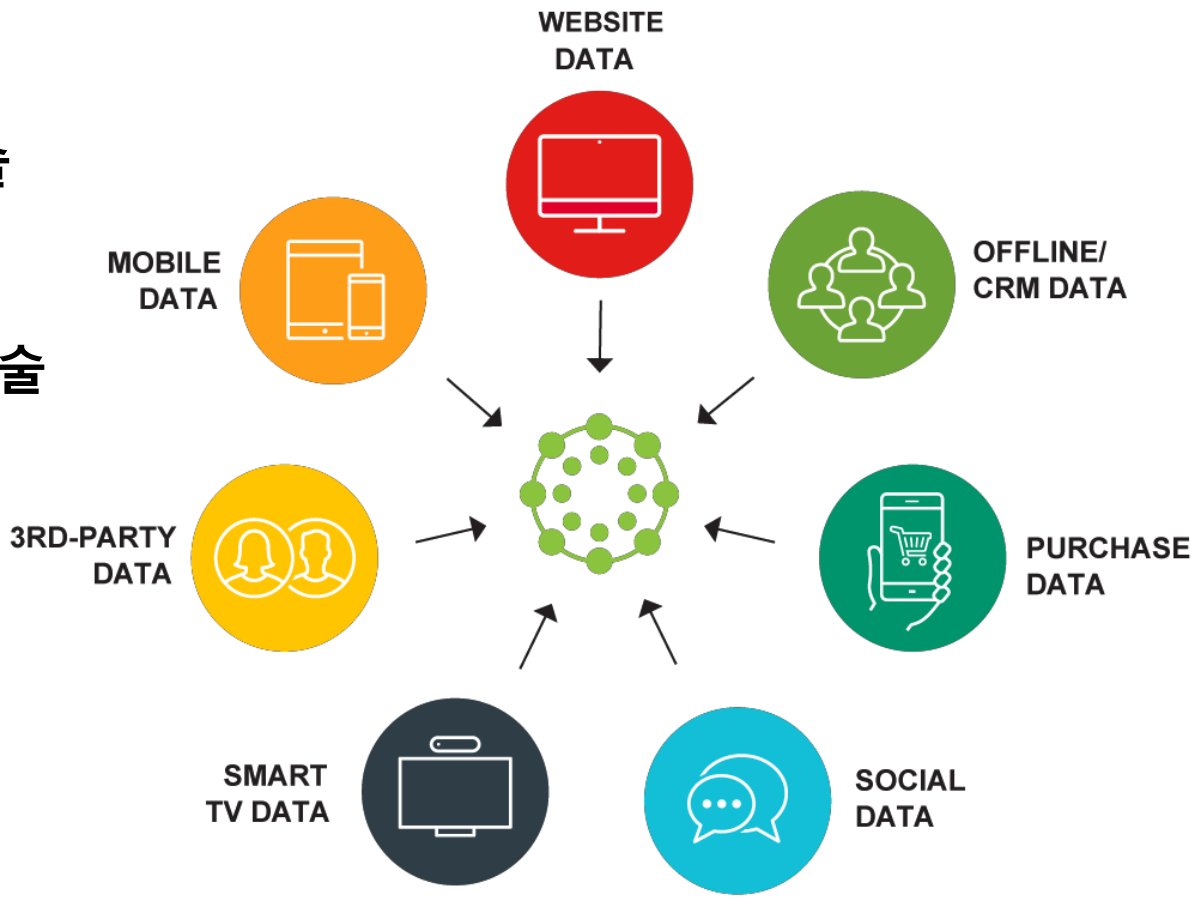
2단계 > 데이터 수집

- 수집 데이터의 저장·관리되는 형태에 따른 분류
  - 정형 데이터
  - 반정형 데이터
  - 비정형 데이터
- 수집 데이터의 저장 위치에 따른 분류
  - 내부 데이터
  - 외부 데이터
- 수집 데이터의 생산 주체에 따른 분류

2. Data gathering 절차

일반적 수집 데이터의 형태와 종류

- HTTP 수집
  - 크롤링 기술
  - Open API 수집 기술
- 로그/센서 수집
  - 로그 수집 기술
  - 센서 데이터 수집 기술
- DBMS 수집
- FTP 수집



출처 : http://xcademy.in/data-management-platform/data-collection/ 이하 동일

2. Data gathering 절차

일반적 수집 데이터의 형태와 종류

- **크롤링 기술**

기능	기능항목 요건
환경 설정기능	▪ 수집할 사이트의 URL 목록을 관리 하는 기능 ▪ 수집주기를 설정하는 기능 ▪ URL, 설정값을 에이전트에 전달하는 기능 ▪ 설정의 수동/자동 조작을 통해 관리할 수 있는 기능
데이터 처리 에이전트 기능	▪ 에이전트별 관리기능 ▪ 에이전트 운영 기능 ▪ 웹문서, 콘텐츠 수집기능 ▪ 수집한 웹문서에서 URL 추출기능 ▪ 머신런닝 기능 ▪ URL 리스트 탐색 기능 ▪ 스크랩한 문서를 DB에 업로드하는 기능 ▪ 병렬 크롤링 기능 ▪ 불가역 데이터를 수집하므로 불필요 데이터를 전처리하는 기능
접근제어기능	▪ 에이전트가 웹 사이트에 접근하는 것을 방지하기 위한 로봇 배제 표 준 규약 권고안을 준수하도록 robots.txt를 탐지하는 기능

2. Data gathering 절차

일반적 수집 데이터의 형태와 종류

- Open API 수집기술

기능	기능항목 요건
환경 설정기능	▪ 수집할 사이트의 URL 목록을 관리 하는 기능 ▪ 수집주기를 설정하는 기능 ▪ URL, 설정값을 에이전트에 전달하는 기능 ▪ 설정의 수동/자동 조작을 통해 관리할 수 있는 기능
데이터 처리 에이전트 기능	▪ 에이전트별 관리기능 ▪ 에이전트 운영 기능 ▪ 에이전트가 mash-up이 가능하도록 RESTFUL 방식의 API 제공기능 ▪ 콘텐츠 자원에 유일한 URI를 부여하는 기능 ▪ CRUD를 에이전트의 메소드로 제공하는 기능 ▪ 반정형 데이터(XML, JSON, RSS)의 정보 제공방식을 지원 ▪ 수집 데이터의 성공/실패에 대한 응답기능
테이블 매핑기능	▪ 수집한 반정형 데이터의 요소와 수집 시스템의 컬럼에 대해 매핑하는 기능

2. Data gathering 절차

일반적 수집 데이터의 형태와 종류

- **로그 수집기술**

기능	기능항목 요건
환경 설정기능	▪ 수집할 사이트의 URL 목록을 관리 하는 기능 ▪ 수집주기를 설정하는 기능 ▪ URL, 설정값을 에이전트에 전달하는 기능 ▪ 설정의 수동/자동 조작을 통해 관리할 수 있는 기능
데이터 처리 에이전트 기능	▪ 에이전트별 관리기능 ▪ 에이전트 운영 기능 ▪ 수집한 파일을 Chunk 단위로 전송하는 기능 ▪ 주기/블록 단위로 잘라서 전송하는 기능 ▪ 압축지원 기능 ▪ 수집대상 파일 체크기능
콜렉터기능	▪ 다수의 에이전트로 받은 데이터를 직렬화된 chunk에 분산 파일시스템의 시퀀스 파일로 전송하는 기능 ▪ 트래픽 밸런싱을 자동 조정하거나 수동으로 관리할 수 있는 기능 ▪ 파일 전송 모니터링 기능

2. Data gathering 절차

일반적 수집 데이터의 형태와 종류

- DBMS 수집기술

기능	기능항목 요건
환경 설정기능	▪ 수집할 사이트의 URL 목록을 관리 하는 기능 ▪ 수집주기를 설정하는 기능 ▪ URL, 설정값을 에이전트에 전달하는 기능 ▪ 설정의 수동/자동 조작을 통해 관리할 수 있는 기능
데이터 처리 에이전트 기능	▪ 에이전트별 관리기능 ▪ 에이전트 운영 기능 ▪ DBMS 메타 정보에서 테이블을 선택하는 기능 ▪ DBMS 메타 정보에서 컬럼을 선택하는 기능 ▪ 레코드 단위로 수집해 분산파일시스템으로 전송하는 기능 ▪ 수집의 성공/실패시 응답 기능
클린징기능	▪ 수집한 데이터를 타겟시스템에 맞게 정제하여 로드하는 기능

2. Data gathering 절차

일반적 수집 데이터의 형태와 종류

- **FTP 수집기술**

기능	기능항목 요건
환경 설정기능	▪ 수집할 사이트의 URL 목록을 관리 하는 기능 ▪ 수집주기를 설정하는 기능 ▪ URL, 설정값을 에이전트에 전달하는 기능 ▪ ACTIVE, PASSIVE 연결에 필요한 통신포트 설정기능
데이터 처리 에이전트 기능	▪ 에이전트별 관리기능 ▪ 에이전트 운영 기능 ▪ 클라이언트 방화벽 운영시 액티브, 패시브 모드 변환기능 ▪ 액티브 연결기능 ▪ 패시브 연결기능 ▪ 파일전송 연결기능
파일전송기능	▪ ASCII, BINARY 파일 전송기능 ▪ Get, put, mget, mput 명령기능 ▪ 수집파일에 대한 무결성 확인기능

3. Data gathering 출처 사례

공개 데이터 셋

- 데이터 과학을 연구하는 데 관심이 있는 사람들을 위해 최근 쉽게 액세스 할 수 있는 행정 기관 및 지역 기관, 공공포털 등이 있다.
- 공개 데이터 셋 구할 수 있는 곳

◦ 유명한 공개 데이터 저장소

① UC 얼바인Irvine 머신러닝 저장소(http://archive.ics.uci.edu/ml)
② 캐글Kaggle 데이터셋(http://www.kaggle.com/datasets)
③ 아마존 AWS 데이터셋(https://registry.opendata.aws)

◦ 메타 포털(공개 데이터 저장소 나열)

① 데이터 포털Data Portals(http://dataportals.org)
② 오픈 데이터 모니터Open Data Monitor(http://opendatamonitor.eu)
③ 퀀들Quandl(http://quandl.com)

◦ 인기 있는 공개 데이터 저장소가 나열되어 있는 다른 페이지

① 위키백과 머신러닝 데이터셋 목록(https://goo.gl/SJHN2k)
② Quora.com(https://homl.info/10)
③ 데이터셋 서브레딧subreddit(http://www.reddit.com/r/datasets)

3. Data gathering 출처 사례

Scikit-learn Datasets

- Scikit-learn 내의 데이터 세트 패키지에는 인기 있는 머신의 다운 샘플링된 버전이 포함되어 있다.Iris, Boston 및 Digits 데이터 세트와 같은 학습 데이터 세트, 이러한 데이터 세트는 장난감 데이터 세트로 자주 참조된다.
- Scikit-learn 내의 데이터 세트 패키지는 다음 속성으로 구성되어 있다.
  - DESCR: 사람이 읽을 수 있는 데이터 세트 설명을 반환합니다.
  - data: 모든 기능에 대한 데이터가 포함된 NumPy 배열을 반환합니다.
  - feature_names: 기능의 이름을 포함하는 NumPy 배열을 반환합니다. 모든 장난감 데이터 세트가 이 속성을 지원하는 것은 아닙니다.
  - target: 대상 변수에 대한 데이터를 포함하는 NumPy 배열을 반환합니다.
  - target_names: 범주형 대상 변수의 값을 포함하는 NumPy 배열을 반환합니다. 숫자, 보스턴 주택 가격 및 당뇨병 데이터 세트는 이 속성을 지원하지 않습니다.

3. Data gathering 출처 사례

Scikit-learn Datasets

- Scikit-learn에 포함된 장난감 데이터 세트(Toy Datasets)목록은 다음과 같다.

① Boston 집값 데이터셋: 회귀 모델 구축에 사용되는 인기 데이터셋입니다. 이 데이터세트의 장난감 버전은 load_boston() 함수를 사용하여 로드할 수 있다. 이 데이터 세트의 전체 버전은 https://archive.ics.uci.edu/ml/ machine-learning-databases/housing/에서 찾을 수 있다.

② iris 식물 데이터세트: 분류 모델 구축에 사용되는 인기 있는 데이터세트입니다. 이 데이터세트의 장난감 버전은 load_iris() 함수를 사용하여 로드할 수 있다. 이 데이터세트의 전체 버전은 https://archive.ics.uci.edu/ml/datasets/iris에서 찾을 수 있다.

③ 당뇨병 데이터 세트의 시작: 회귀 모델 구축에 사용되는 인기 있는 데이터 세트입니다. 이 데이터세트의 장난감 버전은 load_diabetes() 함수를 사용하여 로드할 수 있다. 이 데이터 세트의 전체 버전은 http://www4.stat.ncsu.edu/~boos/var.select/diabetes.html에서 찾을 수 있다.

④ 필기 숫자 데이터셋: 필기 숫자 0 ~ 9의 이미지 데이터셋으로 분류 작업에 사용됩니다. 이 데이터세트의 장난감 버전은 load_digits() 함수를 사용하여 로드할 수 있다. 이 데이터 세트의 전체 버전은 http://archive .ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits에서 찾을 수 있다.

3. Data gathering 출처 사례

Scikit-learn Datasets

- Scikit-learn에 포함된 장난감 데이터 세트(Toy Datasets)목록은 다음과 같다.

⑤ Linnerud 데이터셋: 중년 남성을 대상으로 측정한 운동변수 데이터셋으로 다변량 회귀분석에 사용된다. 이 데이터세트의 장난감 버전은 load_linnerud() 함수를 사용하여 로드할 수 있다. 이 데이터 세트의 전체 버전은 https://rdrr .io/cran/mixOmics/man/linnerud.html에서 찾을 수 있다.

⑥ 와인 인식 데이터셋: 이 데이터셋은 이탈리아에서 생산된 와인에 대한 화학적 분석 결과입니다. 분류 작업에 사용됩니다. 이 데이터세트의 장난감 버전은 load_wine() 함수를 사용하여 로드할 수 있다. 이 데이터 세트의 전체 버전은 https://archive.ics.uci.edu/ml/machine-learning-databases/wine/에서 찾을 수 있다.

⑦ 유방암 데이터세트: 이 데이터세트는 유방암 종양의 세포핵 특성을 설명합니다. 분류 작업에 사용됩니다. 이 데이터 세트의 장난감 버전은 load_breast_cancer() 함수를 사용하여 로드할 수 있다. 이 데이터 세트의 전체 버전은 https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(진단)에서 찾을 수 있다.

3. Data gathering 출처 사례

UCI 머신 러닝 리포지토리

- UCI 머신 러닝 저장소는 UC Irvine의 머신 러닝 및 지능형 시스템 센터에서 유지 관리하는 450개 이상의 데이터 세트로 구성된 공개 모음. 머신 러닝 데이터 세트의 가장 오래된 소스 중 하나이며 초보자와 숙련된 전문가 모두가 자주 찾는 곳. 데이터 세트는 일반 대중이 제공하며 모델 구축에 사용하기 위해 수행해야 하는 사전 처리 수준이 다르다. 데이터 세트를 로컬 컴퓨터에 다운로드 한 다음 Pandas 및 Scikit-learn과 같은 도구를 사용하여 처리할 수 있. https://archive.ics .uci.edu/ml/datasets.php에서 전체 데이터 세트 목록을 탐색할 수 있다.
- 가장 인기 있는 UCI 머신 러닝 저장소 데이터 세트 중 일부는 Kaggle.com에서도 호스팅되며 https://www.kaggle.com/uciml에서 액세스할 수 있다.

3. Data gathering 출처 사례

Kaggle.com 데이터 세트

- Kaggle.com은 머신 러닝 대회를 주최하는 인기 있는 웹사이트. Kaggle.com또한 https://www .kaggle.com/datasets에서 액세스할 수 있는 일반 사용을 위한 많은 데이터 세트가 포함되어 있다. 페이지에 나열된 일반 사용 데이터 세트 외에도 Kaggle.com의 대회에는 대회에 참가하여 액세스할 수 있는 자체 데이터 세트가 있다. 데이터 세트 파일을 로컬 컴퓨터에 다운로드 한 다음 Pandas 데이터 프레임에 로드할 수 있다.
- 현재 및 과거 대회 목록은 https://www.kaggle.com/competitions 에서 확인할 수 있다.

3. Data gathering 출처 사례

AWS 공개 데이터 세트

- Amazon은 AWS에 배포되는 애플리케이션에 쉽게 통합할 수 있는 공개 머신 러닝 데이터 세트의 리포지토리를 호스팅한다. 데이터 세트는 S3 버킷 또는 EBS 볼륨으로 사용할 수 있다. S3 버킷에서 사용 가능한 데이터 세트는 AWS CLI, AWS SDK 또는 S3 HTTP 쿼리 API를 사용하여 액세스할 수 있다. EBS 볼륨에서 사용 가능한 데이터 세트는 EC2 인스턴스에 연결해야 한다.
- 공개 데이터세트는 다음 범주에서 사용할 수 있다.
  - 생물학: 인간 게놈 프로젝트와 같은 인기 있는 데이터 세트를 포함한다.
  - Chemistry: PubChem 및 기타 콘텐츠의 여러 버전을 포함한다. 펍켐은https://pubchem.ncbi .nlm.nih.gov에서 액세스할 수 있는 화학 분자 데이터베이스.
  - 경제: 인구 조사 데이터 및 기타 콘텐츠를 포함합니다.
  - 백과사전: Wikipedia 콘텐츠 및 기타 콘텐츠를 포함합니다.https://registry.opendata.aws에서 AWS 공개 데이터 세트 목록을 찾아볼 수 있다.

3. Data gathering 출처 사례

바이오정보학과 데이터베이스

- Sequence homology search
  - http://blast.ncbi.nlm.nih.gov/Blast.cgi

서열 유사도 분석

  - http://www.ebi.ac.uk/Tools/msa/clustalo/

Multiple sequence alignment tool

  - http://www.ebi.ac.uk/interpro/

Amino acid sequence를 이용 domain/motif 검색

- Sequence 분석
  - http://web.expasy.org/compute_pi/

sequence의 pI와 mw를 계산해 주는 툴

  - http://web.expasy.org/peptide_mass/

tryptic peptide의 mw를 계산

  - http://web.expasy.org/peptide_cutter/

Peptide cleavage site 예측

3. Data gathering 출처 사례

바이오정보학과 데이터베이스

- PTM 예측
  - http://www.cbs.dtu.dk/services/NetAcet/

N-acetyltransferase A 의 substrate site 예측

  - http://www.cbs.dtu.dk/services/NetCGlyc/

C-mannosylation site 예측

  - http://www.cbs.dtu.dk/services/NetNGlyc/

N-Glycosylation site 예측

  - http://www.cbs.dtu.dk/services/NetPhos/

Phosphorylation site 예측

- Protein localization/topology 예측
  - http://www.psort.org/

Subcellular localization 예측

  - http://www.cbs.dtu.dk/services/SignalP/

Signal peptide cleavage site 예측

  - http://www.cbs.dtu.dk/services/TMHMM-2.0/

Trans-membrane domain 예측

3. Data gathering 출처 사례

바이오정보학과 데이터베이스

- Functional annotation tool
  - https://david.ncifcrf.gov/

Functional annotation tool

- Useful sites
  - http://www.expasy.org/proteomics

Proteomics 관련 tool list

  - http://www.oxfordjournals.org/our_journals/n ar/webserver/c/

Nucleic Acid Research의 web-server issue에 나온 tool list

3. Data gathering 출처 사례

화합물DB

- 화합물 DB는 상업화되어 있는 것까지 포함하여 대 략 30여종에 이르고 있다.
  - MDL사의ACD 3D DB의 경우는 현재 판매하는 시약들의 종류 및 판매회사들을 DB화하여 필요한 시약에 대한 정보를 얻을 수 있도록 되어있고. Cambridge 사의 ACX의 경우 비슷한 용도이지만 개인용 p c를 중심으로 사용 가능하도록 바꾼 형태. MDDR은 약효검사를 중심으로, CMC는 실험값들을 중심으로 DB 화 되어 있고, SciFinder는 자연어 검색과 대량의 참고 문헌 정보 및 합성정보 그리고 이들의 fulltext서비스 등을 수행하고 있고, Derwent는 참고문헌과 합성정보를 지니고 있지만 오히려 특허정보를 차별화시킨 DB 라고 할 수 있다. 이러한 DB들은 국내 연구소와 제약 업계에게 널리 사용되고 있으며 대개의 경우 1user가 사용할 때 년간 1000만원 이상의 고가의 사용료를 부담하고 있는 실정.

3. Data gathering 출처 사례

화합물DB

- Drugs.com은 Drugsite Trust 가 소유하고 운영하고 있으며 일반인과 의료전문가에게 무료 약물 정보를 제공하는 데이터베이스입니다.
- Wolters Kluwer, 미국병원약사회(ASHP), Cerner Mulum, IBM Watson Micromedex, Havard University, Mayo Clinic 등으로부터 제공받은 신뢰할 수 있는 정보로 일반인을 위한 정보, 전문가를 위한 정보로 분류되어 있으며 포털사이트처럼 구성

https://www.drugs.com/

- 한국화합물은행

https://chembank.org/

- 화학물질정보시스템: NCIS

https://ncis.nier.go.kr/main.do

- ChemSpider

http://www.chemspider.com/

4. Data acquisition 예

데이터 수집 세부 계획서

● 특허정보 경우 예시

특허 전략에 대한 텍스트 마이닝과 특허와 기업 정보 그룹의 연관성 분석 데이터 수집 세부 계획서							
특허 전략에 대한 텍스트 마이닝과 특허와 기업 정보 그룹의 연관성 분석을 위한 데이터 수집 세부 계획서					담당자		
문서번호		작성자			작성일자		
1. 분석 목적							
2. 수집 데이터 상세 조사 내용							
데이터 유형	통계 | 문자 | 텍스트 | 음성 | 이미지 | 동영상 | GIS | 기타				수집 주기		
위치	수요기업 | 공급기업 보유 또는 수집 | 허브 데이터셋 | 공공데이터				확보 비용		
크기	레코드수	300	레코드단위	장	데이터 이관 절차		
	크기	15	단위	MB			
보관 방식							
3.적절성 검증 방식							
데이터 누락/중복							
데이터 오류							
개인정보 유무	포함 | 미포함						
데이터 저작권							

4. Data acquisition 예

특허문서의 데이터 셋 구할 수 있는 곳

- 특허빅데이터센터 KPBCenter
  - https://biz.kista.re.kr/pbcenter/p/index
- 특허청
  - https://www.kipo.go.kr/ko/MainApp.do
- 특허정보검색서비스 키프리스
  - http://www.kipris.or.kr/khome/main.jsp
- 특허로
  - https://www.patent.go.kr/smart/portal/Main.do
  - 국내특허 문헌번호체계
  - 국내 특허관련 문헌번호는 총 13자리로 이루어져 있으며, 출원, 공개, 공고 및 등록 번호가 이에 해당.
  - 특허 번호체계는 권리의 구분을 의미하는 숫자(2자리)와 연도(4자리), 일련번호(7자리)로 구성되어 있으며,등록번호는 권리(2자리)와 일련번호(7자리), 자릿값(4자리=0000)으로 구성되어 있다.

4. Data acquisition 예

특허문서의 데이터 셋 구할 수 있는 곳

번호	권리번호	연도	일련번호	자릿값
출원번호	2자리	4자리	7자리	-
공개번호	2자리	4자리	7자리	-
공고번호	2자리	4자리	7자리	-
등록번호	2자리	-	7자리	0000

[문헌번호형식]

구분	권리번호	내 용
특허	10	아직까지 없었던 물건이나 그 물건을 만드는 방법을 최초로 발명한 것
실용신안	20	이미 발명된 것을 바꾸어서 보다 편리하고 쓸모 있게 만든 것
디자인	30	보는 사람으로 하여금 아름다움을 느끼도록 모양을 만드는 것
상표	40	제조회사가 자사 제품의 신용을 유지하고, 자기 상품을 다른 상품과 구별시키기 위해서 제품 및 포장에 표시하는 표장

[권리번호형식]

출처: http://kbbs.kipris.or.kr/kbbs/kr/faq.do?act=view&SEQ=28

4. Data acquisition 예

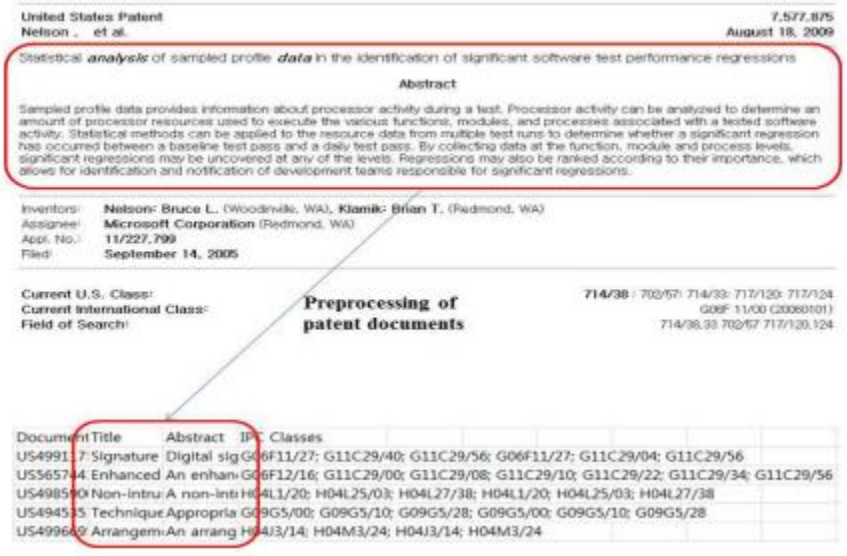
선진특허분류(CPC)코드

- CPC(Cooperative Patent Classification, 협력적 특허분류)는 IPC(International Patent Classification, 국제특허분류)보다 세분화된 특허분류체계로서 IPC(7만여 개소)보다 많은 26만여 개의 특허분류 개소를 갖고 있다.CPC는 효율적인 선행기술조사를 위해 미국특허청과 유럽특허청의 주도로 2012년 개발되었다.
- 2021년 현재 전 세계 중 30개 국가가 특허문헌을 CPC로 분류하고 있으며 특허문헌과 비특허문헌을 포함하여 전 세계 6000만 건 이상의 문헌이 CPC로 분류되어 있다.우리나라는 2015년 1월 이후 신규출원에 CPC, IPC를 함께 부여하고 있다.
  - https://www.kipo.go.kr/ko/kpoContentView.do?menuCd=SCD0200269
- 특허분류 조회
  - https://www.kipro.or.kr/business/cpcService

4. Data acquisition 예

특허문서의 데이터 셋

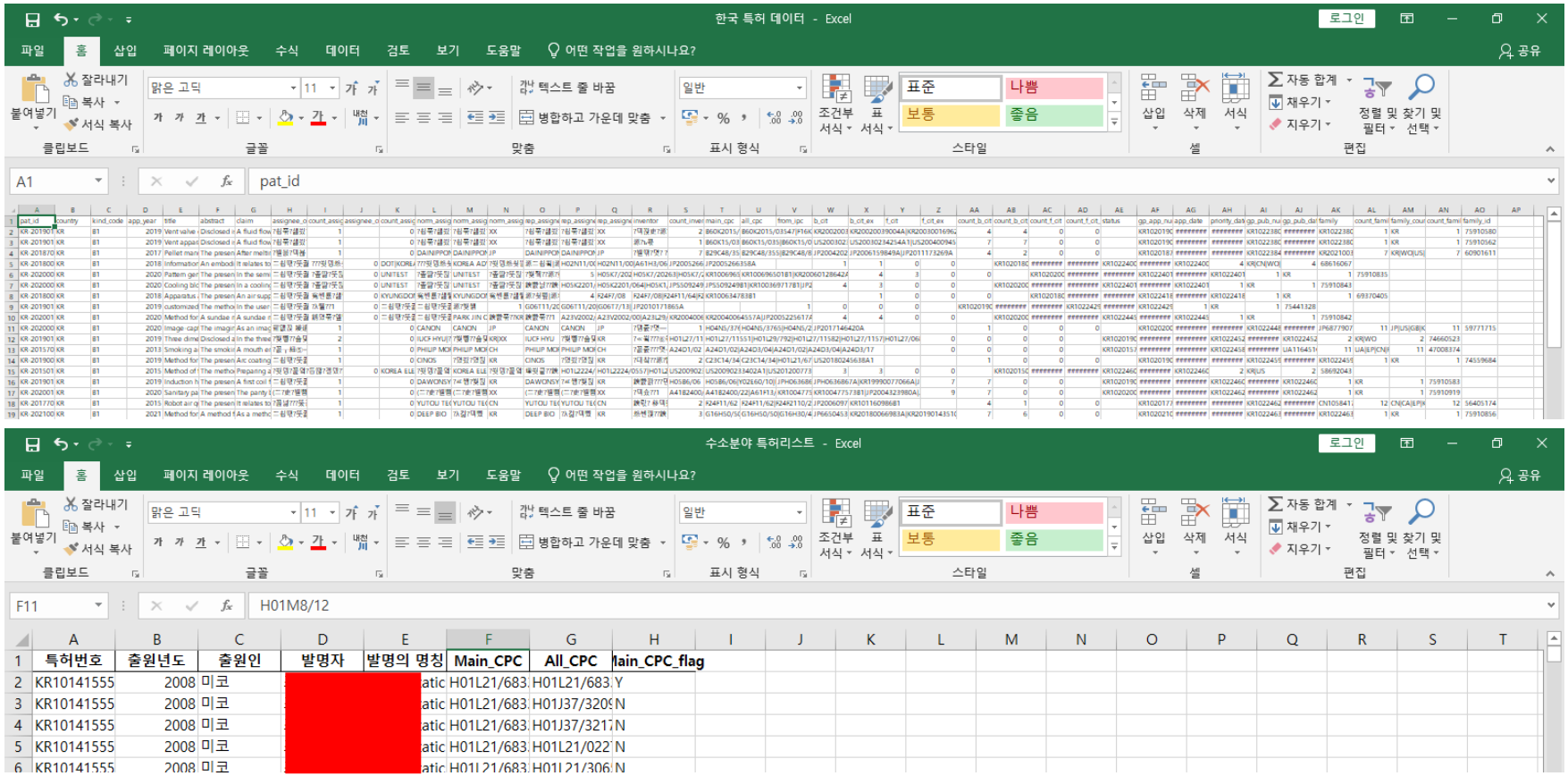
- 특허 데이터의 여러 유형 중에서 문자 데이터를 선택하여 분석한다. 특허 문서를 구성 하는 세부요소들 중에서 특허제목(title)과 기술요약정보 (abstract)만을 선택하여 별도의 데이터 셋(data set)을 구 축한다. 다음 그림은 본 연구에서 사용된 데이터 셋이다
- 특허문서의 데이터 셋
  - 특허문서 데이터는 전 세계에 존재하는 각국의 특허데이 터베이스가 레거시데이터가 되며 이곳으로부터 특정기술에 대한 특허문서를 검색하고 엑셀파일과 같은 특허데이터 셋 을 구축한다



4. Data acquisition 예

특허문서의 데이터 셋

- 데이터 셋: 엑셀파일 예



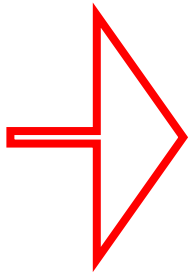
	A	B	C	D	E	F	G	H
1	특허번호	출원년도	출원인	발명자	발명의 명칭	Main_CPC	All_CPC	Main_CPC_flag
2	KR10141555	2008	미코	[redacted]		H01L21/683	H01L21/683	Y
3	KR10141555	2008	미코			H01L21/683	H01J37/320	N
4	KR10141555	2008	미코			H01L21/683	H01J37/321	N
5	KR10141555	2008	미코			H01L21/683	H01L21/022	N
6	KR10141555	2008	미코			H01L21/683	H01L21/306	N

4. Data acquisition 예

특허문서의 데이터 셋

- 특징 공학

pat_id	
country	
kind_code	
app_year	
title	
abstract	
claim	
assignee_original	
count_assignee_original	
assignee_current	
count_assignee_current	
norm_assignee_original	
norm_assignee_original_kr	
norm_assignee_country	
rep_assignee	
rep_assignee_kr	
rep_assignee_country	
inventor	
count_inventor	
main_cpc	
all_cpc	
from_ipc	
b_cit	
b_cit_ex	
f_cit	
f_cit_ex	
count_b_cit	
count_b_cit_ex	
count_f_cit	
count_f_cit_ex	
status	
gp_app_number	
app_date	
priority_date	
gp_pub_number	
gp_pub_date	
family	
count_family	
family_country	
count_family_country	
family_id	



Feature Engineering

특허번호	
출원년도	
출원인	
발명자	
발명의 명칭	
Main_CPC	
All_CPC	
Main_CPC_flag	

4. Data acquisition 예

알고리즘 선정

● 특성 유형 별 머신러닝/ 딥러닝 분류

머신러닝	지도학습	회귀	숫자값 예측	
		분류	항목 분류	
		추천 시스템과 랭킹학습	선호도 예측	
	비지도학습	군집화와 토픽 모델링	군집화 : 비슷한 데이터들을 묶어서 큰 단위로 만드는 기법 토픽 모델링 : 텍스트 데이터에 대한 토픽 관련 정도 표현	
		밀도 추정	관측 데이터로 부터 원래의분포 추측	커널 밀도 추정 , 가우스 혼합 모델
		차원 축소	데이터의 차원을 낮추는 기법	주성분 분석 (PCA), 특잇값 분해
	강화학습		기계가 환경과의 상호작용을 통해 장기적으로 얻는 이득을 최대화하도록 하는 학습 방법	
딥러닝	DNN	SLP, MLP		
	CNN	AlexNet,VGGNet,Inception,ResNet,Inception v4,DenseNet,Xception,MobileNet		
	RNN			
	LSTM			
	GAN			

4. Data acquisition 예

알고리즘 선정

- 문제 유형 별 머신러닝/ 딥러닝 분류

문제 유형	회귀문제	입력을 받아서 가장 적합한 숫자값 예측	주가 예측	선형 회귀 , 확장된 회귀분석 (ex : 다항회귀 , 비선형 회귀 , 벌점화 회귀 등),가우시안 프로세스 회귀 , 칼만 필터 , 인공 신경망 분석 (Artificial Neural Network) ,의사결정트리 (Decision Tree), 서포트 벡터 머신 (회귀) (Support Vector Machine (Regression)), PLS(Partial Least Squares), 앙상블 기법 (랜덤 포레스트 등)
	분류문제	주어진 입력에 대해 선택지 선택	단어로 기사 분류	로지스틱 회 , 인공 신경망 분석 (Artificial Neural Network) , CRF, RNN,의사결정트리 (Decision Tree) 서포트 벡터 머신 (Support Vector Machine),나이브 베이즈 (Naive Bayes) , 앙상블 기법 (랜덤 포레스트 등)
	군집화	주어진 입력을 비슷한 것끼리 군집	비슷한 문서 군집	K-means clustering, mean shift ,LDA(latent dirichlet allocation)
	차원축소 기법			
	연관관계분석			
	자율학습 인공 신경망 (SOM 등)			
	표현형학습	풀고자 하는 문제에 적합한 표현 추출	일기를 통해 글쓴이의 감정 파악	딥러닝 (단어 임베딩) 방법 : word2vec, 행렬 분해

4. Data acquisition 예

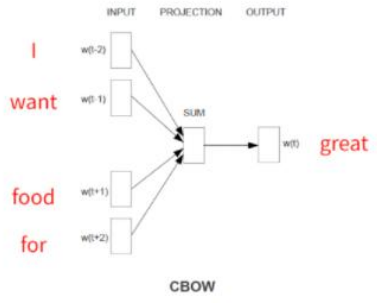
학습 모델

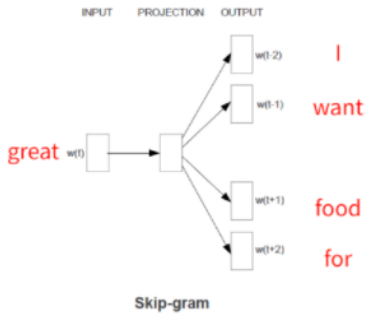
- Doc2Vec으로 특허 유사도 계산
  - 거리 유사도 측적인 유클리안 거리, 문서의 단어 분포 및 문서 집합 전체의 단어 분포를 사용하여 주제와 관련된 단어 확률을 계산하는 LDA(Latent Dirichlet Allocation), 단어빈도-역문서빈도인 TF-IDF(Term Frequency - Inverse Document Frequency), 벡터로 표현하고 가중치를 부여해 유사도 측정인 코사인 유사도, 자카드 계수, 피어슨 상관 계수가 존재한다.
- 두가지전처리 단계가 있을 수 있다.
  - 첫째, 불용어, 특수문자, 최소빈도 단어의 전처리는 정규 표현식, 불용어 제거, 최소빈도단어 제거로 첫째 전처리를 한다.
  - 둘째 전처리는 한국어기준 대표적으로 전처리하는 KoNLPy(Korean NLP in Python)가 있다. KoNLPy는 C/C++, JAVA에서 사용되던한국어 정보처리 라이브러리가 파이썬에서 구동할 수있게 만든 파이썬 패키지이며 패키지 내부에는 5가지한국어 정보처리 모듈이 있다. Hannanum, Kkma, Komoran, Mecab, Twitter 5가지 한국어 정보처리가 존재한다. KoNLPy를 사용해 띄어쓰기, 오타, 조사 제거, 서술어 현 재화를 하여 둘째 전처리를 끝낸다.
  - 전처리가 완료되면 컴퓨터가 한국어 문자를 이해하기 위해 벡터화가 필수적 이기에 단어 임베딩을 실시하여야 한다.

4. Data acquisition 예

학습 모델

- Doc2Vec으로 특허 유사도 계산
  - Word2Vec은 분산 표현 방 법을 채택하고 있다. 즉, 비슷한 위치에서 등장하는 단어 들은 비슷한 의미인 분포 가설을 기반으로 한다. 예를 들 면 바나나, 사과, 카메라의 단어가 존재한다고 하면 바나 나와 사과는 유사도 벡터값이 작고 바나나/사과, 카메라인 경우는 바나나/사과와 카메라의 유사도 벡터값의 차이가 크다. 현재 많은 기업들은 인건비를 줄이기 위해 챗봇을 도입하고 있으며 고객센터 용도로 많이 사용하고 있다. 따 라서 기업은 인건비를 줄일 수 있었으며 사용자들은 24시 간 내내 대기하는 챗봇을 통해 원하는 만족 내용을 들을 수 있다.
  - Word2Vec은 Skip-gram과 CBOW(Continuous Bag of Words)로 나뉘게 된다. Skip-gram은 하나의 단어 에 대해 주변 단어 k개를 찾는 모델이다. CBOW는 주변 단어 k개를 사용하여 가운데 하나의 타겟 단어를 찾는 모델이다.





4. Data acquisition 예

학습 모델

- Doc2Vec으로 특허 유사도 계산
  - CBOW는 설명하면 문자로 되어있는 자연어는 원 -핫 인코딩하여 입력층에 입력이 된다. 이때 주변 단어 k 를 결정하는 것은 window size로 주변 단어를 결정한다. 다음 학습하기 위한 프로젝트 층에 각 단어에 0과 1로 이 뤄진 원-핫 인코딩 값과 가중치를 곱하고 더해 프로젝트 층에 입력이 된다. 마지막, 출력층으로 갈 때는 각 단어의 가중치와 원-핫 인코딩 값을 곱하고 주변 단어 k만큼 평 균과 소프트맥스를 취하여 스코어 벡터(출력값들이 총합 1이 나옴)이 출력된다. 마지막에 나온 스코어 벡터 값들은 cross-entropy함수를 사용하여 정답 단어와 오차를 줄여 정답 단어의 원-핫 인코딩 값이 나오게 되는 원리를 가지 고 있다.

레이블 인코딩

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

원-핫 인코딩

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

4. Data acquisition 예

구현 예

```
[ ] from gensim.models import doc2vec
```

```
[ ] model = doc2vec.Doc2Vec(vector_size=300, alpha=0.025, min_alpha=0.025, workers=8, window=8)

    # Vocabulary 빌드
    model.build_vocab(tagged_corpus_list)
    print(f"Tag Size: {len(model.docvecs.doctags.keys())}", end=' / ')

    # Doc2Vec 학습
    model.train(tagged_corpus_list, total_examples=model.corpus_count, epochs=50)

    # 모델 저장
    model.save('dart.doc2vec')
```

```
Tag Size: 2976 /
```

```
[ ] similar_doc = model.docvecs.most_similar('울산과학기술원')
    print(similar_doc)
```

```
[('주식회사 쉐어라이트', 0.34096598625183105), ('주식회사 나노엑스', 0.3376728892326355), ('안근택', 0.3362635970115661 6), ('주식회사 클로소사이언스', 0.3322572708129883), ('삼성전기', 0.31925415992736816), ('주식회사 마더스제약', 0.313985228538513 2), ('지
```

```
[ ] similar_doc = model.docvecs.most_similar('한국전기연구원')
    print(similar_doc)
```

```
[('김지호', 0.4067040681838989), ('주식회사 스포컴', 0.3959147930145263 7), ('주식회사 레이저모션테크', 0.3786007463932037 4), ('(주)한국비엠아이', 0.367114841938018 8), ('(주)닥터티제이', 0.351535499095916 75), ('(주)에임텍', 0.351435720920562 74), ('(주)제이티
```

4. Data acquisition 예

구현 후 향후 방향성

- 향후 단어 임베딩 부분에서는 Word2Vec 뿐만 아닌 GloVe, FastText도 사용하여 비교 분석하여 높은 성능 모델을 사용하는 것이다. 마지막으로 Attention Mechanism, LSTM 기반인 딥러닝 기술을 적용하기 위한 연구가 필요하다.

웹 스크래핑과 웹 크롤링

- 웹 스크래핑(web scraping)과 웹 크롤링(web crawling)의 차이점은 목표로 하는 특정 웹 페이지 유무이다.
  - 웹 스크래핑(web scraping) : 목표로 하는 특정 웹 페이지가 있다.
  - 웹 크롤링(web crawling) : 목표로 하는 특정 웹 페이지 없다.

5. 실습

실습1 : 공공포털 이용

- **문제:**
  - 행정안전부의 <공공데이터포털 : https://www.data.go.kr/index.do> 과 같은 공공포털 등 직접 관찰한 웹사이트를 찾아 구축된 분석용 데이터 사례를 제시하고, 데이터 취득

5. 실습

실습2 : 한글 웹 페이지 크롤링 후 저장

- **문제:**
  - 도시별 현재날씨 확인하기
  - https://pythondojang.bitbucket.io/weather/observation/currentweather.html

5. 실습

실습3 : 한글 웹 페이지 크롤링 후 저장

- **문제:**
  - 최근 1년 동안 판매량 기준으로 인기 있는 국산 자동차 브랜드를 알고 싶다.
  - https://auto.danawa.com/

실습 4 : Web 스크래핑

- **문제:**

네이버 증권 데이터 수집

6. Data Ingestion

Why Data Ingestion?

…
What is data acquisition? We define it as this:

Data acquisition is the processes for bringing data that has been created by a source outside the organization, into the organization, for production use.

Prior to the Big Data revolution, companies were inward-looking in terms of data. During this time, data-centric environments like data warehouses dealt only with data created within the enterprise. But with the advent of data science and predictive analytics, many organizations have come to the realization that enterprise data must be fused with external data to enable and scale a digital business transformation.

This means that processes for identifying, sourcing, understanding, assessing and ingesting such data must be developed.

This brings us to two points of terminological confusion. First, "data acquisition" is sometimes used to refer to data that the organization produces, rather than (or as well as) data that comes from outside the organization. This is a fallacy, because the data the organization produces is already acquired.

Second, the term "ingestion" is often used in place of "data acquisition." Ingestion is merely the process of copying data from outside an environment to inside an environment and is very much narrower in scope than data acquisition. It seems to be a term that is more commonplace, because there are mature ingestion tools in the marketplace. (These are extremely useful, but ingestion is not data acquisition.)

출처 : https://www.firstsanfranciscopartners.com/blog/defining-data-acquisition-importance/

6. Data Ingestion

What is data ingestion?

Data ingestion is the process of obtaining and importing data for immediate use or storage in a database.

...

Types of data ingestion
There are a few main ways to ingest data:

1. **Batch processing.** In batch processing, the ingestion layer collects data from sources incrementally and sends batches to the application or system where the data is to be used or stored. Data can be grouped based on a schedule or criteria, such as if certain conditions are triggered. This approach is good for applications that don't require real-time data. It is typically less expensive.
2. **Real-time processing.** This type of data ingestion is also referred to as stream processing. Data is not grouped in any way in real-time processing. Instead, each piece of data is loaded as soon as it is recognized by the ingestion layer and is processed as an individual object. Applications that require real-time data should use this approach.
3. **Micro batching.** This is a type of batch processing that streaming systems like Apache Spark Streaming use. It divides data into groups, but ingests them in smaller increments that make it more suitable for applications that require streaming data.

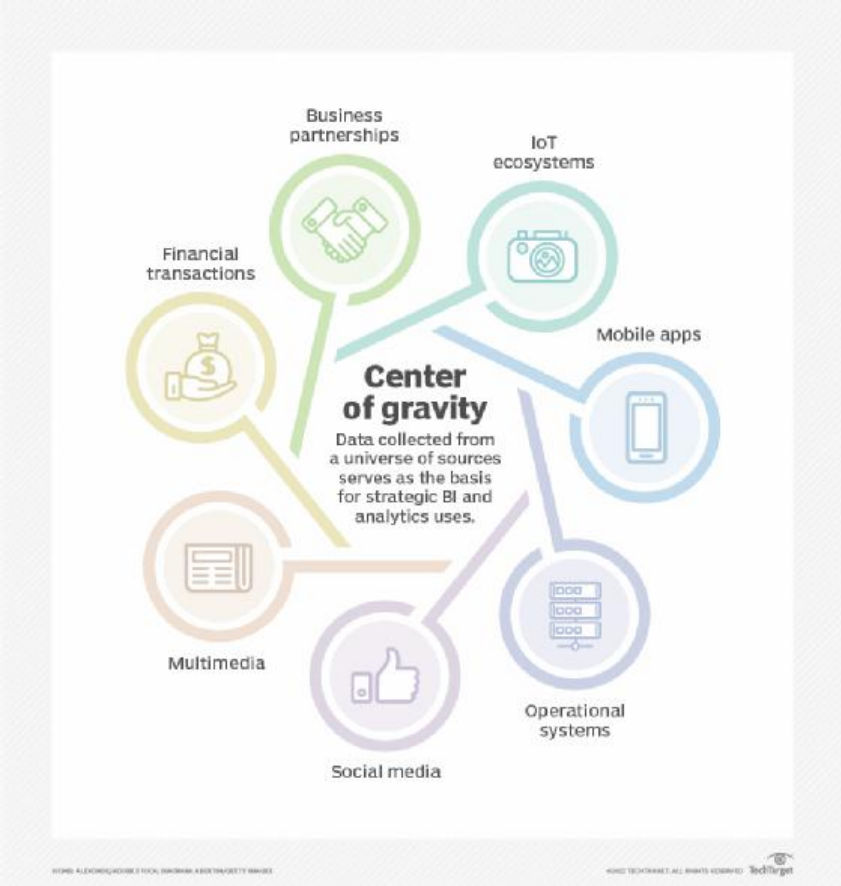
...

출처 : https://www.techtarget.com/whatis/definition/data-ingestion

6. Data Ingestion

What is data ingestion?

Data can be ingested from a variety of sources to be fed into the data pipeline.



출처 : https://www.techtarget.com/whatis/definition/data-ingestion

6. Data Ingestion

What is data ingestion?

Data ingestion tools and features
Data ingestion tools come with a range of capabilities and features, including the following:

- **Extraction.** The tools collect data from a variety of sources, including applications, databases and internet of things devices.
- **Processing.** The tools process data so it's ready for the applications that need it right away or for storage for later use. As mentioned earlier, a data ingestion tool may process data in real time or in scheduled batches.
- **Data types.** Many tools can handle different types of data, including structured, semistructured and unstructured data.
- **Data flow tracking and visualization.** Ingestion tools typically provide users with a way to visualize the flow of data through a system.
- **Volume.** These tools usually can be adjusted to handle larger workloads and volumes, and scale as the needs of the business change.
- **Security and privacy.** Data ingestion tools come with a variety of security features, including encryption and support for protocols such as Secure Sockets Layer and HTTP over SSL.

출처 : https://www.techtarget.com/whatis/definition/data-ingestion

6. Data Ingestion

What is data ingestion?

Benefits of data ingestion
Data ingestion technology offers the following benefits to the data management process:

- **Flexibility.** Data ingestion tools are capable of processing a range of data formats and a substantial amount of unstructured data.
- **Simplicity.** Data ingestion, especially when combined with extract, transform and load (ETL) processes, restructures enterprise data to predefined formats and makes it easier to use.
- **Analytics.** Once data is ingested, businesses can use analytical tools to draw valuable BI insights from a variety of data source systems.
- **Application quality.** Insights from analyzing ingested data enable businesses to improve applications and provide a better user experience.
- **Availability.** Efficient data ingestion helps businesses provide data and data analytics faster to authorized users. It also makes data available to applications that require real-time data.
- **Decision-making.** Businesses can use the analytics from ingested data to make better tactical decisions and reach business goals more successfully.

출처 : https://www.techtarget.com/whatis/definition/data-ingestion

6. Data Ingestion

What is data ingestion?

Challenges of data ingestion and big data sets
Data ingestion also poses challenges to the data analytics process, including the following:

- **Scale.** When dealing with data ingestion on a large scale, it can be difficult to ensure data quality and ensure the data conforms to the format and structure the destination application requires. Large-scale data ingestion can also suffer from performance challenges.
- **Security.** Data is typically staged at multiple points in the data ingestion pipeline, increasing its exposure and making it vulnerable to security breaches.
- **Fragmentation and data integration.** Different business units ingesting data from the same sources may end up duplicating one another's efforts. It can also be difficult to integrate data from many different third-party sources into the same data pipeline.
- **Data quality.** Maintaining data quality and completeness during data ingestion is a challenge. Checking data quality must be part of the ingestion process to enable accurate and useful analytics.
- **Costs.** As data volumes grow, businesses may need to expand their storage systems and servers, adding to overall data ingestion costs. In addition, complying with data security regulations adds complexity to the process and can raise data ingestion costs.

출처 : https://www.techtarget.com/whatis/definition/data-ingestion

6. Data Ingestion

What is data ingestion?

Data ingestion vs. ETL
Data ingestion and ETL are similar processes with different goals.

- Data ingestion is a broad term that refers to the many ways data is sourced and manipulated for use or storage. It is the process of collecting data from a variety of sources and preparing it for an application that requires it to be in a certain format or of a certain quality level. In data ingestion, the data sources are typically not associated with the destination.

- Extract, transform and load is a more specific process that relates to data preparation for data warehouses and data lakes. ETL is used when businesses retrieve and extract data from one or more sources and transform it for long-term storage in a data warehouse or data lake. The intention often is to use the data for BI, reporting and analytics.

출처 : https://www.techtarget.com/whatis/definition/data-ingestion

6. Data Ingestion

Data Ingestion 하기 전

- 데이터 과학자가 되려면 데이터가 필요하다.
  - 인터넷에 들어가면 수많은 데이터를 수집할 수 있고, 명령 줄에서 Python 스크립트를 실행하는 경우 sys.stdin 및 sys.stdout을 사용하여 이를 통해 데이터를 가져올 수 있다. 더욱 더 스마트폰의 보급과 소형 센서의 발달은 수많은 정보를 실시간으로 수집할 수 있는 환경을 마련하는데 기여했다.
  - 인터넷에는 다양한 데이터가 있으나, 이것을 알고만 있어서는 데이터를 활용할 수 없다. 새로운 데이터를 처음 확보 할 때 데이터 냄새 테스트를 수행하여 데이터를 신뢰할지 여부와 신뢰할 수 있는 정보 소스인지 여부를 결정해야 한다.

  - 수집된 가치는 다음 3가지 항목을 유지할 수 있을 때 획득할 수 있다.
    - Readability(가독성)
    - Cleanliness(청결성)
    - Longevity(수명)

6. Data Ingestion

Data Ingestion 하기 전

- 데이터를 신뢰할지 여부와 신뢰할 수 있는 정보 소스인지 여부를 판단하기 위해 다음 질문을 할 수 있다.

① 질문이나 우려 사항이 있는 경우 연락 할 수 있는 확실한 출처가 있는가?
② 데이터가 정기적으로 업데이트되고 오류가 있는지 확인 가능한가?
③ 데이터가 수집 된 방법 및 유형에 대한 정보와 함께 제공됩니까?
④ 수집에 샘플이 사용되었습니까?
⑤ 데이터 세트를 확인하고 검증 할 수 있는 다른 데이터 소스가 있습니까?
⑥ 주제에 대한 전반적인 지식을 감안할 때 이 데이터가 그럴듯하게 보입니까?
⑦ 소스 최신 방법 및 릴리스 확인이 가능한가?
⑧ 비교를 위한 다른 좋은 출처가 있는가?
⑨ 소스 및 / 또는 데이터를 결정하기 위해 주제를 더 조사할 수 있는가?

7. 실습

실습5 : 데이터 인제스트

- 문제:
  - [automobile_sales(2025.01-2205.08).txt] 업로드
  - 이 txt파일은 자동차 판매실적 테이블의 html 이다. html 코드는 모두 제거하고 [순위, 모델, 판매량, 점유율, 이미지 url]만 추출하고 싶어. 데이터를 추출해서 csv 파일로 정리해서 다운로드할 수 있게 해줘.
  - 추출된 파일을 요약해줘

8. Data gathering 고려사항

고려사항

- 수집 가능성
  - 서비스 활용에 아무리 좋은 데이터가 있다고 가정하더라도 수집이 불가능하거나 통제 불가능한 주기를 가지고 있다면 서비스 활용을 원천 데이터의 정책에 의존하게 되므로 바람직하지 않음
  - 수집이 아무리 용이하더라도 서비스 활용측면에서 데이터를 활용하기 위한 전처리・후처리에 비용이 많이 들어가게 되면 좋은 데이터 선정이라 할 수 없음
- 개인정보보호 및 저작권 문제
  - 수집된 데이터에 대해 개인정보보호 문제나 저작권에 대한 문제 발생시 서비스 활용에 대해 심각한 문제가 발생하므로 반드시 살펴보아야 함
- 데이터의 정확성
  - 수집한 데이터의 정확성은 서비스의 활용목적에 세부항목이 정확히 존재하는가에 대해 검토 필요
  - 수집목적에 맞는 데이터를 수집하기 위해서는 사전처리 과정이 필요하고 수집한 데이터의 사후처리 방안 필요

8. Data gathering 고려사항

고려사항

- 수집 난이도
  - 수집 난이도는 데이터 수집 및 처리에 들어가는 구축비용Ongoing Cost이 많이 들어갈 경우와 데이터 수집의 분석 • 설계와 필요한 데이터를 얻기 위해 많은 정제 과정이 필요할 경우로 구분하여 대안을 고려
  - 정성적 기준으로 주로 비용으로 직접 산출이 어려운 경우 수집난이도 측면에서 트래픽량과 저장처리
  - 장치의 용량 등이 고려대상이며, 수집 대상의 대안을 찾아야 함
- 수집비용
  - 수집비용은 데이터를 획득하기 위해 직접적으로 들어가는 획득비용
  - 정량적 기준으로 적용된 수집기술에 들어가는 비용이 발생할 경우에는 수집기술에 대한 검토가 필요

8. Data gathering 고려사항

고려사항

- 새로운 데이터를 처음 확보 할 때 데이터 냄새 테스트를 수행하여 데이터를 신뢰할지 여부와 신뢰할 수있는 정보 소스인지 여부를 결정해야 함.

1. 질문이나 우려 사항이 있는 경우 연락 할 수 있는 확실한 출처가 있는가?
2. 데이터가 정기적으로 업데이트되고 오류가 있는지 확인 가능한가?
3. 데이터가 수집 된 방법 및 유형에 대한 정보와 함께 제공됩니까?
4. 수집에 샘플이 사용되었습니까?
5. 데이터 세트를 확인하고 검증 할 수 있는 다른 데이터 소스가 있습니까?
6. 주제에 대한 전반적인 지식을 감안할 때 이 데이터가 그럴듯하게 보입니까?
7. 소스에 연락하고 최신 방법 및 릴리스 확인이 가능한가?
8. 비교를 위한 다른 좋은 출처가 있는가?
9. 소스 및 / 또는 데이터를 결정하기 위해 주제를 더 조사할 수 있는가?

THANK YOU.

앞으로의 엔지니어는 단순한 '코더'나 '기계 조작자'가 아니라 뇌-기계 인터페이스를 통해 지식과 능력을 즉각 확장하는 존재(뉴로-인터페이스: Neuro Interface)가 될 수 있습니다.

🎯 목표 달성을 위한 여정이 시작됩니다.
✨ 궁금한 점이 있으시면 언제든 문의해주세요!
🚀 함께 코더와 프롬프트 전문가로 성장해 나갑시다!