SAMSUNG SW AI ACADEMY FOR YOUTH

함께가요 미래로!
Enabling People

삼성청년 SW·AI아카데미

관통 프로젝트:
실시간 데이터 수집 및 데이터베이스 연동

SAMSUNG SW AI ACADEMY FOR YOUTH

함께가요 미래로!
Enabling People

실시간 데이터 수집 및 데이터베이스 연동

챕터의 포인트

- 관통 프로젝트 안내

SAMSUNG SW AI ACADEMY FOR YOUTH

함께가요 미래로!
Enabling People

관통 프로젝트 안내

관통 프로젝트 안내

Confidential

- **목표**
  - 해당 프로젝트의 목표는 실시간 데이터를 직접 적재하여 컨텐츠 기반의 서비스를 제공하는 환경을 구축하는 것입니다.
  - 해당 차시 프로젝트의 목표는 최종 프로젝트 주제에 맞는 데이터를 실시간으로 수집하고, PostgreSQL에 저장하여 서비스에 필요한 데이터를 적재할 수 있는 구조를 갖춘 데이터 수집 환경을 구현하는 것입니다.
  - 본 프로젝트를 통해 라이브 데이터를 수집하고, PostgreSQL 데이터베이스와 연동하는 전 과정을 경험할 수 있습니다.
  - 이를 통해 실시간 데이터 처리 및 데이터베이스 연동에 대한 실전 경험을 쌓고, 향후 데이터 기반 의사결정에 활용할 수 있는 기술 역량을 강화할 수 있습니다.
  - 다양한 출처의 데이터를 웹 상에서 크롤링 등을 활용하여 자유롭게 수집할 수 있다.
  - PostgreSQL 등의 RDB를 이용한 데이터베이스 연동 및 저장할 수 있다.
  - 데이터 수집 과정에서 발생하는 문제를 해결하고, 실전 프로젝트를 통해 경험을 축적할 수 있다.

관통 프로젝트 안내

Confidential

- 준비사항
  - 사용 데이터
    - 실시간으로 제공되는 데이터 플랫폼의 어떤 데이터라도 무방
      ('실시간, 준 실시간으로 생성되는 데이터' 라는 속성은 유지되어야 합니다.)
    - 웹 크롤링을 통해 수집한 데이터
  - 개발언어/프로그램
    - Python: 데이터 수집, 전처리 및 연동 구현
    - PostgreSQL 등의 RDB: 데이터 저장 및 관리
    - feedparser, requests, BeautifulSoup 등의 라이브러리

• 구현 방법

1) 실시간 데이터 수집

- 데이터를 실시간으로 수집 (ex. RSS 피드를 활용하여 실시간으로 업데이트되는 컨텐츠 수집)
- Python의 사용해 파싱 및 각 항목 추출 (BeautifulSoup을 활용하는 등)
- 중복 전송 방지를 위한 체크 포인트(ex. News의 url)를 잡은 후, 수집한 데이터를 바로 데이터베이스에 저장

관통 프로젝트 안내

Confidential

• 관통 프로젝트 가이드

This XML file does not appear to have any style information associated with it. The document tree is shown below.


```
▼<rss xmlns:dc="http://purl.org/dc/elements/1.1/" version="2.0">
  <script/>
  <script/>
 ▼<channel>
    <title>경향신문:전체기사</title>
    <link/>
    <description>경향신문 RSS 서비스 | 전체기사</description>
    <lastBuildDate>2025-04-07T21:19:01+09:00</lastBuildDate>
    <copyright>Copyright (C)1996 Kyunghyang Shinmun, All right reserved.</copyright>
    <webMaster>webmaster@khan.co.kr</webMaster>
    <language>ko</language>
    <image/>
   ▼<item>
    ▼<title>
       <![CDATA[ 시장 혼란이 "무역질서 재편 일환" …미 당국자들, 상호관세 옹호 나서 ]]>
     </title>
     <link>https://www.khan.co.kr/article/202504072115015/?utm_source=khan_rss&utm_medium=rss&utm_campaign=total_news</link>
    ▼<description>
       <![CDATA[ 공화당 일각선 "제동 필요" 도널드 트럼프 미국 행정부 고위 당국자들이 주말인 6일(현지시간) 언론과 연쇄 인터뷰를 통해 상호관세 정책을 옹호했다. 세계 각국의 반발과 금융시장 혼란을 야기한 상호관세에 대해 "무역질서 재편의 일환"이라며 강행 의지를 밝혔다. 또한 물가 상승이나 경기침체 전망에 선을 그으면서 파장 축소도 시도했다. 하워드 러트닉 · · · ]]>
     </description>
     <dc:date>2025-04-07T21:15:01+09:00</dc:date>
    ▼<author>
       <![CDATA[ 워싱턴 | 김유진 특파원 yjkim@kyunghyang.com ]]>
     </author>
    ▼<category>
       <![CDATA[ 미국 · 중남미 ]]>
     </category>
   </item>
   ▼<item>
    ▼<title>
       <![CDATA[ 월가는 주말 반납…마트 · 쇼핑센터엔 사재기 행렬 ]]>
     </title>
     <link>https://www.khan.co.kr/article/202504072115005/?utm_source=khan_rss&utm_medium=rss&utm_campaign=total_news</link>
    ▼<description>
       <![CDATA[ 금융인들, 고객 리스크 점검 기업은 아시아 공급사와 연락 시민들 "공황상태…그냥 샀다" 도널드 트럼프 미국 행정부의 고율 상호관세 발표로 증시가 폭락하자 뉴욕 월가의 금융인들은 주말에도 일하며 분노와 불안, 좌절감을 토로했다고 뉴욕타임스(NYT)가 6일(현지시간) 보도했다. 미국 소비자들은 관세가 소비자가격에 전가돼 물가가 오를 것으로 예상하고 여 · · · ]]>
     </description>
     <dc:date>2025-04-07T21:15:00+09:00</dc:date>
```

- **구현 방법**

  2) 데이터베이스 연동

  - PostgreSQL 등의 관계형 데이터베이스의 기본 개념, 사용자 및 권한 관리, 데이터베이스 생성 및 설정 방법 이해
  - 데이터를 저장하기 위한 테이블 설계 원칙 및 데이터 무결성을 유지하는 방법 고민
    (예: UNIQUE 제약 조건 활용)
  - 확장 기능(pgvector) 설치와 활용을 통한 고차원 데이터(텍스트 임베딩 등)의 저장 전략 이해
    (이후 추가적으로 임베딩 기반의 서비스 구축을 위한 과정)
  - 데이터 적재 시, 중복 체크 및 에러 핸들링 전략 고려

• 관통 프로젝트 가이드

1.1. PostgreSQL 설치 (Linux - Ubuntu)

1. **PostgreSQL 설치**
   터미널에서 아래 명령어를 실행하여 PostgreSQL과 추가 패키지를 설치합니다.

```
sudo apt-get update
sudo apt-get install postgresql postgresql-contrib
```

2. **서비스 상태 확인**
   PostgreSQL 서비스가 정상 실행 중인지 확인합니다.

```
sudo service postgresql status
```

1.2. PostgreSQL 데이터베이스 설정

1. **PostgreSQL 접속**
   기본 사용자 `postgres`로 전환한 후 `psql` 셸에 접속합니다.

```
sudo -i -u postgres
psql
```

2. **데이터베이스 생성**
   `news` 데이터베이스를 생성합니다.

```
CREATE DATABASE news;
```

3. **사용자 생성 및 권한 부여**
   SSAFY 전용 사용자 `ssafyuser`를 생성하고, `news` 데이터베이스에 대한 모든 권한을 부여합니다.

```
CREATE USER ssafyuser WITH PASSWORD 'your_password';
GRANT ALL PRIVILEGES ON DATABASE news TO ssafyuser;
```

• 관통 프로젝트 가이드

4. 테이블 생성

i. 데이터베이스 변경

생성한 `news` 데이터베이스로 접속합니다.

```
\c news
```

ii. pgvector 확장 설치 (최초 한 번 실행) 및 테이블 생성

아래 SQL 명령어를 실행하여 `pgvector` 확장을 설치하고, `news_article` 테이블을 생성합니다.

```
-- pgvector 확장이 필요한 경우 (최초 한 번만 실행)
CREATE EXTENSION IF NOT EXISTS vector;

-- news_article 테이블 생성
CREATE TABLE news_article (
    id SERIAL PRIMARY KEY,
    title VARCHAR(200) NOT NULL,
    writer VARCHAR(255) NOT NULL,
    write_date TIMESTAMP NOT NULL,
    category VARCHAR(50) NOT NULL,
    content TEXT NOT NULL,
    url VARCHAR(200) UNIQUE NOT NULL,
    keywords JSON DEFAULT '[]'::json,
    embedding VECTOR(1536) NOT NULL
);
```

- **구현 방법**

  3) 시스템 통합

  - 수집부터 저장까지의 데이터 흐름과 각 단계 간 연동 방법에 대해 체계적으로 이해
  - 실시간 데이터 적재 환경에서 발생 가능한 문제(예: 데이터 지연, 중복 적재, 데이터 손실 등)와 그에 따른 해결책 모색
  - 시스템 전체 아키텍처를 설계하고, 각 구성 요소(데이터 수집, 전처리, 저장)의 역할 및 상호 연관성을 분석

관통 프로젝트 안내

Confidential

• 관통 프로젝트 가이드

1) PJT 내용

RSS 기반 실시간 컨텐츠 데이터 수집 및 DB 연동 프로젝트

목표: 실시간으로 구할 수 있는 데이터를 수집하고, PostgreSQL 데이터베이스에 연동하여 저장하는 환경을 구축

데이터 수집: (ex. 경향신문 RSS를 통해 실시간 컨텐츠 기사 데이터 수집)

데이터 저장: SQL의 테이블 설계, 데이터 무결성 확보 및 확장 기능 활용 방안 숙지

관통 프로젝트 안내

Confidential

• 요구사항

- 실시간으로 데이터를 수집하고, 이를 PostgreSQL 데이터베이스에 연동하여 저장함으로써, 전체 데이터 흐름과 저장 관리 원리를 이해

- 기본 기능
  - 실시간으로 수집한 컨텐츠 데이터를 바탕으로 데이터베이스 연동 및 저장 관리 원리를 이해

- 데이터 수집 및 저장 프로세스 활용
  - 데이터의 구조와 데이터 수집 원리 이해
  - 웹 크롤링을 통한 추출 기법 숙지
  - 수집 데이터의 중복 체크 및 오류 처리 원리 적용
- 요청조건
  - 실시간 컨텐츠 데이터 수집 이론적 배경
  - PostgreSQL 데이터베이스 구축 및 사용자, 권한 관리 이론 적용
  - 테이블 설계, 데이터 무결성 확보 및 확장 기능 활용 방안 이해
  - 수집-저장 과정에서의 데이터 흐름 및 문제 해결 전략 마련

관통 프로젝트 안내

Confidential

- 결과
  - 수집 및 전처리된 컨텐츠 데이터셋 (메타데이터와 본문 포함)
  - PostgreSQL에 저장된 컨텐츠 데이터 구조 및 설계도
  - 해당 내용을 Readme에 정리

관통 프로젝트 안내

Confidential

• 심화기능

- 두 개 이상의 데이터(사이트 다른 곳)를 수집하여, 모든 데이터를 동일한 포맷으로 저장하고 분석하는 기능

- 요청 조건
  - 하나의 사이트 등에서 데이터를 가져오는 것 이상으로 다른 포맷에서 동시에 하나의 데이터 테이블 형태로 적재 (ex.컨텐츠 수집에서 서로 다른 RSS 피드를 2개 이상 활용하여 데이터를 수집)
  - 제목, 본문, 날짜 등 핵심 필드를 동일한 스키마로 정의하고 저장
  - 데이터 중복 제거 및 통합 과정 수행

관통 프로젝트 안내

Confidential

- **산출물**
  - Gitlab에 올라온 baseline 코드 기반의 레포지토리 지속적 커밋
  - 이후 PJT도 백엔드, 프론트엔드 구현을 제외한 내용은 해당 레포지토리에 이어 나가면서 개발

- 정리

구현 기능	체크 포인트
실시간 컨텐츠 데이터 수집 및 전처리	웹 크롤링 데이터 중복 체크 및 오류 처리
PostgreSQL(RDBMS) 연동 및 데이터 저장	테이블 설계 및 데이터베이스 구축
데이터 정확성	데이터 무결성 확보 및 확장 기능 서로 다른 출처의 데이터 정확한 수집 여부 동일한 데이터 포맷으로 일관성 있게 저장

SAMSUNG
SW
AI ACADEMY
FOR
YOUTH

함께가요 미래로!
Enabling People

내일 방송에서 만나요!

삼성청년SW·AI아카데미