

**SK네트웍스 Family AI과정 12기  
 데이터 수집 및 저장 수집 데이터 보고서**



| **산출물 단계** | 데이터 수집 및 저장 |
| --- | --- |
| **평가 산출물** | 수집 데이터 보고서 |
| **제출 일자** | 2025년 7월 11일 |
| **깃허브 경로** | https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN12-FINAL-2TEAM |
| **작성 팀원** | 박진양, 김재현 |

1. **수집 데이터 개요**

| **데이터명** | **수집 대상** | **수집 목적** | **사용 예정 기능** | **출처/저작권** |
| --- | --- | --- | --- | --- |
| Yahoo Finance 뉴스 | S&P500·나스닥·ETF·섹터별 대표 종목 150여 종목 | 글로벌 금융시장 주요 종목 및 ETF의 최신 뉴스 데이터 확보 | VectorDB 저장 및 유사군 서치 | 출처 : Yahoo Finance  <https://finance.yahoo.com>  저작권 : 각 언론사 |
| FMP 금융 데이터 API | 미국 주식/ETF 실시간 시세, 재무정보, 차트 데이터 | 금융시장 데이터 실시간 확보 및 분석 | 주가 조회, 재무정보 분석, 투자지표 제공 | 주가 조회, 재무정보 분석, 투자지표 제공 |
| NewsAPI 뉴스 데이터 | 글로벌 주요 뉴스, 기사 | 최신 금융/경제 뉴스 수집 | 뉴스 알림, 기사 요약 | NewsAPI, 각 언론사 |
| OpenAI GPT-4 모델 | 텍스트, 요약, 번역 | 챗봇, 자동요약, 자연어 처리 기능 | 챗봇, 리포트 요약, 번역 | OpenAI |
| FRED 경제지표 데이터 | 미국 경제지표(금리, CPI 등) | 글로벌 거시 경제 데이터 수집 | 거시경제지표 조회, 분석 | FRED, 연준(미국) |

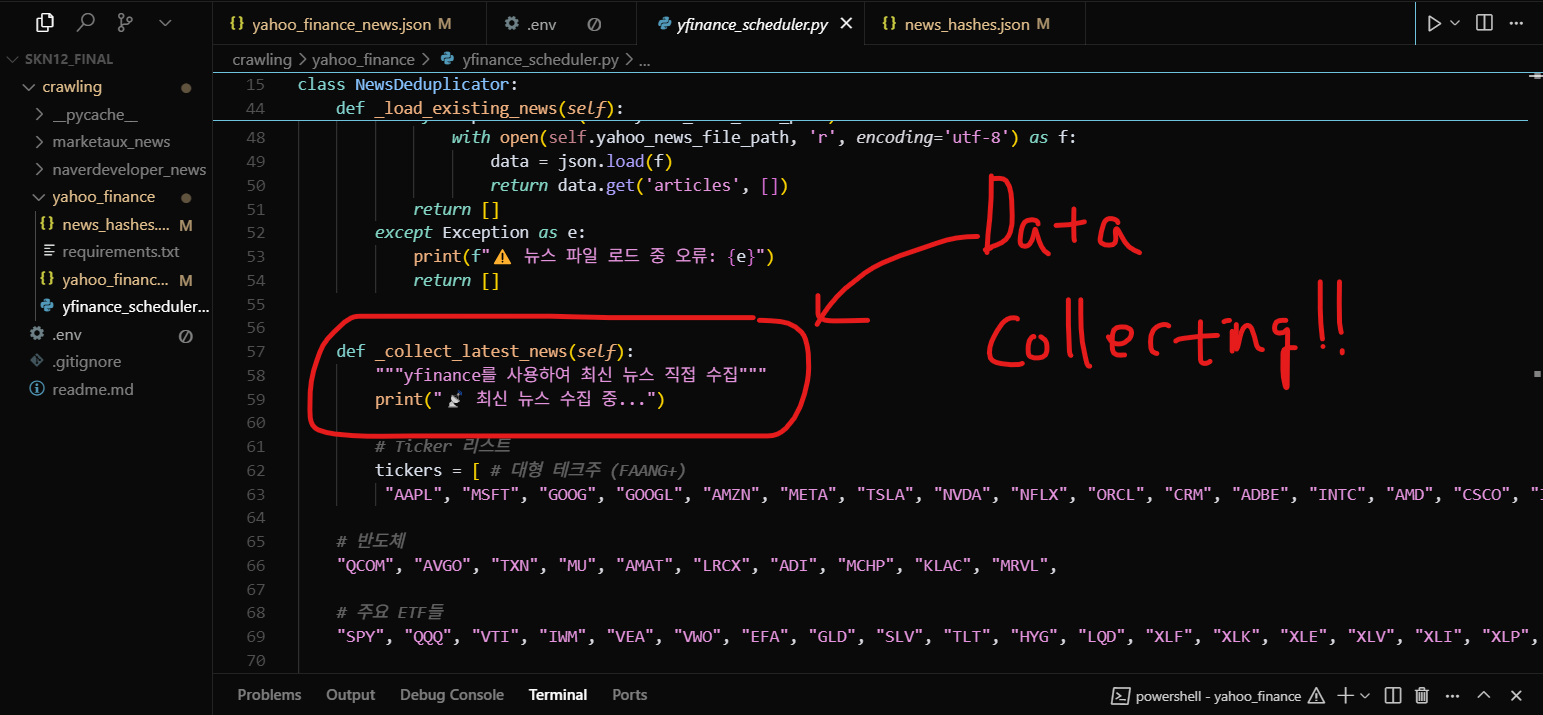
1. **수집 방법 및 자동화 절차**

* 수집 방식  
  + Yahoo Finance API: yfinance 라이브러리 활용
  + Scheduler를 활용하여 1분에 한번씩 데이터 갱신, JSON 파일로 확인 가능
  + FMP (Financial Modeling Prep):  
     HTTP REST API 요청(GET)으로 실시간/과거 주가, 재무 데이터 수집
  + NewsAPI:  
     HTTP REST API 요청(GET)으로 실시간 뉴스 기사 데이터 자동 수집
  + OpenAI:  
     HTTP REST API 요청(POST)으로 텍스트 생성·분석·요약 등 자연어 처리 결과 수집
  + FRED:  
     HTTP REST API 요청(GET)으로 미국 경제지표(금리, CPI 등) 데이터 수집
* 수집 도구 또는 스크립트 설명:  
  + 사용한 언어/라이브러리

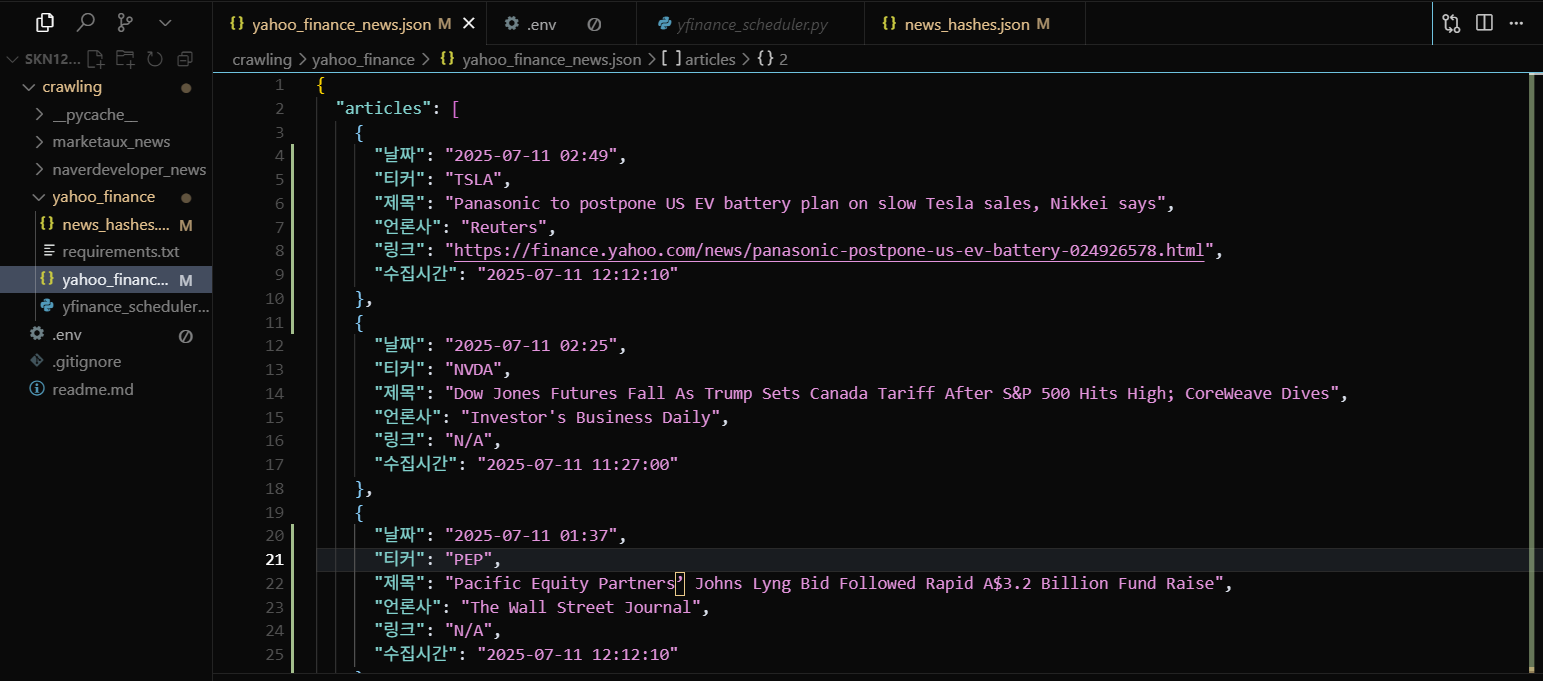
: Python + yfinance, pandas, numpy, schedule, requests, lxml, beautifulsoup4 등..

* + 자동화 여부 및 주기

: schedule 라이브러리를 이용해서 주기적으로 실행되도록 설계됨.  
 뉴스 데이터 호출 주기는 1분마다 1회 실행

* + 오류 발생 시 예외 처리 전략: 로깅
* 예시 스크립트 또는 흐름도 첨부:   
  

1. **데이터 설명 및 구성**

****

**데이터 포맷: JSON**

**구조: articles 배열 (뉴스 데이터 객체 리스트)**

**각 객체의 필드:**

* **날짜: 뉴스 발행 일자 (문자열)**
* **티커: 종목 코드 (문자열)**
* **제목: 뉴스 제목 (문자열)**
* **언론사: 뉴스 제공 언론사 (문자열)**
* **링크: 뉴스 URL (문자열)**
* **수집시간: 실제 수집 시각 (문자열)**

**4.1 파일 및 필드 설명**

| **파일명 또는 테이블명** | **필드명** | **데이터 타입** | **설명** | **예시** |
| --- | --- | --- | --- | --- |
| yahoo\_finance\_news.json | 날짜 | string | 뉴스 발행 일자 | "2025-07-11 02:49" |
| yahoo\_finance\_news.json | 티커 | string | Ticker | "TSLA" |
| yahoo\_finance\_news.json | 제목 | string | 뉴스 제목 | "Panasonic to postpone US EV battery plan on slow Tesla sales" |
| yahoo\_finance\_news.json | 언론사 | string | 뉴스 제공 언론사 | "Reuters" |
| yahoo\_finance\_news.json | 링크 | string | 뉴스 URL | "<https://finance.yahoo.com/news/panasonic-postpone-us-ev-battery-024926578.html>" |
| yahoo\_finance\_news.json | 수집시간 | string | 데이터 수집 시각 | "2025-07-11 12:12:10" |

**4.2 데이터 양**

* 전체 수집 데이터 건수: 2516건

**4.3 저장 위치 및 포맷**

* 저장 경로: SKN12\_FINAL/crawling/yahoo\_finance/yahoo\_finance\_news.json
* 저장 포맷: JSON
* 인코딩: UTF-8 / 기타

**5. 법적·윤리적 검토**

* 개인정보 포함 여부:  
  + [ ] 미포함
* 비식별화 조치 여부:  
  + 해당 없음  
    (개인정보가 없으므로 별도의 비식별화 조치 불필요)
* 출처 및 사용권:  
  + 공개 여부: 공개(비영리 연구, 분석, 내부 테스트 용도)
  + 라이선스 또는 약관 검토 여부:  
     검토 완료 (뉴스 콘텐츠 저작권은 각 언론사에 귀속되며, 인용 시 출처 표기 필요)
* 검토자 및 검토 일자: 박진양 / 2025년 7월 11일
* 비고:

**FMP (Financial Modeling Prep)**:  
 실시간 API 호출(REST API)로 금융 데이터 수집

**NewsAPI**:  
 실시간 API 호출(REST API)로 뉴스 기사 데이터 수집

**OpenAI**:  
 실시간 API 호출(REST API)로 자연어 처리 결과 획득

**FRED:** 실시간 API 호출(REST API)로 경제지표 데이터 수집

**6. 데이터 품질 및 정합성 관리 방안**

* 중복 제거 기준: 뉴스 제목을 MD5 해시로 변환하여 기존 해시값과 비교해 중복 제거

(예: hash(title) 기준 중복 식별)

* 정합성 검증 방법:  
  + 발행일, 제목, 링크 필수 필드 존재 여부 확인
  + 발행일 포맷 검증 (yyyy-MM-dd HH:mm)
  + 언론사 정보 누락 시 기본값 N/A 처리
* Null 처리 및 결측치 전략:
  + 필수 필드 누락 데이터는 수집 제외
  + 일부 비필수 필드(언론사, 링크)는 N/A로 대체 저장
* 표준화 전략:  
  + 발행일 타임존 일관성(UTC)
  + 텍스트 변환 및 특수문자 필터링은 별도로 수행 안함

**7. 변경 이력 및 보완 내역**

| **변경일** | **변경자** | **변경 내용** | **비고** |
| --- | --- | --- | --- |
| 2025-07-11 | 박진양 | 현재까지는 없음 |  |