SK네트웍스 Family AI과정 14기

데이터 수집 및 저장 수집 데이터 보고서

산출물 단계	데이터 수집 및 저장
평가 산출물	수집 데이터 보고서
제출 일자	2025.08.22
깃허브 경로	SKN14-Final-1Team
작성 팀원	김준기, 김재우, 안윤지, 이나경, 이원지희, 정민영

1. 수집 데이터 개요

데이터명	수집 대상	수집 목적	사용 예정 기능	출처/저작권
Google API Docs	Google API Docs 의 11개 API 공식 문서	프롬프트 튜닝, 모델 응답 테스트	키워드 검색, 문맥 기반 질의응답	Google for developers
회사 사내 문서	사원용 10장 대리용 10장 과장용 10장 부장용 10장 사장용 10장	사내문서 축적	사내 규정 기반 지침 QA, 권한 기반 문서 검색	OpenAI 합성 데이터셋 (자체 생성)

2. 수집 방법 및 자동화 절차

- 수집 방식
  - Google api 문서 : 웹 크롤링 → 텍스트 추출 → 텍스트 파일 저장
  - 사내문서 : OpenAI API 프롬프트 합성 → 텍스트 파일 저장

- 수집 도구 또는 스크립트 설명:
  - 사용한 언어/라이브러리

    - 구글 API 문서 크롤링 : python, Selenium, lxml, beautifulsoup 등

    - 사내 내부 문서 : python, openai 등
  - 자동화 여부 및 주기: 현재 1회 수집

    (이후 자동화 파이프라인 구축 필요)
  - 오류 발생 시 예외 처리 전략: 로깅

- 예시 스크립트 또는 흐름도 첨부 (이미지, 순서도 또는 코드)
  - 구글 API 문서 Selenium 웹 크롤링

```python
# ===== 메인 =====
try:
    full_start_url = ensure_hl_ko(urljoin(BASE_URL, START_URL))
    driver.get(full_start_url)

    print("사이드바의 링크를 수집 중…")
    urls_to_crawl = collect_sidebar_links()
    urls_to_crawl.insert(0, full_start_url)
    # 중복 제거
    urls_to_crawl = sorted(list(dict.fromkeys(urls_to_crawl)))
    print(f"총 {len(urls_to_crawl)}개의 유효한 페이지 링크를 수집했습니다.")

    for i, url in enumerate(urls_to_crawl, 1):
        try:
            print(f"\n({i}/{len(urls_to_crawl)}) 크롤링 중: {url}")
            driver.get(url)

            try:
                article_element = wait.until(EC.presence_of_element_located((By.TAG_NAME, "article")))
            except TimeoutException:
                # 폴백
                article_element = wait.until(EC.presence_of_element_located((By.TAG_NAME, "body")))

            final_page_text = extract_page_text()

            # 파일 저장
            path = url.split("?")[0].replace(BASE_URL, "")
            filename = re.sub(r'[/\\?%*:|"<>]', "_", path).strip("_") + ".txt"
            filepath = os.path.join(OUTPUT_DIR, filename)

            with open(filepath, "w", encoding="utf-8") as f:
                f.write(f"Source URL: {url}\n\n{final_page_text}")
            print(f"저장 완료: {filepath}")

        except Exception as e:
            print(f"페이지 처리 중 오류 발생: {url} - {e}")

        time.sleep(0.8)

finally:
    driver.quit()
    print("\n크롤링 완료! 브라우저를 종료합니다.")
```

○ 사내 내부 문서 OPENAI API 생성 흐름도

- **문서 사양 정의** → 직급별로 10개씩 카테고리·제목을 리스트에 정리.
- **프롬프트 생성** → 각 문서 사양에 맞춰 실행 지침과 섹션 구조를 포함한 프롬프트 작성.
- **OpenAI API 호출** → 프롬프트를 넣어 문서 본문을 생성.
- **파일 저장** → 생성된 문서를 직급/제목 규칙에 맞게 .txt 파일로 저장.
- **인덱스 집계** → 생성된 파일 목록을 INDEX.txt로 정리.
- **반복 구조** → 직급별(5개) × 문서별(10개) 루프를 돌며 전체 50개 문서를 자동 생성.

○ 사내 내부 문서 OPENAI API 생성 **프롬프트 코드**

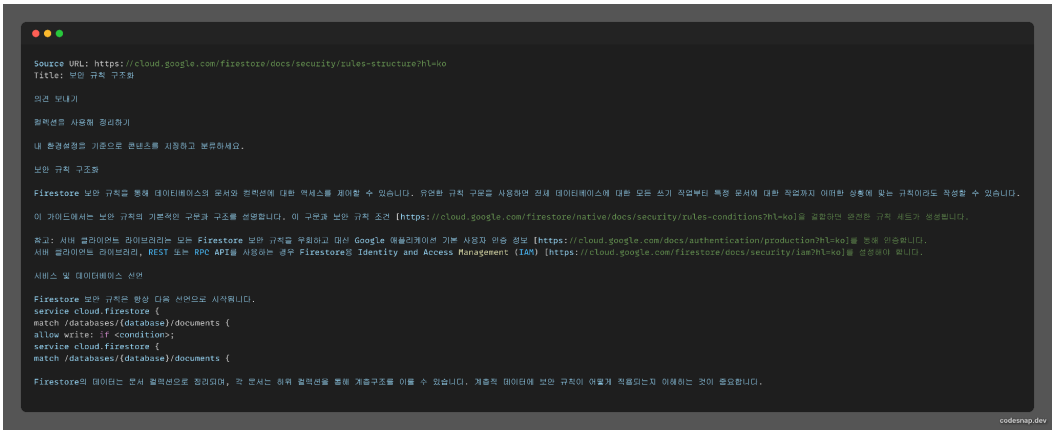
```
def make_user_prompt(category: str, title: str, today: str) -> str:
    base = f"""{title}
(분류: {category}) | 회사: M-Core | 버전: v1.0 | 작성일: {today}
{"-"*70}
작성 지침:
- 언어: 한국어, 평문 텍스트(마크다운/표 금지)
- 분량: A4 1장(약 550~750단어) 내외
- 독자: 사장(CEO). 전략/재무/법률/위기 의사결정 관점으로 작성
- 모든 수치·고유명사는 범위형 또는 더미 표기로 표기(예: XX억, YY% 등)
- 결론 중심의 결정 항목, 책임 주체, 검증 포인트를 포함
- 문서 끝에 '다음 개정 제안' 2~3줄
"""

    if category == "전사 경영 전략":
        body = dedent("""
            포함 섹션:
            1) 경영 요약(핵심 목표 3~5개, 우선순위/투자원칙)
            2) 전략 축(제품/시장/수익화/조직)과 자본 배분 원칙
            3) 분기 운영 프레임(OKR 상한/하한, 피벗 규칙)
            4) 주요 리스크/가정과 선제 대응
            5) 승인 라인/거버넌스(CEO/이사회/BU장 역할)
            6) 커뮤니케이션 원칙(대내/대외 메시지)
        """).strip()
```

3. 데이터 설명 및 구성

✔️ 파일 및 필드 설명

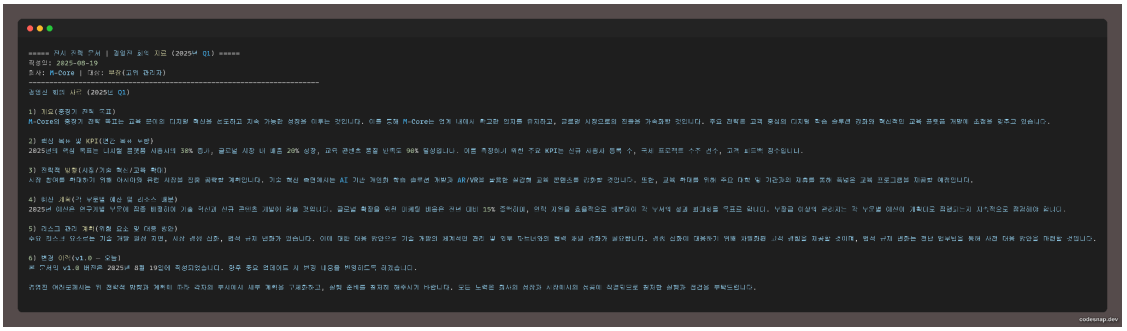
- 구글 API 문서 : txt 파일 (문서 상단에 Source URL / Title 포함)



- 구글 API QA 문서

파일명 또는 테이블명	필드명	데이터 타입	설명	예시
Google API Docs	-	TEXT	원문	
Google API Docs QA	question	string	질문 텍스트	Firestore 감사 로그에서 요청 호출자를 식별하기 위해 어떤 필드를 참조해야 하나요?
	answer	string	답변 텍스트	요청 호출자를 식별하기 위해 `AuditLog` 객체 내의 `AuthenticationInfo` 필드를 참조해야 하며, 여기에는 사용자의 `principalEmail`이 포함될 수 있습니다.
	sources	array of string	출처 URL 리스트	["https://cloud.google.com/firestore/docs/audit-logging?hl=ko"]
	tags	string	태그(주제/분류)	firebase
	last_verified	string	최종 검증일	2025-08-19
	source_file	string	원문 파일명	cloud.google.com_firestore_docs_audit-logging_hl=ko.txt

- 사내 내부 문서 : 아래와 같이 직급별 폴더 안에 문서별 **txt** 파일 존재



✔️ 데이터 양

- 전체 수집 데이터 건수:
  - 구글 API 문서: 약 2000개 문서 (txt 기준)
  - 회사 내부 문서: 50개의 txt 문서 (각 직급별 10개씩)
- 추출된 고품질 데이터 건수 (필터링 후 기준)
  - 구글 API 문서 QA: 약 10000개의 QA 데이터셋 (jsonl)

✔️ 저장 위치 및 포맷

- 저장 경로: SKN14-Final-1Team
- 저장 포맷: txt(문서) / JSONL(QA) / Vector DB (임베딩)
- 인코딩: UTF-8

5. 법적·윤리적 검토

- 개인정보 포함 여부
    - 미포함 (Google API 문서는 기술 문서로, 개인 식별 정보 없음)

- 비식별화 조치 여부
    - 해당없음 (개인정보가 없으므로 별도의 비식별화 불필요)

- 출처 및 사용권
    - Google API 공식 문서 활용
    - 원문 그대로 재배포하지 않고 QA 데이터셋으로 가공하여 내부 연구/교육 목적으로만 사용

- 공개 여부
    - 내부사용 한정

- 라이선스 또는 약관 검토 여부
    - Google 개발자 문서 이용 약관 및 robots.txt 확인 완료
    - 허용된 범위 내에서만 수집 및 활용

6. 변경 이력 및 보완 내역

변경일	변경자	변경 내용	비고
2025-08-20	김준기	구글 API 문서를 QA 형식으로 1차 전처리 완료	청킹 및 QA 가공 방식 변경 예정 (누락 없도록)
2025-08-21	정민영	사내 내부 문서를 QA 형식으로 전처리 예정	청킹 및 가공 방식 논의 필요