# FOSSLight: 오픈 소스, 오픈 소스 라이선스, 그리고 보안취약점을 관리해주는 오픈 소스 프로젝트

20220049 기채운, 20253882 전해광

## Problem Definition & Project Overview

### Problem Definition

오픈소스는 외부 데이터셋, 베이스 모델 참조, 파생 버전 배포 등으로 의존성이 복잡합니다. 이로 인해 사용자가 모든 출처의 라이선스를 수동으로 식별하고 검증하기 어려우며, 검증 누락 시 재현성 저하나 법적 리스크가 발생할 수 있습니다.

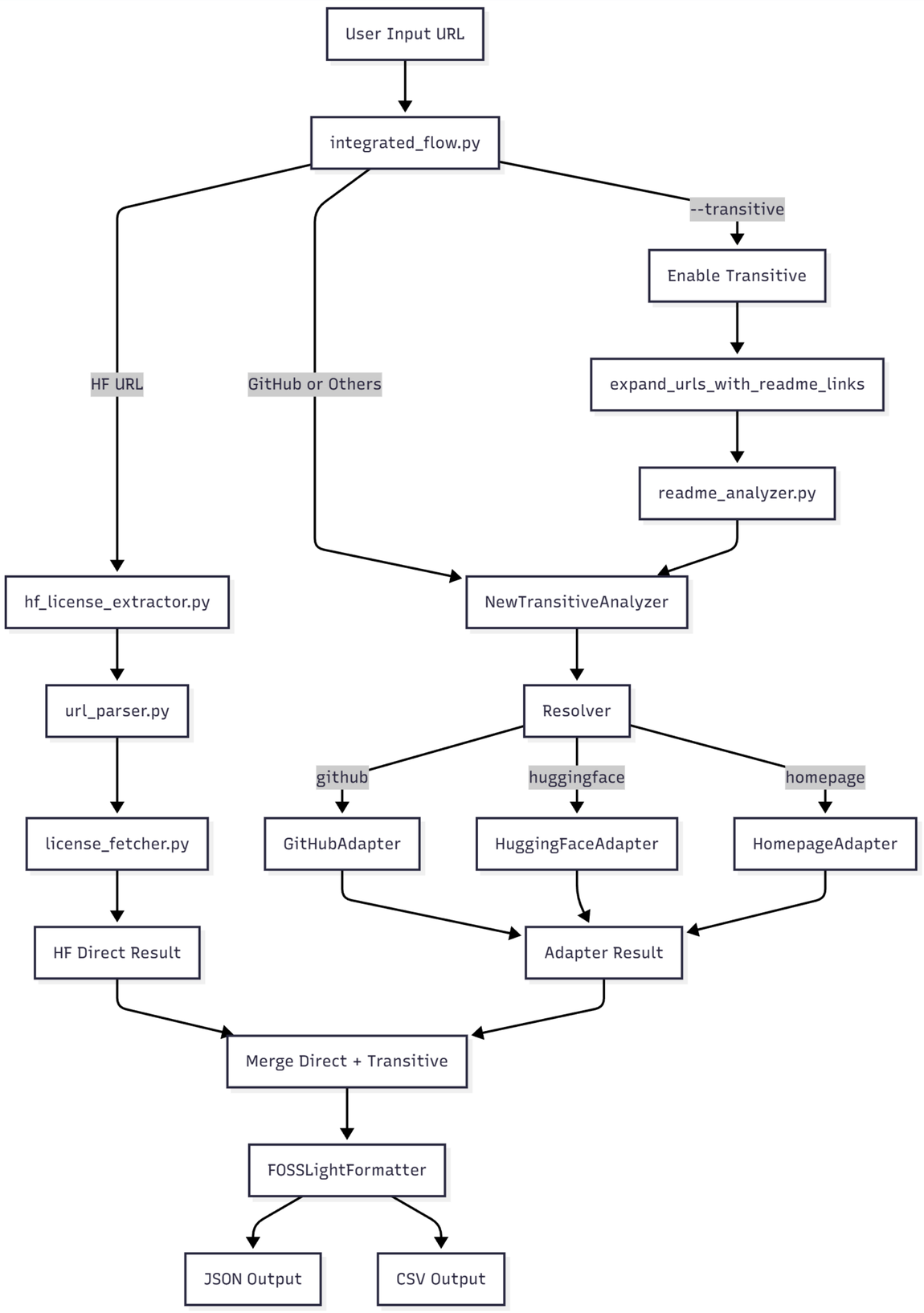
### Project Overview

- FOSSLight Dataset Scanner는 오픈소스 모델 및 데이터셋의 의존성과 라이선스를 자동 분석하는 프로젝트입니다.
- Hugging Face/GitHub URL 입력 시, README 분석과 링크 추적을 통해 참조 소스를 식별하고 재귀적으로 전체 의존성을 추적
- 이를 통해 모든 출처의 라이선스 정보를 종합적으로 제공하여, 개발자와 연구자의 수동 분석 비효율을 해소하고 법적 리스크를 감소시킵니다.

## Core Function Implementations

| | |
|---|---|
| **License 정보 추출** | • 사용자가 Hugging Face/GitHub 모델 URL을 입력하면 license 정보 추출 |
| **README 분석** | • LLM을 활용해 README 파일을 분석해서 데이터셋과 모델을 식별 |
| **재귀적 License 추적** | • 재귀적 분석을 수행해서 출처 모델/데이터셋의 모든 라이선스를 추적 |
| **종합적인 결과 제공** | • 각 모델 및 데이터셋의 출처 링크, 라이선스 정보, 의존성 정보를 통합 제공 |

## System Architecture



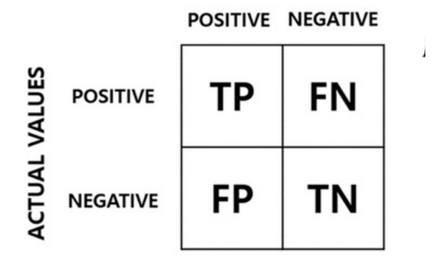

## Results

```
## 📋 Test Case Validation Results
```

| No. | Name | Expected License | Actual License | Main | References |
|---|---|---|---|---|---|
| 1 | kubernetes | Apache-2.0 | (none) | ✅ | – |
| 2 | protobuf | others | (none) | ✅ | – |
| 3 | fastapi | MIT | (none) | ✅ | ✅ |
| 4 | star_coder2 | Apache-2.0 | (none) | ✅ | – |
| 5 | ydata-profiling | MIT | (none) | ✅ | ✅ |
| 6 | matplotlib | other(다중) | (none) | ✅ | – |
| 7 | vlcj | GPL | (none) | ✅ | – |
| 8 | ansible | GPL-3.0 | (none) | ✅ | – |
| 9 | readme-generator | MIT | (none) | ✅ | – |
| 10 | stanford_alpaca | Apache License 2.0 | (none) | ✅ | ❌ |
| 11 | yolov5 | AGPL-3.0, Enterpris | (none) | ✅ | ✅ |
| 12 | stable-diffusion | CreativeML Open RAI | (none) | ✅ | ✅ |
| 13 | eleutherai_pile | other | other | ✅ | ✅ |
| 14 | alpaca_lora | MIT | mit | ✅ | ✅ |
| 15 | e5-base-v2 | MIT | mit | ✅ | – |
| 16 | pile | other | other | ✅ | – |
| 17 | multi_nli | cc-by-sa-3.0, cc-by | cc-by-3.0 | ✅ | – |
| 18 | mnist | MIT | mit | ✅ | – |
| 19 | stylegan3 | NVDIA Source Code L | (none) | ✅ | ✅ |
| 20 | FastChat | Apache-2.0 | (none) | ✅ | ✅ |
| 21 | bert | Apache-2.0 | (none) | ✅ | ✅ |

```
## 🏆 Final Results
   - Total test cases: 21
   - Passed (Main+References): 20
   - Failed (Main+References): 1
   - Accuracy: 95.24%

## 📊 Detailed Evaluation Metrics
   - True Positive (TP): 20
   - False Positive (FP): 0
   - False Negative (FN): 1
   - True Negative (TN): 0
   - Accuracy: 95.24%
   - Precision: 100.00%
   - Recall: 95.24%
   - F1 Score: 97.56%
```




$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

### Results

- 오픈소스의 URL과 참조 정보를 수집해 JSON 테스트 케이스셋 구축
- 라이선스 분석 모듈의 처리 결과와 실제 정보를 비교해 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1 score를 계산함
  (1) 라이선스 분석 정확도
  (2) 의존성 정확도