한국컴퓨터정보학회 동계학술대회 논문집 제32권 제1호 (2024. 1)

인공지능과 관련된 오픈 소스 파이썬 소프트웨어 프로젝트에서 자주 사용되는 파이썬 API들에 대한 연구

김정일[O]
[O]경북대학교 자율군집소프트웨어연구센터
e-mail: 2009307043@knu.ac.kr[O]

An Empirical Study on Frequently used Python APIs in AI-Related Open Source Python Software Projects

Jungil Kim[O]
[O]Center of Self-Organizing Software, Kyungpook National University

● 요 약 ●

전통 소프트웨어 프로젝트 개발과 AI 관련된 소프트웨어 프로젝트 개발에 큰 차이가 있어서 AI 관련된 소프트웨어 프로젝트 개발 환경을 이해하려는 많은 노력이 있었지만 AI 관련 소프트웨어 프로젝트 개발에서 어떤 API들이 자주 사용되는지에 대해서 아직 충분히 조사되지 않았다. 본 논문에서는 "AI 관련 오픈 소스 소프트웨어 프로젝트에서 어떤 파이썬 API들이 자주 사용되는가?"에 대한 연구 질문의 해답을 알아보는 경험 연구를 소개한다. 이 경험 연구의 결과로 AI 관련 오픈 소스 소프트웨어 프로젝트에서 파이썬 표준 라이브러리와 관련된 API들이 가장 자주 사용된다는 것을 확인했다. 또한 기계 학습을 포함해서 데이터 처리, 이미지 처리, 테스팅, 웹 서비스와 관련된 라이브러리들에 있는 API들도 AI 관련 오픈 소스 소프트웨어 프로젝트들에 자주 사용된다는 것을 알아냈다.

키워드: 인공지능, 소프트웨어 공학, 소프트웨어 저장소, 데이터 마이닝, 깃허브, 파이썬

I. Introduction

인공 지능 (Artificial Intelligence : AI) 응용 소프트웨어가 주목받으면서 AI와 관련된 소프트웨어 프로젝트 개발에 많은 관심이 쏠리고 있다. 기존 전통적인 소프트웨어 프로젝트 개발과 AI 관련된 소프트웨어 프로젝트 개발에 큰 차이가 있어서 AI 관련된 소프트웨어 프로젝트 개발 환경을 이해하려는 많은 노력이 있었지만 AI 관련 소프트웨어 프로젝트 개발에서 어떤 API들이 자주 사용되는지에 대해서 아직 충분히 조사되지 않았다 [1, 2].

본 논문에서는 "AI 관련 오픈 소스 소프트웨어 프로젝트에서 어떤 파이썬 API들이 자주 사용되는가?"에 대한 연구 질문의 해답을 알아보는 경험 연구를 소개한다. 깃허브에서 AI 관련 오픈 소스 소프트웨어 프로젝트 저장소들을 수집했다. 수집한 저장소들에 있는 파이썬 소스 파일들에서 API 모듈 임포트 문장과 API 호출 문장을 추출하고, 추출된 각 API 호출 문장이 파이썬 소스 파일들에 나타난 횟수를 계산해서 자주 호출된 API 들을 뽑았다. 그런 다음 그 API들을 API 범주로 분류했다. 그 분류 결과에서 파이썬 표준 라이브러리와 관련된 API들이 가장 자주 사용된다는 것을 확인했다. 또한 기계 학습을 포함해서 데이터 처리, 이미지 처리, 테스팅, 웹 서비스와 관련된 라이브러리들에 있는 API들도 AI 관련 오픈 소스 소프트웨어 프로젝트들에 자주 사용된다는 것을 알아냈다. 앞으로 추가 연구를 통해서 어떤 API 함수가 자주 사용되는지를 면밀히 조사해볼 것이다.

II. Related works

Islam 들 [1]은 딥 러닝 버그의 특징을 알아보기 위한 연구를 했다. 그들은 2716개 스택 오버플로우와 깃허브에서 유명한 딥러닝 관련된 5개 저장소들에서 수집한 버그 고침 커밋 (Bug fix commits) 500개를 조사해서 데이터 버그와 로직 버그가 가장 흔하게 나타난다는 것을 밝혔다. Zhang 들 [2]은 기계 학습 응용 소프트웨어들을 조사해서 기계 학습과 관련된 코드 냄새 (Code smells) 22가지 코드 냄새를 찾았다. Simmons 들 [3]은 데이터 과학 프로젝트들에서 쓰는 코딩 표준을 이해하기 위해서 1048개 오픈 소스 데이터 과학 프로젝트들과 1099개 전통 소프트웨어 프로젝트의 코딩 표준을 조사했다. 그 조사 결과로 전통 소프트웨어 프로젝트와 데이터 과학 프로젝트의 코딩 표준에 차이가 있다는 것을 찾았다. 본 연구는 파일 수준에서 AI 관련된 오픈 소스 소프트웨어 프로젝트의 파이썬 API 사용을 조사한 연구로써 이들 앞선 연구들과 차이가 있다.

III. Research method

1. Overview

이 연구의 전체 연구 방법은 <Fig. 1>에서 보여준다. 가장 먼저 깃허브에서 연구 대상이 되는 AI 관련 깃허브 저장소들을 수집한다. 그런 다음 그 연구 대상 저장소들에서 자주 쓰이는 파이썬 API들을 찾는다. 마지막으로 자주 쓰이는 파이썬 API들을 범주로 나눈다.

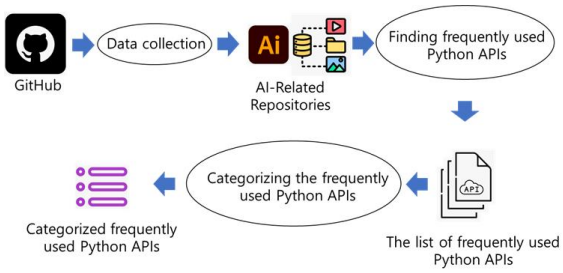

Fig. 1. 전체 연구 방법

2. Data collection

깃허브는 호스팅하고 있는 저장소들의 내부 데이터를 공유하기 위해서 다양한 API들을 제공한다. 이 연구에서는 연구 대상 AI 관련 깃허브 저장소를 수집하기 위해서 깃허브 검색 API (GitHub search API)를 사용했다. 앞선 연구들 [3,4]에서도 특정 깃허브 저장소 데이터를 수집하기 위해서 깃허브 검색 API를 썼다. 깃허브 검색 API는 주어진 질의 문자열(Query string)과 관련된 깃허브 저장소들을 최대 1000개 까지 찾아주는 웹 API이다. 이 연구에서는 "deep learning, machine learning, artificial intelligence, python 같은 AI와 관련이 있는 단어들을 깃허브 검색 API의 질의 문자열로 만들어서 깃허브 저장소들을 검색했다. Jebnoun 들의 연구를 따라서 [4] 검색 결과에 포함된 저장소들의 README.md 파일을 보고 다음 저장소들을 이 연구 대상에서 제외했다.

- 파이썬 언어로 작성되지 않은 저장소
- 파이썬 소스 파일이 없는 저장소
- AI와 관련되지 않은 소프트웨어 프로젝트 저장소

최종적으로 721개 저장소들을 이 연구의 대상 저장소들로 선정했다.

3. Finding frequently used Python APIs

파이썬 소스 파일 수준에서 파이썬 API 사용 정보는 두 가지로 나타난다. 파이썬 소스 코드 작성자는 특정한 파이썬 API를 사용하기 위해서 먼저 임포트 (Import) 문장으로 해당 API 모듈을 작업 중인 소스 파일에 포함시키고, 포함된 모듈에 있는 API 들을 호출하는 문장 (Call statement)를 작성한다. 이 두 문장이 직접적인 파이썬 API 사용 정보가 된다.

이 API 사용 정보를 바탕으로 이 연구에서는 AI 관련 저장소들인 연구 대상 저장소들에서 자주 쓰이는 파이썬 API 들을 조사했다. 연구 대상 저장소들에 포함된 모든 파이썬 소스 파일들에서 API 모듈 임포트 문장과 API 호출 문장을 체계적으로 추출하기 위해서 파이썬 표준 라이브러리인 ast 모듈을 썼다. 이 모듈은 파이썬 소스 파일 스트링을 파이썬 추상 구문 트리로 변환하고, 변환된 추상 구문 트리를 파싱하는 방문자(Visitor) API를 제공한다. 이 API들을 사용해서 연구 대상 저장소들에 있는 각 파이썬 소스 파일에 포함된 임포트 문장들과 API 호출 문장들을 추출했다. 그리고 다음 식을 바탕으로 추출된 각 파이썬 API의 사용 횟수를 계산했다.

$$APIUsage_a = \left|\{f_i | APICall_a \in f_i\}\right|$$

여기서, a는 파이썬 API, $APICall_a$는 파이썬 API 호출 문장, f_i는 파이썬 파일이다. 이 수식은 특정 파이썬 API의 사용 횟수를 파일 수준에서 정량적으로 나타낸다. 예를 들어, 파이썬 API list.append의 API 호출 문장이 10개 파이썬 소스 파일들에서 나왔다면 그 API의 APIUsage는 10이 된다.

위 식으로 연구 대상 저장소들의 소스 파일들에 포함된 파이썬 API들의 APIUsage를 계산한 다음 그 API들 가운데 자주 사용된 파이썬 API들을 결정했다. APIUsage 값으로 연구 대상 저장소들에 포함된 파이썬 API들을 내림차순으로 정렬하고, 그 정렬 결과에서 상위 1000개 파이썬 API들을 자주 쓰인 파이썬 API들로 결정했다.

4. Categorizing the frequently used Python APIs

이 연구에서 제시하는 연구 질문의 해답을 알아보기 위해선 3.3절에서 설명한 절차로 찾은 파이썬 API들을 이해할 수 있는 범주로 나누어 조사해야 한다. 이 연구에서는 수동 분류 방법으로 그 파이썬 API들을 API 범주별로 분류했다. 이 수동 분류 방법을 수행하기 위해서 컴퓨터 과학 대학원에 박사 과정으로 재학 중인 대학원생 1명을 초청해 이 논문의 주저자와 대학원생 2명으로 분류자 집단을 구성했다. 분류자 집단을 구성한 다음에 전체 분류 대상 API 집합을 두 부분으로 쪼개고 각 분류자에게 한 부분씩 할당했다. 그런 다음 분류자들은 자신에게 할당된 각 파이썬 API가 사용된 파이썬 소스 파일 부분과 임포트 문장을 직접 눈으로 확인하면서 그 대상 API의 모듈을 찾았다. 만약 그 대상 API의 모듈을 정확하게 찾았다면 그 모듈을 그 대상 API의 범주로 결정하고, 만약 모듈을 찾을 수 없거나 모호하다면 그 대상 API의 범주를 "알수 없음"으로 분류했다. 분류자들은 자신에게 할당된 파이썬 API들의 분류를 마치고 서로 할당된 API 집합을 바꿔서 같은 작업을 반복했다. 이 2라운드 수동 분류 작업의 결과를 코헨 카파 계수 (Cohen's Kappa Coefficient) 로 평가했다. 코헨 카파 계수는 두 명 이상의 분류자들이 분류한 결과를 얼마나 동의하는가를 나타낸다. 이 수동 분류 결과의 코헨 카파 계수는 92.6%이다. 이 수동 분류 결과에서 두분류자들이 "알수 없음"으로 분류한 API들은 이 연구의 분석 대상에서 제외했다.

IV. Result

<Table 1>은 3.4절에서 설명한 파이썬 API 수동 분류 결과이다. 이 연구에서 찾은 모든 API 범주들 가운데 파이썬 표준 라이브러리 (Python Standard Library : PSL)가 연구 대상 저장소들에서 가장 자주 사용된다는 것을 확인했다. 기계 학습과 관련된 라이브러리로 Tensorflow, PyTorch, PennyLane, Keras, neuralgym, mlflow,

ONNXRUNTIME, Gymnasium, transformers 들이 사용되고 있고, 그 중에서 Tensorflow와 PyTorch가 가장 자주 사용되고 있다는 것을 확인했다. 데이터 처리를 효율적으로 할 수 있도록 사용될 수 있는 서드 파티 유틸리티 라이브러리들 가운데에는 NumPy, JAX, Pandas, PyYAML, protobuf, flatbuffers, tqdm, click, albumentations, HDF5 들이 자주 사용되고, 그 중에서 NumPy가 가장 자주 사용된다는 것을 확인했다. 이미지 처리를 위해서 openCV, Pillow 라이브러리들이 그리고 그래프 그리기를 위해서 matplotlib 라이브러리가 각각 자주 사용되고 있다는 것을 알 수 있다. 그 외에도 테스팅, 웹 서버 구현을 위해서 pytest, Django, Flask 들이 사용된다는 것을 확인했다.

Table 1. 자주 사용된 파이썬 API들

API category	Description	#. Used APIs
PSL	파이썬 표준 라이브러리	197
Tensorflow	오픈 소스 기계학습 라이브러리	155
PyTorch	딥러닝 최적화된 텐서 라이브러리	133
NumPy	과학 컴퓨팅을 위한 기본 패키지	110
JAX	수치 함수를 변환하기 위한 기계 학습 프레임워크	30
matplotlib	NumPy 라이브러리를 활용한 그래프 라이브러리	25
PennyLane	양자 기계 학습, 자동 차별화 등을 위한 Python 라이브러리	17
openCV	오픈소스 컴퓨터 비전 라이브러리	16
pytest	테스팅 프레임워크	8
Pandas	오픈 소스 데이터 분석 및 조작 도구	5
Pillow	파이썬 이미징 라이브러리	4
Django	오픈 소스 웹 프레임워크	3
PyYAML	Python YAML 파서	3
protobuf	데이터 직렬화 도구	8
Keras	오픈 소스 신경망 라이브러리	3
flatbuffers	직렬화 형식 지원 Python 런타임 라이브러리	3
tqdm	진행 상태 표시 유틸리티 라이브러리	2
abseil	파이썬 용 Abseil 표준 라이브러리	2
click	명령줄 인터페이스 파이썬 라이브러리	2
Flask	오픈 소스 웹 프레임워크	1
albumentations	오픈 소스 이미지 확대 라이브러리	1
neuralgym	딥러닝 도구	1
HDF5	HDF5 데이터 포맷 패키지	1
google.cloud	구글 클라우드	1
mlflow	기계학습 라이프 사이클 관리 패키지	1
ONNXRUNTIME	크로스 플랫폼 기계학습 모델 가속기	1
Gymnasium	오픈 소스 강화 학습 라이브러리	1
transformers	딥 러닝 라이브러리	1
absl	파이썬 용 C++ absl 라이브러리	1

V. Discussion

1. Implication

이 연구의 결과로 AI 관련 소프트웨어 프로젝트 개발에 파이썬 표준 라이브러리와 기계 학습 라이브러리들이 가장 자주 사용되지만 그것들과 함께 데이터 처리, 그래프, 이미지 처리, 테스팅, 웹 프레임워크 라이브러리들이 사용된다는 것을 알았다. 이 결과는 소프트웨어 오류나 효율성 저하로 인한 유지보수 문제를 줄이기 위해서 AI 관련 소프트웨어 프로젝트 개발의 참여자들은 이들 라이브러리들의 사용 방법을 잘 이해할 필요가 있다는 것을 나타낸다. 또한 이들 라이브러리들에 익숙하지 못한 사용자들의 참여율과 개발 성공률을 높이기 위해서 이 라이브러리 사용을 지원할 수 있는 추천 도구를 개발할 필요가 있다.

2. Limitation

이 연구의 외부 타당성 위험은 이 연구에서 고려한 연구 대상 저장소 표본에 있다. 이 연구에서는 깃허브에 있는 AI 관련 소프트웨어 프로젝트 저장소들을 연구 대상 저장소로 선택했다. 이는 깃허브가 아닌 다른 호스트 서버에 있는 오픈 소스 AI 관련 소프트웨어 프로젝트들을 고려하지 못한다는 위험이 있다. 하지만 깃허브는 지금 가장 유명한 온라인 소프트웨어 프로젝트 호스트 웹 사이트이다. 또한 이 연구에서는 깃허브에서 인기가 많은 저장소들을 연구 대상으로 선정하도록 노력했다. 이러한 노력으로 이 외부 타당성 위험을 많이 줄였다고 생각한다.

VI. Conclusions

이 연구 결과로 파이썬 표준 라이브러리와 기계 학습 라이브러리들이 AI 관련 오픈 소스 소프트웨어 프로젝트 개발에 자주 사용된다는 것을 알았다. 또한 데이터 처리, 이미지 처리, 그래프 그리기, 테스팅, 웹 서비스와 관련된 라이브러리들도 함께 자주 사용된다는 것도 알았다. 이 연구 결과를 바탕으로 향후에 더 많은 AI 관련 오픈 소스 소프트웨어 프로젝트 표본을 대상으로 이 연구를 확장할 것이다. 특히 API 범주보다 더 낮은 수준에서 AI 관련 오픈 소스 소프트웨어 프로젝트 개발에 자주 사용되는 API들이 무엇인지 조사해볼 것이다.

ACKNOWLEDGEMENT

본 연구는 과학기술정보통신부의 재원 (1711194613, RS-2023-00213733)과 교육부의 재원 (NRF-2018R1A6A1A03025109)으로 한국연구재단의 지원을 받아 수행된 연구임.

REFERENCES

[1] Islam, M. J., Nguyen, G., Pan, R., & Rajan, H., A comprehensive study on deep learning bug characteristics. In: Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. 2019. p. 510-520.

[2] Zhang, H., Cruz, L., & Van Deursen, A., Code smells for machine learning applications. In: Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI. 2022. p. 217-228.

[3] Simmons, A. J., Barnett, S., Rivera-Villicana, J., Bajaj, A., & Vasa, R., A large-scale comparative analysis of coding standard conformance in open-source data science projects. In: Proceedings of the 14th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM). 2020. p. 1-11.

[4] Jebnoun, H., Ben Braiek, H., Rahman, M. M., & Khomh, F., The scent of deep learning code: An empirical study. In: Proceedings of the 17th International Conference on Mining Software Repositories. 2020. p. 420-430.