프로그램 소스코드 학습을 위한 오픈 소스 데이터 구축

빅데이터 혁신공유대학 데이터 셋 개발 프로젝트 공모전

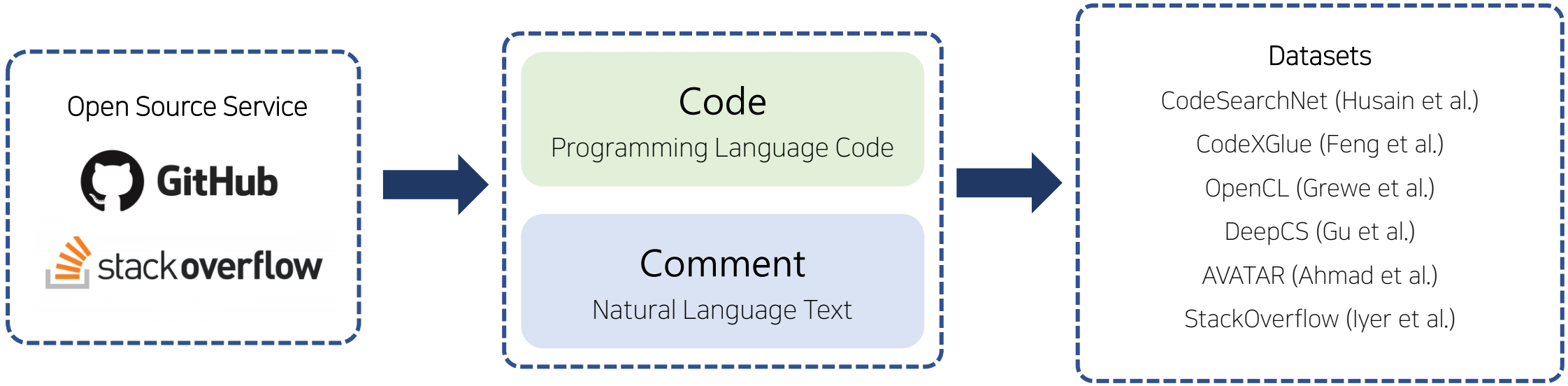
관련 연구

프로그램 이해는 소프트웨어 유지, 코드 검색, 코드 분류 등 소프트웨어와 관련된 작업

소프트웨어 프로젝트의 복잡성과 업데이트 빈도가 증가하면서 프로그램 이해의 중요성이 증가

최근 프로그램 이해와 관련된 연구는 활발한 연구 주제 → 다양한 데이터셋이 필요해짐

코드 주석은 프로그램 이해를 위한 가장 중요한 문서 형식으로 간주 → 대부분 코드-주석 데이터 활용



딥러닝 모델로 코드와 주석의 관계를 캡처하는 방법으로 다양한 분야에 활용 가능

관련 연구

CodeSearchNet Dataset

- 6가지 프로그래밍 언어 Python, Java, JavaScript, PHP, Ruby, Go 포함
- 210만 개의 bimodal 데이터, 640만 개의 unimodal 코드가 포함

TRAINING DATA	*bimodal* DATA	*unimodal* CODES
GO	319,256	726,768
JAVA	500,754	1,569,889
JAVASCRIPT	143,252	1,857,835
PHP	662,907	977,821
PYTHON	458,219	1,156,085
RUBY	52,905	164,048
ALL	2,137,293	6,452,446

Table 1: Statistics of the dataset used for training CodeBERT.

문제 인식

하지만 기존 데이터는 function-level의 데이터만 존재

지역적인 부분에서는 잘 동작하지만, project-level에 대한 이해가 부족

오픈소스 프로젝트 데이터를 구축하고 소스코드를 통해 학습할 수 있는 다양한 기계학습 모델의 훈련 및 개발에 활용

- 소스코드 기반 기계학습의 최종 종착점 = 프로그래밍 자동화
- 프로그래밍에 관련된 모델의 학습 및 개발에 활용 가능

데이터를 활용해 해결할 수 있는 문제

- 자동주석 생성, 소스코드 자동분류, 소스코드 자연어검색 등에 활용 가능
- 더 도전적인 연구로는 소스코드 자동생성 연구를 위한 학습데이터로 활용 가능

데이터 수집 전략

데이터 수집 시나리오

- GitHub, stack overflow 등의 오픈소스 아카이브 플랫폼을 크롤링하여 타켓 언어로 구현된 프로젝트를 수집

데이터 수집 장비(HW/SW)투입계획

- HW: 클라우드 서버를 사용하여 데이터 크롤링 진행
- SW: 크롤러는 오픈소스를 기반으로 개발팀에서 독자 개발

데이터 정제 및 분석 시나리오

- 수집된 프로젝트의 구조를 분석하고 특정 수준 이상의 project description과 source code comments가 포함된 프로젝트를 선별
- 선별된 프로젝트의 resource를 정의한 데이터 형식의 구조에 맞춰 JSON 형식으로 변환

데이터 분석 장비(HW/SW)투입계획

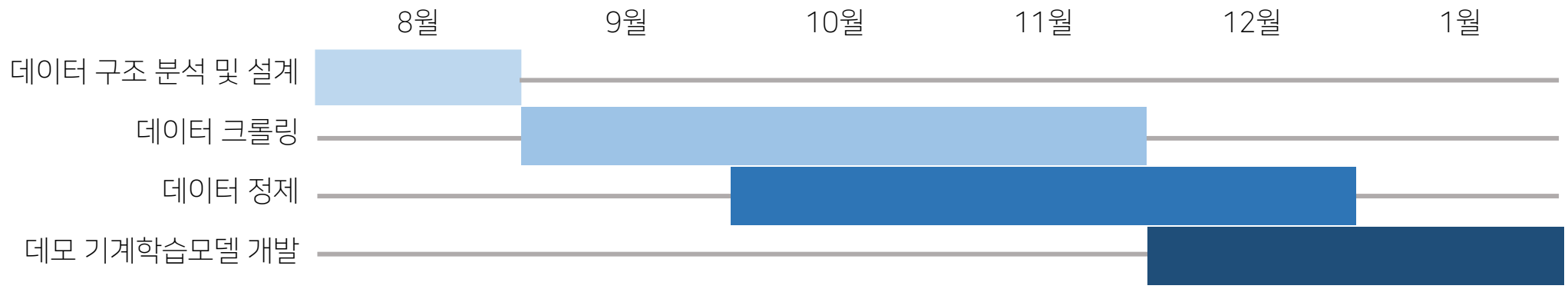
- HW: 클라우드 서버를 사용하여 데이터 분석 작업을 진행
- SW: 개발팀에서 독자 설계하는 머신러닝 모델을 구현

데이터 응용 및 활용 시나리오

- 자동 주석생성 기계학습 모델의 개발
- 소스코드 검색 모델의 개발
- 소스코드 분류 모델의 개발

데이터 수집 전략

추진계획



결과 데이터 및 이관 전략(예상용량, 데이터 포맷)

- 데이터 포맷: JSON형식
- 예상용량: 1,000,000개 프로젝트, 1TB

감사합니다

빅데이터 혁신공유대학 데이터 셋 개발 프로젝트 공모전