프로그램 소스코드 학습을 위한 오픈 소스 데이터 구축

빅데이터 혁신공유대학 데이터 셋 개발 프로젝트 공모전 중간점검

숙명여자대학교

목차

1. 관련연구
2. 문제 인식
3. 데이터 크롤링
4. 데이터 파싱
5. 향후 계획

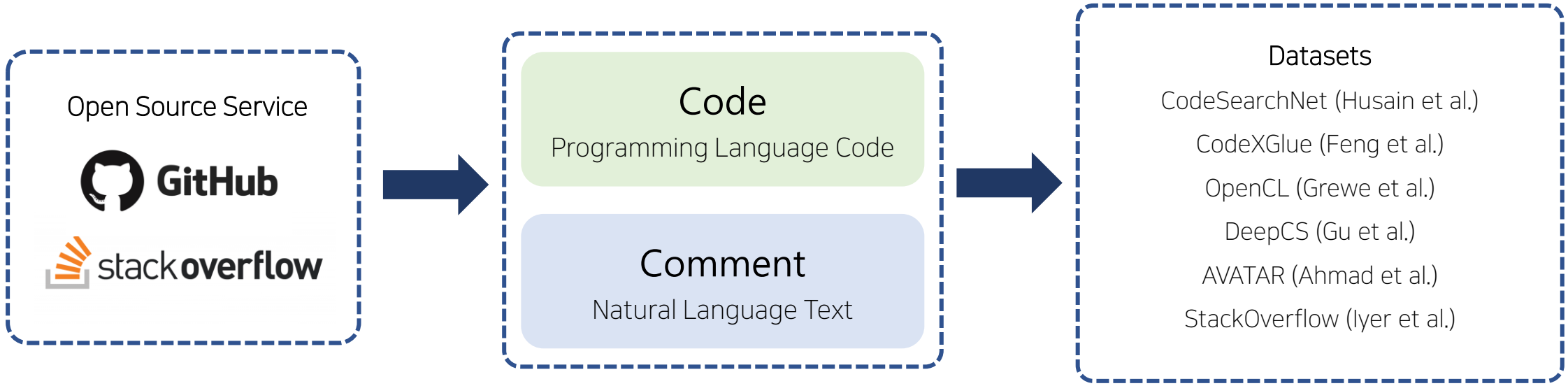
관련 연구

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

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

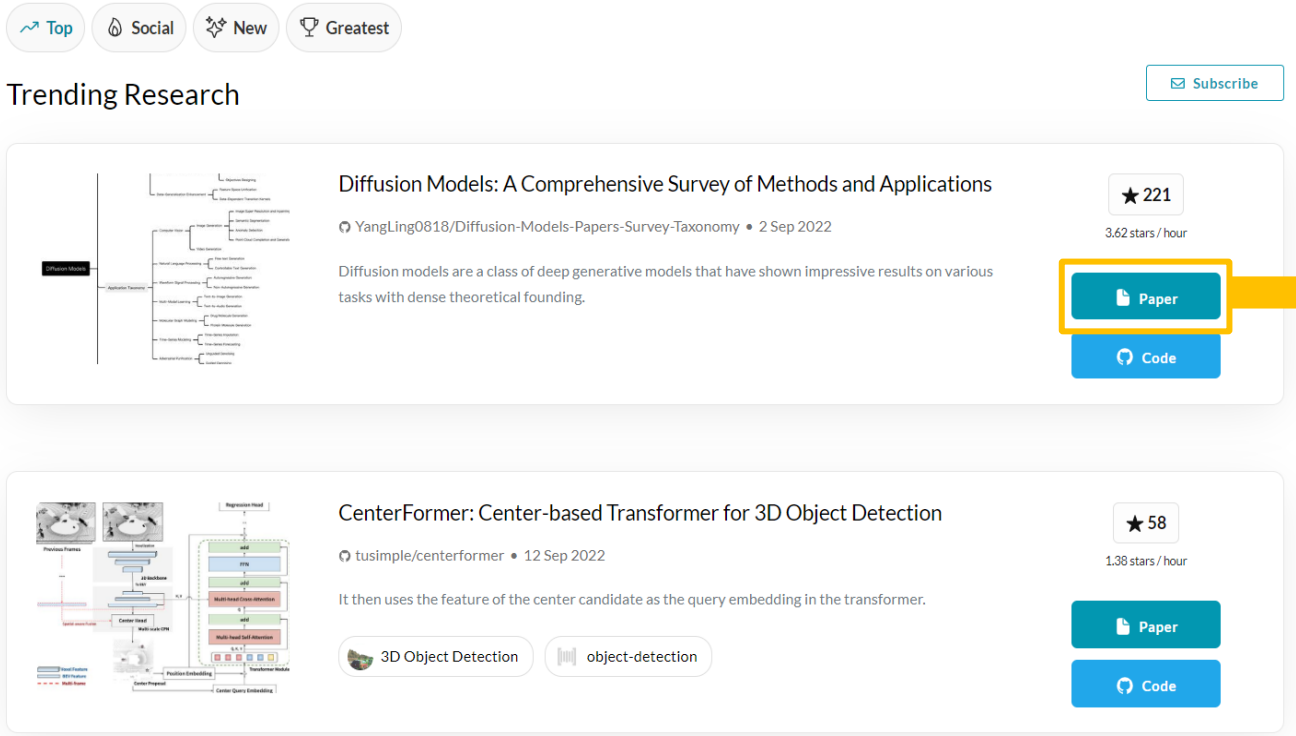
데이터를 활용해 해결할 수 있는 문제

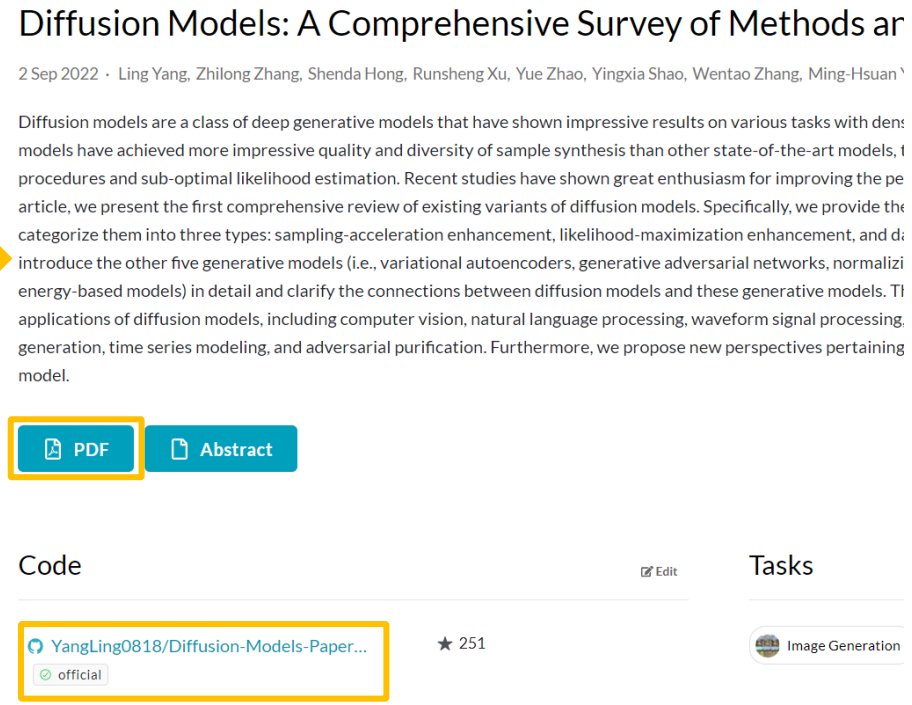
- 프로그래밍에 관련된 모델의 학습 및 개발에 활용 가능
- 자동주석 생성, 소스코드 자동분류, 소스코드 자연어검색 등에 활용 가능
- 더 도전적인 연구로는 소스코드 자동생성 연구를 위한 학습데이터로 활용 가능
- 소스코드 기반 기계학습의 최종 종착점 = 프로그래밍 자동화

데이터 크롤링

오픈 소스 코드 데이터

1. Papers with Code
   - Papers with Code 사이트에서 제공하는 코드와 논문을 함께 크롤링 → 약 700개 수집



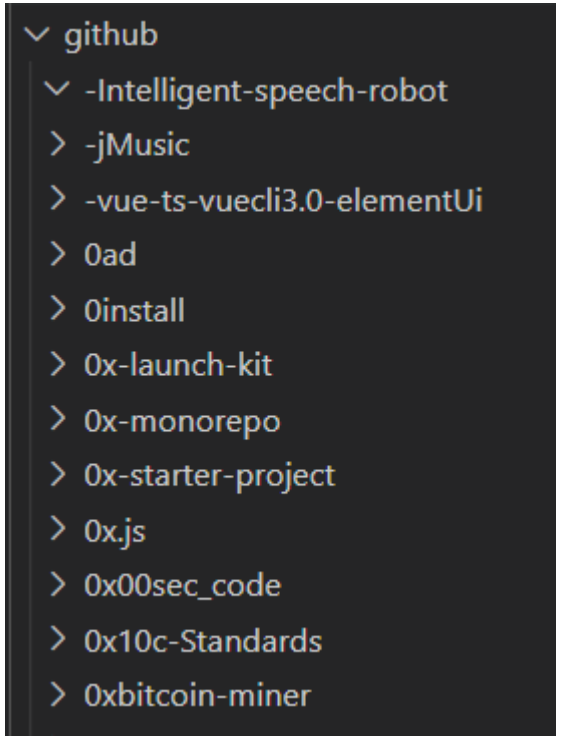


데이터 크롤링

오픈 소스 코드 데이터

2. Github open source
    - 직접 수집한 repository url, CodeSearchNet에서 제공하는 repository url 등 → 약 500,000개 수집

```
0169e931-6889-4682-e296-ee129af55614,0-1-0/lightblue-0.4
0169e931-6ece-b911-f568-024bc5e35aa3,0-14N/NDroid
0169e931-6f3d-5557-da41-828124d25c07,0-duke/wdpassport-utils
0169e931-6f5a-bb2a-9abe-c763cf157988,0-l/dawn
0169e931-6f73-1734-0ed2-8d62ba9ad52f,00-Evan/shattered-pixel-dungeon-gdx
0169e931-6fbd-7d6b-c251-729651db6b88,00-Evan/shattered-pixel-dungeon
0169e931-704f-cb9b-67bd-4944f8dbf8b6,0011001011/vizuka
0169e931-706c-8ca3-59ea-5748fd7ea91d,00111000/Imports-in-Python
0169e931-70a9-f6c5-27a4-944d90b8e08d,001SPARTaN/aggressor_scripts
0169e931-7107-0dd5-b783-f4e91b6ccdc5,003random/003Recon
0169ecd0-9fdc-5793-e113-bbff1b0c4467,007gzs/dingtalk-sdk
0169ecd0-9fe6-ca7b-5514-20795fecbab0,008chen/InterpolatorShow
0169ecd0-9fef-f723-9bf3-2e5102000632,008karan/Face-recognition
0169ecd0-9ff6-34de-64a9-f91f48844943,00StevenG/NSString-Japanese
```



데이터 파싱

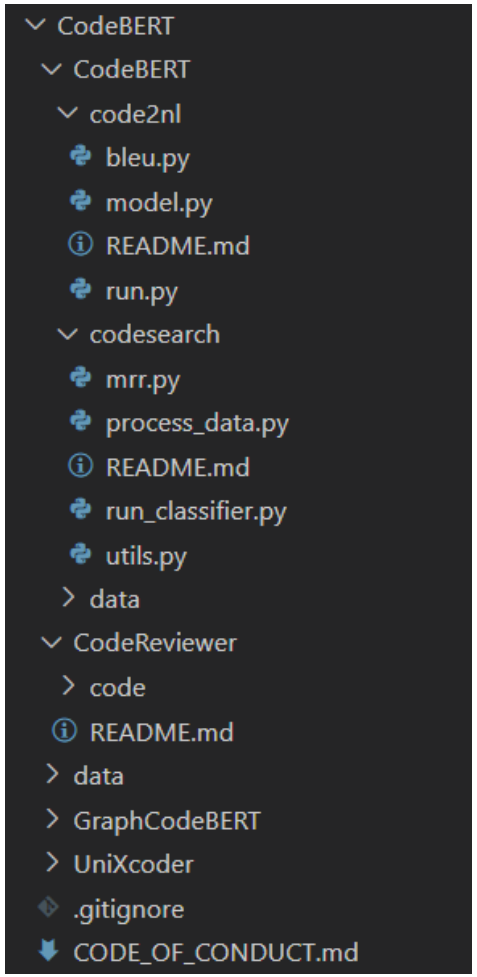
데이터 파싱 규칙

- 소스코드 이해를 위한 데이터로 바로 실행하거나 컴파일 목적이 아니다.
    - 프로젝트 내에서 소스코드 파일과 설명 파일만 포함한다.
    - 압축 파일, 데이터 파일, 오브젝트 파일 등 컴파일을 위한 요소는 포함하지 않는다.
- 프로젝트 하나에 대해서 한 개의 json파일 만든다.
- 데이터셋에 포함되는 소스코드 파일의 프로그래밍 언어는 9개이다.
    - java, python, ruby, c, c++, c#, javascript, html, php
- 데이터셋에 포함되는 설명 파일은 markdown 파일과 논문 url이다.
- 소스코드 파일은 'code ' 로, 설명은 ‘description’으로 구분하여 저장한다.
- 크롤링한 원본데이터는 유지하고 해당 파일의 경로를 포함하여 사용자가 찾기 쉽도록 한다.

데이터 파싱

데이터 파싱 예시

- 프로젝트 소스코드에서 파일만 읽어와서 순서대로 저장하는 형태




```python
import torch
import torch.nn as nn
import torch
from torch.autograd import Variable
import copy
import torch.nn.functional as F
from torch.nn import CrossEntropyLoss, MSELoss

class RobertaClassificationHead(nn.Module):
    """Head for sentence-level classification tasks."""

    def __init__(self, config):
        super().__init__()
        self.dense = nn.Linear(config.hidden_size*2, config.hidden_size)
        self.dropout = nn.Dropout(config.hidden_dropout_prob)
        self.out_proj = nn.Linear(config.hidden_size, 2)

    def forward(self, features, **kwargs):
        x = features[:, 0, :]  # take <s> token (equiv. to [CLS])
        x = x.reshape(-1,x.size(-1)*2)
        x = self.dropout(x)
        x = self.dense(x)
        x = torch.tanh(x)
        x = self.dropout(x)
        x = self.out_proj(x)
        return x

class Model(nn.Module):
    def __init__(self, encoder,config,tokenizer,args):
```

parsing →

Project별 JSON 파일

데이터 파싱

데이터 파싱 예시

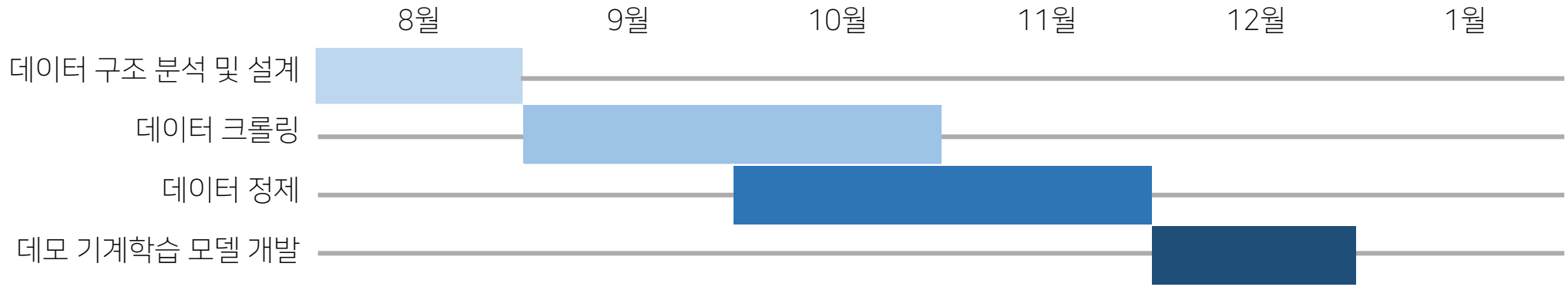
- 프로젝트 소스코드에서 파일만 읽어와서 순서대로 저장하는 형태


```
{'project_name': 'CodeBERT',
 'repo_url': 'https://github.com/microsoft/CodeBERT.git',
 'description': {'language': 'markdown',
  'path': 'CodeBERT/NOTICE.md',
  'content': 'NOTICES AND INFORMATION\n\nDo Not Translate or Localize\n\nThis software ir
 'description': {'language': 'markdown',
  'path': 'CodeBERT/CONTRIBUTING.md',
  'content': '# Contributing\n\nThis project welcomes contributions and suggestions. Most
 'description': {'language': 'markdown',
  'path': 'CodeBERT/CODE_OF_CONDUCT.md',
  'content': '# Microsoft Open Source Code of Conduct\n\nThis project has adopted the [Mi
 'code': {'language': 'python',
  'path': 'CodeBERT/UniXcoder/unixcoder.py',
  'content': '# Copyright (c) Microsoft Corporation. \n# Licensed under the MIT license.\
 'description': {'language': 'markdown',
  'path': 'CodeBERT/UniXcoder/README.md',
  'content': '# UniXcoder\n\nThis repo will provide the code for reproducing the experime
 'code': {'language': 'python',
  'path': 'CodeBERT/UniXcoder/downstream-tasks/code-search/model.py',
  'content': '# Copyright (c) Microsoft Corporation.\n# Licensed under the MIT License.\n
 'code': {'language': 'python',
  'path': 'CodeBERT/UniXcoder/downstream-tasks/code-search/run.py',
  'content': '# coding=utf-8\n# Copyright 2018 The Google AI Language Team Authors and Th
```


저장하는 데이터

- project_name: 프로젝트 명
- repo_url: 프로젝트 url 주소
- language: 소스코드 파일 프로그래밍 언어
- path: 소스코드 파일의 경로
- content:소스코드 파일의 내용

향후 계획



진행 예정 사항

- 데이터 크롤링 및 데이터 정제
- 데모 기계학습 모델 개발
- 구축한 데이터셋으로 모델 성능 테스트(테스트 할 프로그래밍 언어: Java)

최종 데이터

- 오픈소스 프로젝트 1,000,000개, 1TB
    - 오픈소스 프로젝트 데이터 원본
    - 프로젝트 별로 정제한 JSON 파일

감사합니다

빅데이터 혁신공유대학 데이터 셋 개발 프로젝트 공모전

숙명여자대학교