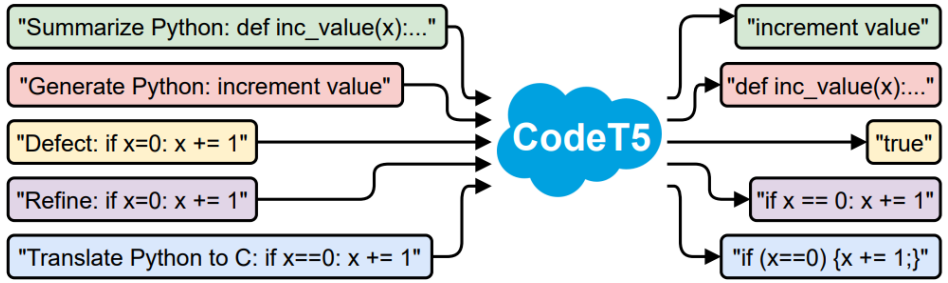
코드 유사성 판단 AI 경진대회

Code NLP



Team ETC
박혜진 장정인 김예지

1. Challenge
2. Data Preprocessing
3. Model
4. Training Strategy
5. Result

코드 유사성 판단 AI 경진대회

두 코드간 유사성(동일 결과물 산출 가능한지) 여부를 판단할 수 있는 AI 알고리즘을 개발
심사기준 : Accuracy

구구단1.py

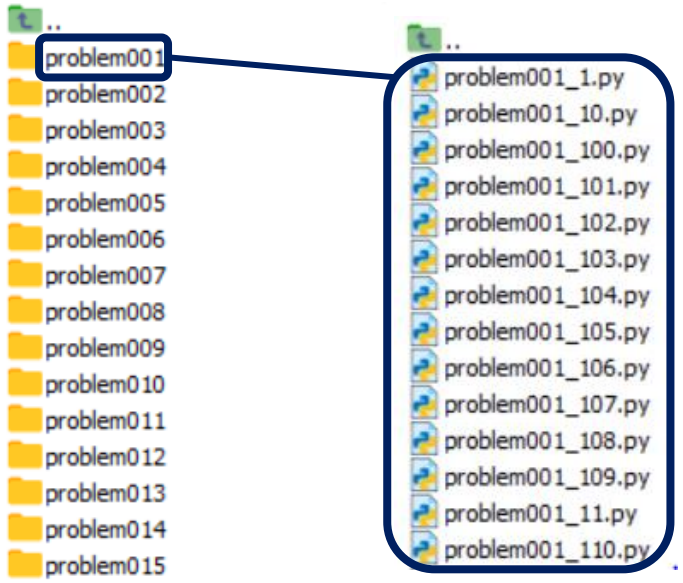
```
M = 9
N = 9

def main():
    for i in range(1,M+1,1):
        for j in range(1,N+1,1):
            mult = i * j
            print(str(i) + "x" + str(j) + "=" + str(i * j))
main()
```

구구단2.py

```
for i in range(0,9):
    for j in range(0,9):
        print("{0}x{1}={2}".format(i+1,j+1,(i+1)*(j+1)))
```

동일 결과물을
산출 하는가?

- Dataset



300개 150개

Pair 구성

	code1	code 2	similar
1	flag = "go" cnt = 0 while flag == "go": ...	# Python 3+ #----------...	1
2	b, c = map(int, input().split()) print(b * c)	import numpy as np n = int(input()) a ...	0
3	import numpy as np import sys read =...	N, M = map(int, input().split()) if M%2 !...	0
4	b, c = map(int, input().split()) print(b * c)	n,m=map(int,input().split()) h=list(map(i...	0
5	s=input() t=input() ans=0 for i in range...	import math a,b,h,m=map(int,input().sp...	0
6	n,m = map(int,input().split()) l=1 if n%2...	N = int(input()) L = list(map(int, input()...	0
7	N = int(input()) P = list(map(int, input()...	input() numbers = input().split() numbe...	0
8	N = int(input()) A = list(map(int, input()...	sum = 1 input() for i in map(int,input()...	1
9	from sys import stdin, setrecursionlimit...	from statistics import median #import ...	1
10	n,k = map(int,input().split()) a = (n-k+2)...	def make_divisors(n): divisors = [] for i ...	0

▪ Pair composition

> Train data: 5,161,457 개
> Validation data: 50,000 개
> Test data: 179,700 개

- Positive pair
  - 2,584,375 개
  - 같은 폴더(같은 기능을 하는)에 있는 코드 조합으로 구성

- Negative pair
  - 2,577,082 개
  - BM25Okapi를 사용하여 같은 폴더에 있는 코드를 제외한 가장 유사성이 높은 코드들로 구성
  - BM25Okapi : 입력 값과 토큰화된 코드 리스트 각각의 유사도를 계산

▪ Preprocessing

원본.py

```python
import bisect
import copy
import heapq
import math
import sys
from collections import *
from functools import lru_cache
from itertools import accumulate, combinations, permutations,
product
def input():
    return sys.stdin.readline()[:-1]
def ruiseki(lst):
    return [0]+list(accumulate(lst))
sys.setrecursionlimit(5000000)
mod=pow(10,9)+7
al=[chr(ord('a') + i) for i in range(26)]
direction=[[1,0],[0,1],[-1,0],[0,-1]]
s=input()
lns=len(s)
lst=[0]*(lns+1)

start=[]
if s[0]=="<":
    start.append(0)
for i in range(lns-1):
    if s[i]==">" and s[i+1]=="<":
        start.append(i+1)
if s[lns-1]==">":
    start.append(lns)

for i in start:
    d=deque([[i,0],[i,1]])
    while d:
        now,lr=d.popleft()
        # print(now)
        if now-1>=0 and lr==0 and s[now-1]==">":
            lst[now-1]=max(lst[now-1],lst[now]+1)
            d.append([now-1,0])
        if now+1<=lns and lr==1 and s[now]=="<":
            lst[now+1]=max(lst[now+1],lst[now]+1)
            d.append([now+1,1])
    # print(lst)
# print(start)
# print(lst)
print(sum(lst))
```

- ✓ # 포함하여 주석 내용 삭제
- ✓ ' '-> tab 변환
- ✓ 다중개행을 한번으로 변환

```python
import bisect
import copy
import heapq
import math
import sys
from collections import *
from functools import lru_cache
from itertools import accumulate, combinations,
permutations, product
def input():
   return sys.stdin.readline()[:-1]
def ruiseki(lst):
   return [0]+list(accumulate(lst))
sys.setrecursionlimit(5000000)
mod=pow(10,9)+7
al=[chr(ord('a') + i) for i in range(26)]
direction=[[1,0],[0,1],[-1,0],[0,-1]]
s=input()
lns=len(s)
lst=[0]*(lns+1)
start=[]
if s[0]=="<":
   start.append(0)
for i in range(lns-1):
   if s[i]==">" and s[i+1]=="<":
      start.append(i+1)
if s[lns-1]==">":
   start.append(lns)
for i in start:
   d=deque([[i,0],[i,1]])
   while d:
      now,lr=d.popleft()
      if now-1>=0 and lr==0 and s[now-1]==">":
         lst[now-1]=max(lst[now-1],lst[now]+1)
         d.append([now-1,0])
      if now+1<=lns and lr==1 and s[now]=="<":
         lst[now+1]=max(lst[now+1],lst[now]+1)
         d.append([now+1,1])
print(sum(lst))
```

Roberta Tokenizer

```
['import', 'Ġbisect', 'Ċ', 'import',
'Ġcopy', 'Ċ', 'import', 'Ġheap', 'q',
'Ċ', 'import', 'Ġmath', 'Ċ',
'import', 'Ġsys', 'Ċ', 'from',
'Ġcollections', 'Ġimport', 'Ġ*', 'Ċ',
'from', 'Ġfunctools', 'Ġimport',
'Ġlru', '_', 'cache', 'Ċ', 'from',
'Ġitertools', 'Ġimport',
'Ġaccumulate', ',', 'Ġcombinations',
',', 'Ġpermutations', ',',
'Ġproduct', 'Ċ', 'def', 'Ġinput',
'():', 'Ċ', 'ĉ', 'return', 'Ġsys',
'.', 'stdin', '.', 'read', 'line',
'()', '[:', '-', '1', ']', 'Ċ',
'def', 'Ġru', 'ise', 'ki', '(',
'lst', '):', 'Ċ', 'ĉ', 'return',
'Ġ[', '0', ']+', 'list', '(', 'acc',
'umulate', '(', 'lst', ''))', 'Ċ',
'sys', '.', 'set', 'recursion',
'limit', '(', '5', '000000', ')',
'Ċ', 'mod', '=', 'pow', '(', '10',
',', '9', ')+', '7', 'Ċ', 'al', '=[',
'chr', '(', 'ord', "('", 'a', "')",
'Ġ+', 'Ġi', ')', 'Ġfor', 'Ġi', 'Ġin',
'Ġrange', '(', '26', ')]', 'Ċ',
'direction', '=[', '[', '1', ',',
'0', '],', '[', '0', ',', '1', '],',
'[-', '1', ',', '0', '],', '[', '0',
',', '-', '1', ']]', 'Ċ', 's', '=',
'input', ...
```

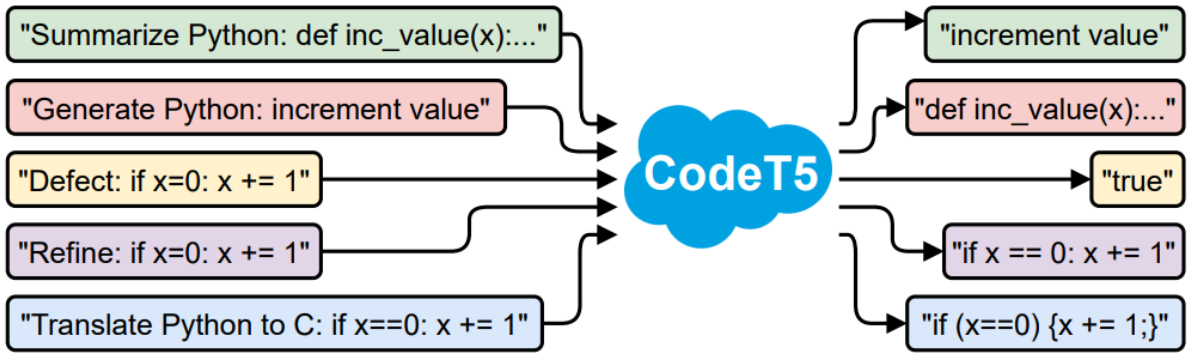
CodeT5

- T5 (Transfer Learning with a Unified Text-to-Text Transformer)
  - 모든 NLP task를 text-to-text 하나로 통합
  - Encoder-decoder(seq2seq) architecture
  - text-to-text 변경만으로도 성능이 좋으며 강건함

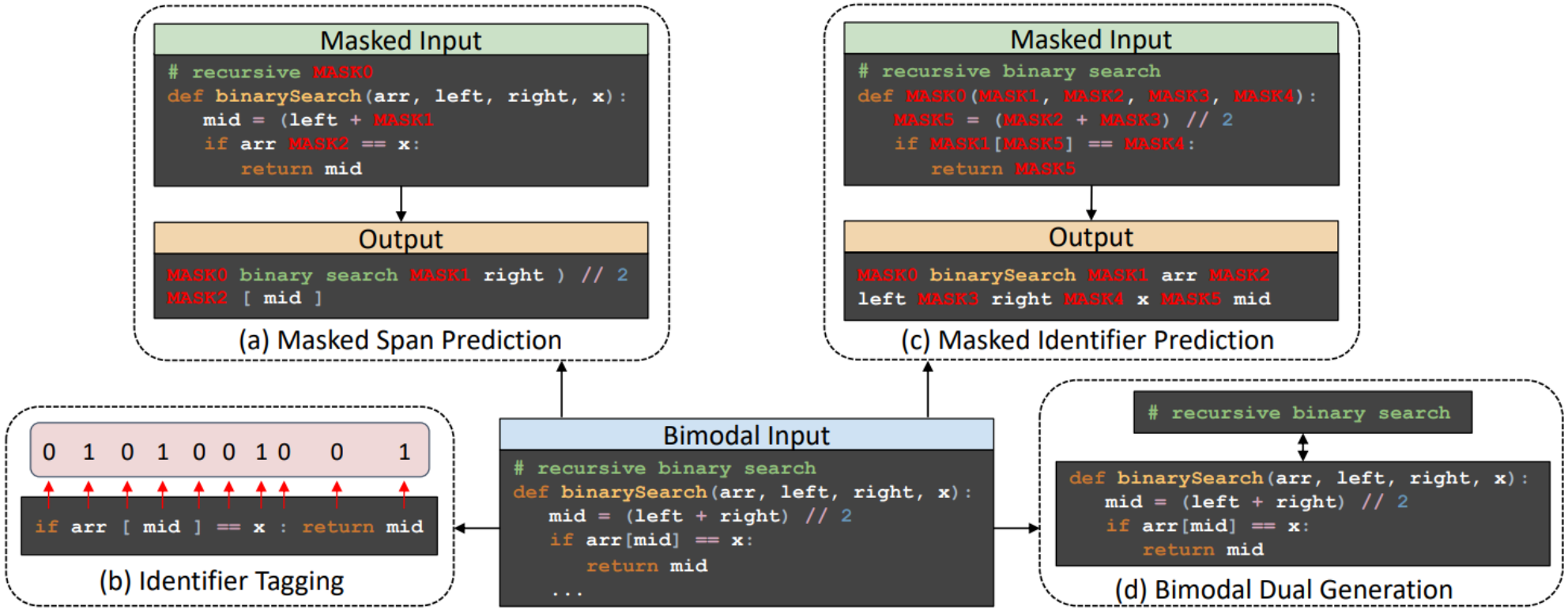
- CodeXGLUE benchmark SOTA 모델

Benchmarks

Trend	Task	Dataset Variant	Best Model
	Code Summarization	CodeXGLUE - CodeSearchNet	CodeT5
	Defect Detection	CodeXGLUE - Devign	CodeT5
	Code Translation	CodeXGLUE - CodeTrans	CodeT5
	Text-to-Code Generation	CodeXGLUE - CONCODE	CodeT5
	Clone Detection	CodeXGLUE - BigCloneBench	CodeT5

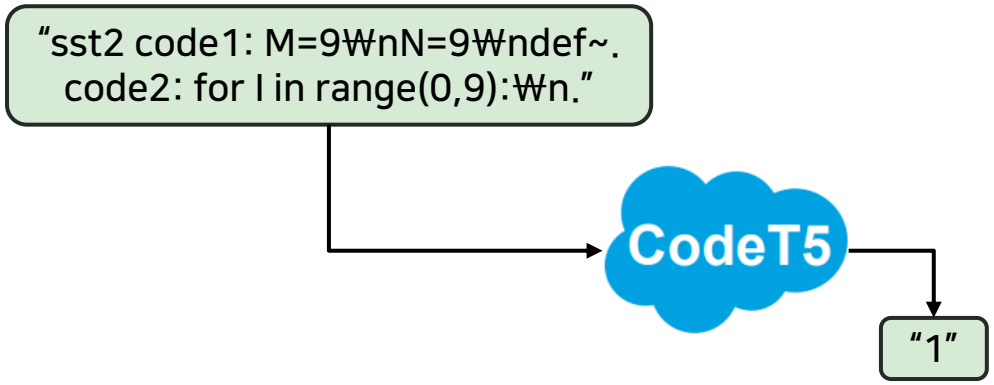


- CodeT5



Pre-training tasks of CodeT5

- T5ForConditionalGeneration 😂
- Pretrained: CodeT5-base
- GLUE sst2: binary classification



SST2

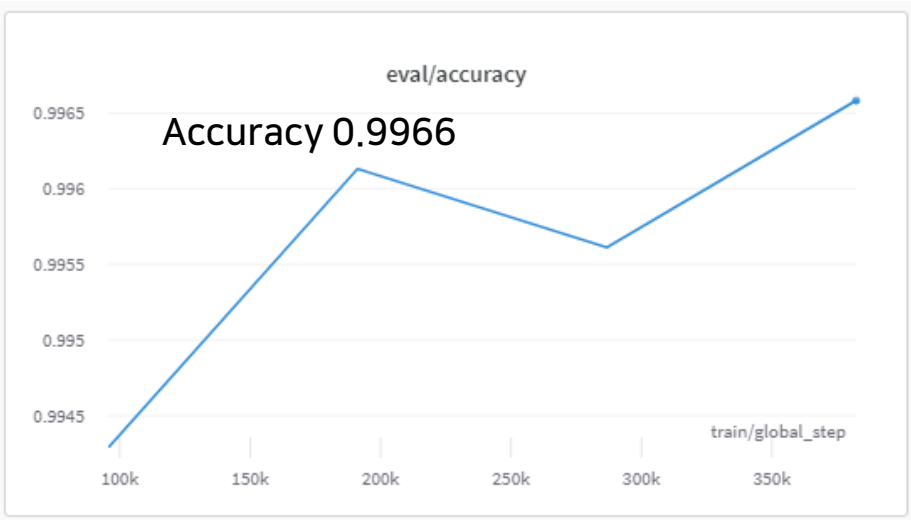
Original input:
code1: M = 9\nN = 9\ndef main():\n\tfor i in range(1,M+1,1):\n\t\tfor j in range(1,N+1,1):\n\t\t\tmult = i * j\n\t\t\tprint(str(i) + "x" + str(j) + "=" + str(i * j))\nmain(). for i in range(0,9):\n\tfor j in range(0,9):\n\t\tprint("{0}x{1}={2}".format(i+1,j+1,(i+1)*(j+1)))

Processed input: sst2 code1: M = 9\nN = 9\ndef main():\n\tfor i in range(1,M+1,1):\n\t\tfor j in range(1,N+1,1):\n\t\t\tmult = i * j\n\t\t\tprint(str(i) + "x" + str(j) + "=" + str(i * j))\nmain(). code 2: for i in range(0,9):\n\tfor j in range(0,9):\n\t\tprint("{0}x{1}={2}".format(i+1,j+1,(i+1)*(j+1)))

Original target: 1

Processed target: "1"

- Result



pair_id	code1	code2	similar
140001	N = int(input()) b = [int(x) for x in input().split()] val = b[0] val += b[-1] for i in range(N-2): val += min(b[i], b[i+1]) print(val)	import sys read = sys.stdin.buffer.read readline = sys.stdin.buffer.readline readlines = sys.stdin.buffer.readlines n, *b = map(int, read().split()) a = [b[0]] for i in range(1, n-1): a.append(min(b[i - 1], b[i])) a.append(b[-1]) print(sum(a))	1

- Leaderboard #9
  - 심사 기준: accuracy
  - 1차 평가(Public Score): 테스트 데이터 중 랜덤 샘플 된 30%로 채점, 대회 기간 중 공개 #8
  - 2차 평가(Private Score): 테스트 데이터 중 나머지 70%로 채점, 대회 종료 직후 공개 #9

1	beretta92x	0.9909
2	vecxoz	0.98379
3	상하목장스누피	0.98061
4	포스빌런	0.97946
5	내가누구게	0.97848
6	하르딘	0.97735
7	ms_kim	0.97707
8	dkseho	0.97612
9	ETC	0.97532

Thank you for your attention....!!
QnA