Code plagiarism

Code clones

- 1) Type 1: Identical code ใส่เข็บช่องว่าง comment หรือ layout
- 2) Type 2: Identical ยกเว้นชื่อเรือแปร, กรประกศตัวแปร
- 3) Tupe 3: คล้ายๆ มีการแก้ไข ลบหรือเพิ่มปรรกัด
- 4) Type 4: หลักการหรือผลลัพช์ เหมือนกัน แต่มี Syntax ต่างกัน

Machine learning: Decision tree, Random forest, SVM, SVM+SMO

Features: 11 Syntactic Code metrics +

12 Semantic Code Metrics +

(Code 2 Vec)

Merry engine (Web-based)

Building Merry Engine

1) Data Collection and preparation

BigCloneBench โดยแบ่วเป็น 1) JAVA Source Code

2) Clone Ground-truth

Train-Test split

2) Code metrics extraction

Syntactic metrics – จำนวนข้างแปร จำนวนบรรทัด (ดูโค๊ด)

Semantic metrics - ใช้ code 2 vec -> แบ่ว Vector และหาค่าความเหมือนแบบ cosin (ลูกรทำงาน)

3) Machine learning models

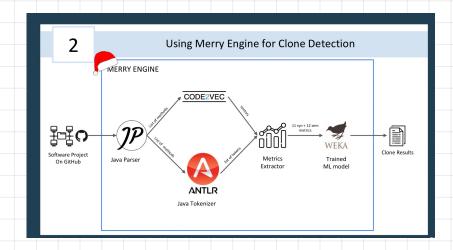
Decision Tree

(Library)

Random Forest

SVM / Using SMO Opt-

How to use ML



Evaluation

 > ใหเรียความแม่นย้า
 True Positive

 precision
 =

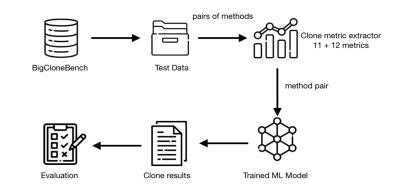
 TruePositive + FalsePositive

Recall = True Positive

True Positive + False Negative

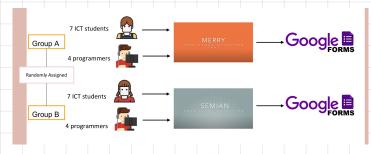
F1-score = 2. Precision - Recall
Precision + Recall

Work Process



Real Project		
	roc	Precision \
JUnit4	7550	0.41
Natty	1761	0.77
Merry	931	0.75

>ในเช็วเครื่องมือ



ทดสอบว่า จะใช้เครื่องมือใดมากกว่ากัน โดยการทำแบบสอบผาม

Limitations

ใช้ได้เฉพาะ JAVA , code 2 vec ใช้เลลาทำอานหาน