Code Clone Detector

ระบบตรวจวัดความเหมือนของโค้ดและตรวจจับการคัดลอกโค้ด

โค้ดที่เหมือนกันคืออะไร?

มีโค้ดสองอันถ้าชุดความเหมือนเพียงพอ

- โค้ดที่ชุดความเหมือนกันชนิดต่างๆ
  1. มีความเหมือนกันแบบ 100%
  2. ชุดความเหมือนกันเกือบทั้งหมดแต่ต่างกัน คือ ตัวแปร, ค่าคงที่ต่างๆ
  3. มีบางบรรทัดที่มีการเพิ่มหรือลดโค้ดที่มีความเหมือนกันแบบคร่าวๆ ล้อมรอบ
  4. หน้าที่การทำงานเหมือนหรือคล้ายกันเหมือนกัน

วิธีการค้นหาโค้ดที่เหมือนกัน

นำโค้ดมาทำให้อยู่ในรูปแบบเดียวกัน → การเปลี่ยนโค้ดให้อยู่ในรูปแบบอื่นๆ → นำโค้ดมาเปรียบเทียบว่าเหมือนกันหรือ

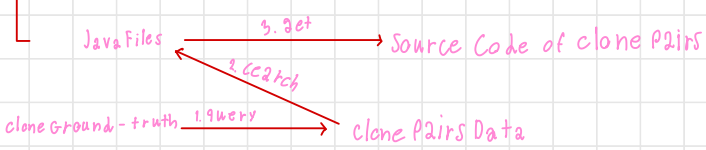
↓

ส่งให้ผู้ใช้ ← คัดข้อมูลที่ไม่จำเป็นออก ← จัดเรียงลำดับให้เป็นระเบียบ

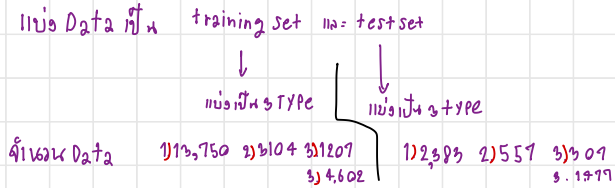
วิธีการสร้าง

1. เก็บข้อมูล
   - ใช้ฐานข้อมูลจาก BigCloneBench

นำข้อมูลออกมาอย่างไร?



แบ่ง Data เป็น training set และ test set



2. Metrics

- Syntactic metrics
  - กฎและวิธีการเขียนของโค้ด
    - code 1 → ก่อต่างของโค้ด
    - code 2 → ก่อต่างของโค้ด
    - เพื่อมา-ลบ และดูความแตกต่าง
    - token, Uni, identifier, operator...

No.	Metric
1	Token No [9]
2	Unique Token No [10]
3	Identifier No [10]
4	Unique Identifier No [10]
5	Operator No [10]
6	Unique Operator No [10]
7	Token Types Diversity [9]
8	Diff File Name Score
9	Diff Method Name Score
10	Similar Return Type
11	DiffLOC

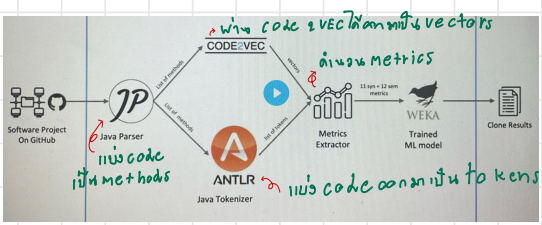
- Semantic metrics
  - ดูวิธีการทำงาน

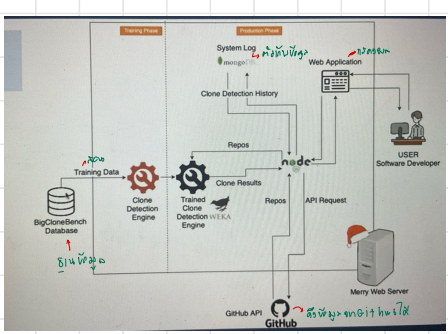
code 1 → code 2 vec → vector → นำมาแบ่งเป็นส่วน
code 2 → code 2 vec → vector → นำมาแบ่งเป็นส่วนๆ
→ ดูความแตกต่างของเวิกเตอร์

3. ใช้ machine Learning models ในการวัดความเหมือนหรือต่างของ code 4 ตัว

1. Decision Tree
2. Random Forest
3. Support Vector Machine
4. SVM using SMO

วิธีการทำงานของ ML ที่สร้างมาตรวจจับ code





การวัดประสิทธิภาพ เชิงความแม่นยำ

ตัวชี้วัด ของ

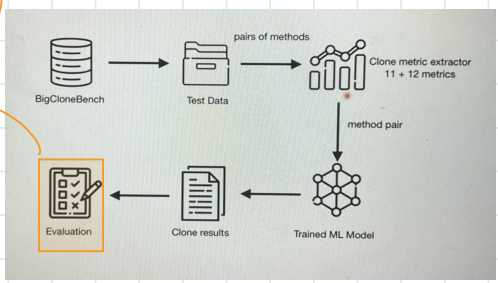
1. Precision = $\frac{\text{True Positive}}{\text{True Positive + False Positive}}$

2. Recall = $\frac{\text{True Positive}}{\text{True Positive + False Negative}}$

3. F1-Score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{precision + Recall}}$ ค่าเฉลี่ยของ ทั้ง

วิธีการที่ใช้งาน

วิธีการที่ใช้ในการ



ผลความแม่นยำ

Model	Metrics	Precision	Recall	F1-Score
Randomization (baseline)		0.20	0.49	0.28
Decision Tree	Syntactic + Semantic	0.89	0.86	0.87
	Syntactic	0.95	0.72	0.86
	Semantic	0.68	0.87	0.76
Random Forest	Syntactic + Semantic	0.97	0.86	0.91
	Syntactic	0.97	0.80	0.87
	Semantic	0.70	0.87	0.78
SVM	Syntactic + Semantic	0.97	0.85	0.91
	Syntactic	0.97	0.79	0.87
	Semantic	0.62	0.90	0.73
SVM using SMO	Syntactic + Semantic	0.98	0.89	0.93
	Syntactic	0.97	0.69	0.81
	Semantic	0.63	0.90	0.74

→ ความแม่นยำต่ำ

→ ความแม่นยำสูง