

Podobnosti v zdrojových kódech

Marek Roštár
Vedúci: Michal Kompan

Motivácia

- S nárastom počtu programovacích jazykov sa stretávame s problémom ako nájsť podobné úseky v kódach z rôznych jazykov.
- Podobné úseky vieme použiť pri odhaľovaní plagiátorov a zlepšovaní zdrojových kódov.
- Navrhnuť metódu odhaľovania podobnosti v zdrojových kódach s dôrazom na rôznorodosť programovacích jazykov a výpočtovú efektívnosť

Hypotéza

- Abstrahovaním zdrojového kódu a jeho následným porovnávaním cez abstraktné syntaktické stromy vieme zrýchliť a zefektívniť proces porovnávania zdrojových kódov, bez veľkej stráty presnosti porovnávania

Metóda

- Predspracovanie pomocou tokenizátora – 2 úrovne abstrackie:
 - Prvá len pripraví súbor na parsovanie ale neupravuje jeho obsah
 - Druhá odtraňuje mená a hodnoty premenných a nahradzuje ich tokenmi
- Parsovanie spracovaného zdrojového kódu do abstraktného syntaktického stromu
- Rozšírenie uzlov stromu o potrebné informácie
- Porovnávanie uzlov podľa ich hash hodnoty

Overenie

- Porovnanie výsledkov s nástrojom MOSS
- Najprv na mnou vyrobených jednoduchých zdrojových kódach
- 1 dataset na podobnosť z Internetu a dáta z niektorých predmetov na škole

Aktuálny stav a čo budem dopĺňať

- Aktuálne len C a len 2 úrovne abstrakcie
- Podpora viacerých jazykov
- Pridanie aspoň 1 ďalšej úrovne abstrakcie

Problémy

- Problémy s parsermi
- Nekonzistentosť ruby funckie hash
- Online prekladače zdrojových kódov?

```
TranslationUnit
  entities:
    - FunctionDef
      type: Function
      type: Void
      name: "main"
      def: Block
      stmts:
        - ExpressionStatement
          expr: Call
            expr: Variable
              name: "printf"
            args:
              - StringLiteral
                val: "Hello World\\n"
```

```
s(:send, nil, :void,
  s(:block,
    s(:send, nil, :main),
    s(:args),
    s(:send, nil, :printf,
      s(:str, "Hello World\\n")))))
```

Zhrnutie

- Metóda založená na abstraktných syntaktických stromoch a abstrakcii zdrojového kódu
- Problém predstavujú výstupy parserov
- Možná práca do budúcnosti vytvoriť parser pre AST ktorý by vedel spracovávať všetky jazyky