

MASARYKOVA UNIVERZITA
FAKULTA INFORMATIKY



Nástroje na analýzu kódu v jazyku Python a vizualizácia ich výstupu

BAKALÁRSKA PRÁCA

Ján Vorčák

Brno, 2012

Prehlásenie

Prehlasujem, že táto bakalárska práca je mojím pôvodným autorským dielom, ktoré som vypracoval samostatne. Všetky zdroje, pramene a literatúru, ktoré som pri vypracovaní používal alebo z nich čerpal, v práci riadne citujem s uvedením úplného odkazu na príslušný zdroj.

Ján Vorčák

Vedúci práce: Mgr. Marek Grác

Pod'akovanie

Zhrnutie

Klíčové slová

Obsah

1	Úvod	2
2	O jazyku Python	3
2.1	Čo je to Python?	3
2.2	Introspekcia v jazyku Python	3
	Funkcia dir() 4	
	Atribút __doc__ 4	
	Atribút __name__ 4	
	Funkcia type() 5	
	Funkcia id() 5	
	Funkcie hasattr() a getattr() 5	
	Funkcia callable() 6	
	Funkcie isinstance() a isinstance() 6	
3	Nástroje na analýzu kódu pre jazyk Python	7
3.1	Nástroje analýzu kódu Python projektu	7
3.1.1	PEP8	7
3.1.2	PyChecker	7
3.1.3	Pylint	8
3.1.4	Zhrnutie	8
3.2	Nástroje na vizualizáciu Python projektu	9
3.2.1	Pyreverse	9
3.2.2	Pylint-gui	9
3.2.3	Implementácia programu Pylint v IDE	10
3.2.4	Zhrnutie	10
4	Analýza nástrojov potrebných k implementácii	11
4.1	Analýza nástroja Gaphas	11
4.1.1	Základná charakteristika nástroja Gaphas	11
4.1.2	Popis Gaphas API	11
4.2	Zhrnutie	12
5	Projekt gpylint	13
5.1	Aplikácia jednotlivých nástrojov	13
5.2	Vyhodnotenie	13
6	Záver	14

1 Úvod

Pri vývoji aplikácií musíme klásť dôraz nielen na samotné programovanie, ale najmä na analýzu a návrh systému, dôkladné otestovanie ale aj spätnú analýzu samotného kódu. Z tohto dôvodu vzniká pre takmer každý programovací či značkovací jazyk množstvo nástrojov, ktoré nám pomáhajú automaticky objaviť potencionálne chyby či porušenie konvencií v zdrojových kódach. Často krát sú tieto nástroje síce efektívne a ľahko konfigurovateľné, no najmä výstupom nie príliš užívateľsky prívetivé. Asi najpoužívanejším nástrojom na analýzu kódu v jazyku Python je konzolová utilita Pylint. Pokiaľ si chceme vďaka tomuto programu vytvoriť obraz o väčšom projekte nestačí nám iba textový výstup, no musíme tieto dáta automaticky spracovať.

Cieľom tejto bakalárskej práce je analyzovať nástroje na analýzu a vizualizáciu Python kódu ako aj vlastnosti jazyka ako je napríklad introspekcia, ktoré nám túto analýzu umožňujú. Z dôvodu vizualizácie je nutné taktiež nastudovať základy modelovania hierarchických štruktúr v UML. Výstupom bakalárskej práce je nástroj, ktorý poskytuje funkcionality, ktorú v existujúcich nástrojoch nenájdeme. Nástroj teda zjednoduší orientáciu najmä vo väčšom projekte pomocou vizualizácie dát najmä za pomoci vygenerovania interaktívnych UML diagramov. Program bude taktiež dopĺňať funkcionality, ktorú nám klasický program Pylint neumožňuje ako je napríklad filtrácia falošných poplachov.

Práca pozostáva z piatich kapitol. V druhej kapitole si predstavíme jazyk Python a jeho vlastnosti, neskôr rozanalyzujeme existujúce nástroje na analýzu kódu v jazyku Python a ich nedostatky. Pred implementačnou časťou si predstavíme knižnice, ktoré nám neskôr pomôžu pri implementácii nástroja ako je napríklad knižnica Gaphas na vykresľovanie grafiky v GTK programoch alebo samotný Pylint. Nasledovať bude kapitola o návrhu nástroja a samotná implementácia. Nakoniec zhodnotíme, či implementovaný nástroj spĺňa určené požiadavky či už po stránke funkcionality alebo škálovateľnosti a navrhujeme zmeny na jeho zlepšenie.

2 O jazyku Python

2.1 Čo je to Python?

Python je vysoko úrovňový, objektovo orientovaný dynamický jazyk, ktorý vytvoril holandský programátor Guido van Rossum ako následníka jazyka ABC. Vyznačuje sa prehľadnou syntaxou, jednoduchosťou a modulárnosťou. Python podporuje viacero programátorských paradigmat, najmä objektovo orientované, imperatívne a čiastočne aj funkcionálne paradigma.

Existuje viacero implementácií jazyka Python. Medzi najznámejšie patria:

- CPython
- Jython
- Python pre .NET
- IronPython
- PyPy

Najpožívanejšou a najpodporovanejšou je ale CPython. Každá z implementácií sa môže líšiť špecifickými informáciami mimo štandardnej Python dokumentácie. Zdrojové kódy sú preložené do byte kódu a zvyčajne uložené v .pyc alebo .pyo súboroch, ktoré sú neskôr spúšťané virtuálnou

V štandardnej knižnici je dostupné množstvo datových typov ako napríklad reálne a komplexné čísla, celé čísla s neobmedzenou dĺžkou, znakové reťazce, zoznamy a slovníky. Datové typy sú silno a dynamicky typované. Použitie nekompatibilného typu spôsobí vyvolanie výnimky. Python podporuje objektovo orientované programovanie vrátane viacnásobnej dedičnosti. Kód je sústredený do modulov a balíkov s možnosťou importovať špecifický modul, triedu, funkciu alebo iný objekt. Za účelom ošetrovania chýb Python podporuje vyvolávanie a odchyťovanie výnimiek. Automatická správa pamäti nahrádza nutnosť manuálne alokovať a uvoľňovať pamäť v kóde.

2.2 Introspekcia v jazyku Python

Introspekcia je schopnosť preskúmať daný objekt a rozhodnúť o jeho identite, vlastnostiach a schopnostiach. Jazyk Python podporuje rozsiahlu introspekciu objektov. Medzi hlavné informácie, ktoré potrebujeme o objektoch v jazyku Python zistiť patrí ich meno, typ, identita, vlastnosti, schopnosti a ich

pôvod. Jazyk Python ponúka množstvo nástrojov, ktoré nám tieto vlastnosti umožňujú zistiť. Môžeme ich rozdeliť do dvoch hlavných skupín. V prvej skupine sú funkcie či už zo štandardnej knižnice alebo z pomocných modulov akým je napríklad modul `inspect`, do druhej skupiny radíme atribúty objektov, ktoré priamo v objekte uchovávajú užitočné informácie.

Funkcia `dir()` Funkcia `dir()` je jedným z hlavných nástrojov introspekcie v jazyku Python a vracia zotriedený zoznam mien atribútov objektu, ktorý bol uvedený ako jej argument. Funkcia je súčasťou štandardnej knižnice, takže nemusíme importovať žiaden modul pre jej použitie. Na výpis funkcií zo štandardnej knižnice môžeme teda využiť samotnú funkciu `dir()`.

```
In [1]: print dir(__builtin__)[-10:]
['str', 'sum', 'super', 'tuple', 'type', 'unichr', '
  unicode', 'vars', 'xrange', 'zip']
```

V prípade, že je funkcia `dir()` použitá bez argumentov vracia zoznam mien, ktoré sú momentálne definované.

```
In [2]: print dir()
['In', 'Out', '_', '__', '___', '__builtin__', '
  __builtins__', '__name__', '_dh', '_i', '_il', '_i2
  ', '_ih', '_ii', '_iii', '_oh', '_sh', 'exit', '
  get_ipython', 'help', 'quit']
```

Atribút `__doc__` Atribút `__doc__` obsahuje komentáre, ktoré popisujú objekt. V prípade, že prvý v module, triede alebo v metóde je znakový reťazec, je automaticky považovaný za `__doc__` atribút. V opačnom prípade sa jeho hodnota nastaví na `None`. Hodnoty `__doc__` sa z dôvodu kompaktnosti nevkladajú do byte kódu.

```
In [3]: print __doc__.__doc__
str(object) -> string
```

```
Return a nice string representation of the object.
If the argument is a string, the return value is the
same object.
```

Atribút `__name__` Atribút `__name__` obsahuje názov objektu odvodený z jeho typu. Niektoré objekty, ako napríklad objekty typu `string` tento atribút

neobsahujú. Tento atribút obsahujú napríklad moduly. V prípade, že spúšťame script priamo pomocou Python interpretu je atribút `__name__` nastavený na `'__main__'` nakoľko Python interpret je považovaný za hlavný modul. Tak isto v prípade, že Python skript spúšťame z príkazového riadku. Je teda často používaný na rozpoznanie či daný modul len importujeme alebo priamo spúšťame.

```
In [4]: def my_function():
      ...:     pass
      ...:
```

```
In [5]: my_function.__name__
Out[5]: 'my_function'
```

```
In [6]: __name__
Out[6]: '__main__'
```

Funkcia type() Funkcia `type()` zo štandardnej knižnice vracia typ jej argumentu. Ten vracia v podobe typového objektu, ktorý môže byť porovnávaný s typmi definovanými v module `types`.

```
In [7]: type(my_function)
Out[7]: function
```

```
In [8]: type(1)
Out[8]: int
```

Funkcia id() Funkcia `id()` vracia unikátnu identitu objektu. Táto funkcia je užitočná nakoľko viacero premenných môže odkazovať na rovnaký objekt. Funkcia `id()` konkrétne vracia pamäťovú adresu objektu.

```
In [9]: id('string')
Out[9]: 3078023712L
```

```
In [10]: id(id)
Out[10]: 3078187660L
```

Funkcie hasattr() a getattr() V prípade, že je potrebné zistiť prítomnosť alebo hodnotu atribútu štandardná knižnica ponúka funkcie `hasattr()` a `getattr()`.

```
In [11]: hasattr(id, '__name__')
Out[11]: True
```

```
In [12]: getattr(id, '__name__')
Out[12]: 'id'
```

Funkcia callable() V niektorých prípadoch môžu objekty slúžiť na vyvolanie určitého druhu udalosti. Pomocou funkcie callable() sa dá overiť, či je daný objekt spustiteľný.

```
In [13]: callable(id)
Out[13]: True
```

```
In [14]: callable(1)
Out[14]: False
```

Funkcie isinstance() a subclass() Funkcia isinstance() vracia hodnotu v závislosti na tom, či je objekt inštanciou danej triedy. Funkcia vracia True aj v prípade, že je objekt inštanciou jej predka.

Funkcia subclass() vracia True v prípade, že objekt reprezentujúci triedu je podtriedou druhého argumentu.

```
class Person(object):
    pass
class Student(Person):
    pass
p=Person()
s=Student()
```

```
In [15]: isinstance(s, Student)
Out[15]: True
```

```
In [16]: isinstance(s, Person)
Out[16]: True
```

```
In [17]: subclass(Student, Person)
Out[17]: True
```

3 Nástroje na analýzu kódu pre jazyk Python

3.1 Nástroje analýzy kódu Python projektu

3.1.1 PEP8

Ide o nástroj, ktorý vyvíja Johann C. Rocholl. Tento nástroj slúži na otestovanie kódu Python podľa konvencií definovaných v PEP 8 [?]. Je to konzolový program, ktorý pomáha zvyšovať prehľadnosť a lepšiu štruktúru kódu.

Program PEP8 umožňuje programátorovi pomocou rôznych prepínačov ovládať analyzátor. V predvolenom móde vypíše program každú chybu len raz, čo je pre analýzu rozsiahlejšieho projektu nevhodné. Voľba `-repeat` vypíše každú chybu bez ohľadu na to, či sa už v kóde vyskytuje viackrát alebo nie. Voľba `-filename=patterns` obmedzuje spracovávané súbory len na tie, ktoré vyhovujú zadanému regulárnemu výrazu. Zaujímavým prepínačom je `-show-pep8`, ktorý k danej chybe vypíše na štandardný výstup aj text z definície PEP 8, ktorý daná časť kódu porušuje. Utilita `Pep8.py` je nástroj, ktorý výrazne dopomáha k prehľadnosti a ucelenosti kódu. Chýbajú mu však možnosti detekovania chýb v kóde [?].

3.1.2 PyChecker

PyChecker je program zachytávajúci problémy, na ktoré upozorňuje u nedynamických jazykov, ako je napríklad C kompilátor v podobe varovaní a chýb. Medzi problémy, ktoré tento program zachytí, patrí napríklad zlý počet parametrov predaný funkcii, metóde či konštruktoru, používanie neexistujúcich metód a tried, ako aj použitie premennej pred jej inicializáciou. Detekuje taktiež nepoužité inštancie globálnych či lokálnych premenných, kontroluje definíciu `self` ako prvej premennej u metód tried, ako aj úroveň dokumentácie pre triedy, moduly a metódy [?].

Jednou z predností je aj priamy import do zdrojového kódu. Pokiaľ má užívateľ prístup a práva k modifikácii zdrojového kódu, je veľmi jednoduché PyChecker nainštalovať priamo. Jeho hlavnou nevýhodou je, že daný kód spustí a vykoná. PyChecker je teda nevhodný na detekciu chýb v aplikáciách, ktoré napríklad pracujú nad databázou, alebo pracujú v ostrom prostredí, ktoré je nevhodné na testovanie.

3.1.3 Pylint

Pylint je jeden z najpoužívanějších, ľahko konfigurovateľných nástrojov na analýzu kódu Python, využívaný či už manuálne alebo automaticky. Je vyvíjaný za podpory Logilab.org a vydaný pod licenciou GNU GPL [?]. V zdrojových kódach analyzuje potencionálne chyby a varovania, ale aj porušenie konvencií podľa štandardu PEP 8. S projektom Pylint je dodávaný aj program Pyreverse, ktorý pre daný kód vygeneruje UML diagram v podobe *dot* [?] formátu. Daný *dot* súbor dokáže priamo pretransformovať do PNG alebo iného formátu. Výhodou programu Pylint oproti PyChecker je, že daný súbor nespúšťa, ale analyzuje staticky. Je teda vhodný do každého prostredia. Pylint dopĺňa PyChecker hlavne v kontrole dĺžky riadkov a kontrole názvov premenných za pomoci regulárnych výrazov. Overuje taktiež implementáciu deklarovaných rozhraní a detekuje duplicitný kód. Obzvlášť zaujímavou vlastnosťou je generovanie metrík a externých závislostí [?].

Správanie sa programu je možné upraviť pomocou prepínačov špecifikovaných na príkazovom riadku alebo pomocou konfiguračného súboru. V konfiguračnom súbore je možná filtrácia varovaní a chýb rôznych druhov, ignorovanie rôznych typov súborov. Taktiež je možné načítať rôzne typy doplnkov, nastavenie minima riadkov kontrolovaných pri zisťovaní duplicity, obmedzenie veľkosti modulu, nastavenie znakov využitých na odsadenie, definície vlastných zastaralých modulov.

Najmä u väčších projektov s požiadavkami na prehľadnosť kódu a dobrý návrh užívateľ ocení možnosť nakonfigurovať maximálny počet atribútov triedy pomocou prepínača *max-attributes*, rodičov triedy pomocou prepínača *max-parents*. Počet výrazov *return* a *yield* je možné obmedziť s využitím prepínača *max-returns*. U niektorých existuje možnosť špecifikovať aj minimálny počet, napríklad u počtu verejných metód triedy pomocou prepínačov *min-public-methods* alebo *max-public-methods* [?].

3.1.4 Zhrnutie

Pylint patrí medzi najpoužívanější program na analýzu kódu Python, nakoľko je ľahko konfigurovateľný a kombinuje funkcionality ostatných nástrojov. Do každého prostredia je vhodné vybrať nástroj na analýzu vždy unikátne. Aj keď Pylint môžeme nazvať najpoužívanším nástrojom, projekty ako Moap, Savon alebo Flumotion využívajú pre testovanie PyChecker.

Program PEP8 má taktiež široké použitie, no len s obmedzením na porušenie konvenčných chýb, ktoré sú v niektorých prípadoch akceptovateľné,

hlavne za účelom sprehľadnenia kódu alebo z historických dôvodov projektu.

3.2 Nástroje na vizualizáciu Python projektu

Nástroje na vizualizáciu nám slúžia na vygenerovanie UML diagramov ale aj na zobrazenie dodatočných informácií o projekte, akými sú napríklad metríky alebo chyby v kóde. V tejto kapitole si predstavíme dostupné nástroje na vizualizáciu Python kódu.

3.2.1 Pyreverse

Pyreverse je program dodávaný spoločne s utilitou Pylint, jeho úlohou je vygenerovať UML diagramy pre Python projekt. Vykonáva syntaktickú analýzu Python balíkov a vygeneruje UML diagram vo formátoch dot alebo vgc. Pyreverse je za pomoci programu dot schopný vygenerovať aj UML diagram vo formáte PNG.

Medzi hlavné výhody programu Pyreverse patrí jeho rýchlosť a široká ponuka prepínačov, ktoré umožňujú ignorovať moduly a súbory, filtrovať typy atribútov alebo ovládať výzor výsledného UML diagramu. Zaujímavou vlastnosťou je vygenerovanie diagramu tried súvisiacich s jednou konkrétnou triedou. Túto vlastnosť umožňuje prepínač `-class=<class>`

Nevýhodou programu Pyreverse je jeho neinteraktívnosť. Pri editácii projektu musíme daný graf pregenerovať. Navyše nemôžeme pomocou programu Pyreverse editovať projekt.

```
$ pyreverse -o png --ignore=gaphas .
```

3.2.2 Pylint-gui

Pylint-gui je program, ktorý bol vytvorený najmä pre užívateľov operačného systému Windows ako alternatíva k spúšťaniu Pylint pomocou príkazového riadka. Rozhranie programu je veľmi jednoduché a jedná sa len o prevedenie konzolového výstupu do grafického rozhrania. Pylint-gui teda neodstránil základné nedostatky programu Pylint. Pri editácii projektu treba výstup znova pregenerovať, nakoľko chyby sú identifikované číslom riadku, ktorý sa pri editácii môže zmeniť.

3.2.3 Implementácia programu Pylint v IDE

Program Pylint bol implementovaný do väčšiny IDE prostredí, ktoré ho graficky aj dynamicky interpretujú. Táto implementácia je často len v podobe doplnkov a modulov bez priamej podpory vývojárov IDE. Ďalšou z nevýhod je veľká robustnosť IDE prostredí a obmedzenosť na dané vývojové prostredie.

Jedným z príkladov týchto doplnkov je modul PyDev pre vývojové prostredie Eclipse. Jeho výhodou je jeho automatické spustenie po zmene súboru, ktoré však ale pri väčších projektoch môže spôsobiť spomalenie. Výhodou je aj možnosť prefiltrovať typ chyby, ktorú Pylint zaznamená, obmedzuje sa ale len na nasledujúcich päť skupín, nie však na konkrétne správy.

- FATAL
- ERRORS
- WARNINGS
- CONVENTIONS
- REFACTOR

Nepodporuje taktiež filtráciu falošných poplachov ani generáciu UML.

3.2.4 Zhrnutie

Existuje množstvo utilít a programov, ktoré pomáhajú výsledky analýzy Python kódu vizualizovať. Okrem vyššie uvedených môžeme spomenúť aj PyCAna, Pymetrics alebo Pypants. Medzi základné problémy ktoré tieto programy majú patrí ich neinteraktívnosť, nutnosť výsledky regenerovať po editácii kódu, robustnosť alebo chýbajúca funkcionality. Pri implementácii projektu sa pokúsím zamerať práve na tieto chýbajúce vlastnosti. Výsledný program by mal byť rýchly a neobmedzovať programátora na jeden druh editoru ako je to napríklad pri doplnkoch IDE.

4 Analýza nástrojov potrebných k implementácií

4.1 Analýza nástroja Gaphas

4.1.1 Základná charakteristika nástroja Gaphas

Gaphas predstavuje zoskupenie knižníc a nástrojov na vykresľovanie grafických objektov na určené elementy grafického rozhrania GTK. Je naprogramovaný v jazyku Python a vydaný pod ??? licenciou.

4.1.2 Popis Gaphas API

Gaphas API využíva MVC návrhový vzor a môžeme ho teda rozdeliť na 3 hlavné časti - model, view a controller.

- Model - Je časť API, ktorá obsahuje
 - api/canvas - plátno na vykresľovanie, ktoré sa správa ako kontajner pre vykresľované položky. Canvas v sebe zahŕňa atribúty, ktoré súvisia s vykresľovaním napríklad atribút gaphas.Solver alebo atribút __connections uchovávajúci spájania medzi jednotlivými položkami.
 - api/items - položky gaphas.item.Element a gaphas.item.Line odvodené od triedy gaphas.item.Item ktoré sú uchovávané a vykresľované na plátne. Jednoduchým odvodením novej triedy z gaphas.item.Item môžeme vytvoriť vlastný prvok na vykresľovanie.
 - api/connectors - v tejto kategórii sa jedná o triedy gaphas.connector.Handle a gaphas.connector.Port, ktoré umožňujú spájať jednotlivé položky na plátne.
 - api/solver - umožňuje definovať podmienku medzi aspoň dvomi premennými a a udržiavať túto podmienku ako pravdivú pri zmene premenných.
 - api/constraint - obmedzenia pre premenné, ktoré sú zaregistrované na plátne. Každé obmedzenie obsahuje zoznam premenných, ktoré sú zaregistrované v objekte typu gaphas.solver.Solver.
 - api/utils - obsahuje pomocné funkcie týkajúce sa najmä vykresľovania textu na plátno

- View - Obsahuje všetko súvisiace so zobrazovaním a vykresľovaním jednotlivých elementov.
 - api/view
 - api/painters
 - api/gtkview
- Controller - tools

4.2 Zhrnutie

5 Projekt gpylint

5.1 Aplikácia jednotlivých nástrojov

5.2 Vyhodnotenie

6 Záver