Dokumentace úlohy SYN: Zvýraznění syntaxe v Python 3 do IPP 2011/2012
Jméno a příjmení: Radek Ševčík
Login: xsevci44

# 1 Úvod

Tento dokument popisuje implementační řešení pythonovského skriptu, který slouží ke zvýraznění částí textu pomocí HTML tagů.

# 2 Popis řešení

## 2.1 Parametry příkazové řádky

Zpracování parametrů zajišťuje modul `getopt`. Skript pracuje s těmito parametry: `format input output br`. Kódování souborů musí být v UTF-8.

Skript se chová jako textový filtr, tzn. není-li zadán vstupní/výstupní soubor parametrem, provede se čtení/zápis standartním vstupem/výstupem.

## 2.2 Formátovací soubor

Formátovací soubor sestává z jednotlivých záznamů ve formátu `<reg. výraz>\t<form. parametry>`, kde jednotlivé parametry jsou mezi sebou odděleny čárkou.

Formátovací parametry:
`bold` - tučný text
`italic` - kurzíva
`underline` - podtrhnutí
`teletype` - strojopis
`size:[cislo]` - velikost textu `1..7`
`color:[hex]` - barva textu v `RRGGBB 000000..FFFFFF`

Regulární výraz:
`A.B` - vyhovuje řetězec A následovaný řetězcem B
`AB` - zkratka pro `A.B`
`A|B` - vyhovuje řetězec A nebo B
`!A` - vyhovují řetězce neobsahující A, aplikovatelná na jeden znak nebo speciální znak `%X`
`A*` - vyhovuje řetězec A opakovaný `0..x` krát
`A+` - vyhovuje řetězec A opakovaný `1..x` krát
`(A)` - závorky určují prioritu
Priorita operátorů: `! > *, + > . > |`

Speciální výrazy:
`%s` - bílé znaky ( \t\n\r\f\v)
`%l` - malá písmena `a..z`
`%d` - číslice `0..9`
`%W` - písmena a číslice `a..z, A..Z, 0..9`
`%n` - nový řádek \n
`%|` - znak `|`
`%*` - znak `*`
`%(` - znak `(`
`%%` - znak `%`
`%a` - libovolný znak
`%L` - velká písmena `A..Z`
`%w` - malá a velká písmena `a..z, A..Z`
`%t` - tabulátor \t
`%.` - znak `.`
`%!` - znak `!`
`%+` - znak `+`
`%)` - znak `)`

Zpracování formátovacího souboru zajišťuje funkce `parse_fmtfile`, která načítá jednotlivé záznamy po řádcích. Je-li řádek prázdný, pak jej přeskočí a pokračuje na dalším. Regulární výrazy převádí funkce `parse_expr` do podoby srozumitelné modulem `re`. O jednotlivé formátovací parametry se stará funkce `parse_fmt`, která kontroluje rozsah hodnot a syntaxi. Výsledek se uloží ve třídě `FormatItem`, která je tvořená kompilovaným regulárním výrazem a dvěmi `deque`, do kterých se vkládají jednotlivé HTML tagy zleva a zprava v pořadí uvedeném ve formátovacím souboru. Byl-li zadán parametr `br`, pak je také instancí této třídy. Výsledkem je pole těchto tříd, které se předají k dalšímu zpracování.

### 2.3 Vstup a výstup

Zpracování probíha ve funkci `process`. Ke každému formátovacímu záznamu se naleznou všechny výskyty a uloží se jejich pozice začátek-konec a příslušné tagy do dvou polí. Seřadí se pole tagů zleva a zprava a poté se spojí dohromady převrácené pole zprava a k němu zleva a opět se seřadí. Tímto způsobem se dá vyřešit překrývání, kde je zapotřebí aby otevřené a uzavřené tagy byly uspořádané podle standartu. Pak se jednoduše zapíše text ze vstupu a na určitých pozicích se připíší tagy.

## 3 Závěr

Tento skript byl vyvíjen a otestován na operačním systému CentOS 5.8 x64 (GNU/Linux) sadou přiložených testů s nastaveným prostředím `LC_ALL=cs_CZ.utf8`.