Dokumentace úlohy SYN: Zvýraznění syntaxe v Python3 do IPP 2015/2016
Jméno a příjmení: Petr Flajšingr
Login: xflajs00

Úvod

Předmětem projektu je zvýraznění syntaxe dokumentu pomocí HTML značek.

Postup řešení

Kontrola argumentů

Pro zpracování argumentů jsem si napsal funkci `getParams()`. Ta využívá `getops` pro získání argumentů programu. Výstup je poté procházen pomocí cyklu a na základě argumentů jsou nastavovány hodnoty v datové struktuře Pythonu - dictionary, který ukládá data pro další funkce programu. Pokud byl zadán neznámý argument nebo byl některý argument zadán chybně, je skript ukončen s chybovou hláškou a návratovým kódem 1.

Datové struktury

Nejdůležitější strukturou celého skriptu je list info. Obsahuje informace o: regulérní výraz, značky, velikosti značek, indexy pro uložení značek.

Načtení vstupních dat

Jak formátovací tak vstupní soubor jsou otevřeny v kódování utf-8 pomocí modulu `codecs`. Data vstupního souboru jsou načítány pomocí funkce `read()`, ale formátovací data pomocí `readlines()`, kvůli jednodušší práci v následujících funkcích. V případě že vstupní soubor nelze otevřít, končí skript s chybovou hláškou a návratovým kódem 2. Pokud je formátovací soubor prázdný nebo nelze otevřít, je rovnou vypsán vstupní soubor (s případným zpracováním parametru --br).

Zpracování formátovacího souboru

Pokud je formátovací soubor správně zadán, je dále zpracován pomocí funkce `formatFile()`. Zpracování probíhá po řádcích. Každý řádek je rozdělen podle `\t`.

První část (regulérní výraz) je poslána funkci `getReg()`. Ta nejdříve obstará tečky v regulérním výrazu a poté změní výraz na takový, kterému bude Python rozumět. Pomocí `re.compile()` probíhá i kontrola platnosti daného výrazu. Druhá část (typ značky) je zpracována funkcí `getTag()`. Ta odstraní bílé znaky a rozdělí značky podle čárek a převede požadované značky do HTML podoby. Pokud je výraz nebo značka neplatná, je program ukončen s chybovou hláškou a návratovým kódem 4. Po zpracování jsou oba údaje uloženy do listu `info`.

Aplikování požadovaných změn

Po zpracování formátovacího souboru je volána funkce `findPos()`, která v textu najde pozice řetězců, na které budou později aplikovány značky. Vyhledávání je realizováno pomocí funkce `re.finditer()`. Informace o indexu řetězců jsou ukládány do listu info.

Dalším krokem je využití funkce `addTags()`. V této funkci jsou postupně aplikovány všechny tagy na všechny požadované řetězce. Při každé aplikaci značky jsou aktualizovány indexy nezpracovaných řetězců v `info`. Funkce pracuje tak, že nejdříve aplikuje jedna značka na celý text a až poté se posune na další. Toto a změna indexů zaručuje, že značky se budou korektně překrývat.

Výstup

Před vytisknutím textu do souboru (pokud byl zadán) nebo na `stdout` je aplikován případný parametr --br pomocí funkce `re.sub()`. Výpis do souboru probíhá v kódování utf-8 pomocí modulu `codecs`. V případě neplatného nebo nedostupného výstupního souboru je skript ukončen s chybovou hláškou a návratovým kódem 3.