Dokumentace úlohy CST: C Stats v Python 3 do IPP 2014/2015
Jméno a příjmení: Roman Halík
Login: xhalik01

Úvod

Tato dokumetace se věnuje principu fungování skriptu, napsaném v jazyce Python, jehož účelem je výpis statistik souborů s příponami „.c“ a „.h“, tedy souborů se zdrojovým textem v jazyce C. Skript umožňuje výpis informací o počtu klíčových slov, identifikátorů, operátorů, počtu znaků v komentářích a vyhledání konkrétního řetězce. Prohledávání probíhá v zadaném souboru, případně pro všechny podsoubory a podadresáře zadaného adresáře, nebo pouze pro soubory zadaného adresáře bez prohledávání podadresářů.
Skript lze rozdělit na několik částí. Nejprve se zpracují vstupní parametry, které se pak předávají funkci, která určí, co se bude vyhledávat. Následně proběhne vyhledání a vrácení výsledků funkci, která výsledky zpracuje a pošle na standardní výstup, nebo je zapíše do souboru.

Zpracování vstupních parametrů

Zpracování parametrů je implementováno za pomocí funkce getopt, avšak bylo třeba dopsat kontrolu opakování parametrů (žádný nesmí být zadán vícekrát) a kolize přepínačů –k, –o, –i, –w, –c (musí být zadán právě jeden z nich). Jestliže nastane při zpracování parametrů chyba, je provádění skriptu ukončeno s odpovídající návratovou hodnotou, a je vypsána chybová hláška na stderr. Pokud vše proběhne v pořádku, pak se u parametrů s hodnotou tato hodnota uloží pro pozdější použití. Stejně tak se uchovává informace o tom, zdali byl parametr zadán. Pořadí parametrů není pevně dané.

Zpracování zdrojových souborů

Skript umožňuje výpis celkem 5 různých statistik zdrojových souborů. Konkrétně počet klíčových slov, počet identifikátorů, počet operátorů, počet výskytů konkrétního zadaného řetězce a počet znaků v komentářích. Pro první tři statistiky je společné odfiltrování řetězců, maker a komentářů, využívají tedy stejnou funkci. Jinak se zpracování děje odděleně pro každou jednu úlohu.

Klíčová slova (-k)

Nejprve se odstraní nevyhovující části zdrojového textu, a následně se pomocí regulárního výrazu vyhledají všechna slova, která se nacházejí v seznamu klíčových slov. Jestliže je nalezena shoda a toto slovo ještě není v seznamu nalezených klíčových slov, je do tohoto seznamu přidáno. Nakonec se spočítá počet prvků seznamu a vrátí se jako výsledek vyhledávání.

Operátory (-o)

Stejně jako u klíčových slov, nejprve dojde k odstranění nevyhovujících částí zdrojového textu, a následně je ještě třeba odstranit deklarace ukazatelů, jelikož „*“ není v takovém případě uvažována jako operátor. Následně jsou operátory vyhledány regulárním výrazem a je vrácen jejich počet.

Identifikátory (-i)

Vyhledání identifikátorů je principielně velmi podobné jako vyhledávání klíčových slov, avšak namísto vyhledávání shody nalezeného slova v seznamu klíčových slov je hledána neshoda. Tedy pokud se nejedná o klíčové slovo, musí se jednat o identifikátor.

Přesný řetězec (-w=pattern)

Vrací počet výskytů zadaného řetězce (pattern) v celém zdrojovém textu. Toto vyhledávání je case-insensitive.

Komentáře (-c)

Vyhledání a spočtení znaků komentářů je implementováno jako konečný automat. Prohledává se v celém zdrojovém textu znak po znaku, včetně maker (rozšíření COM). Konečný automat má celkem 6 stavů, které zabezpečují korektní chování pro každou situaci, která může nastat.

Zpracování výsledků a výstup

Výsledky vrácené od funkcí obsluhující přepínače je třeba správně vypsat a zarovnat a seřadit. Jelikož jsou výsledky uloženy v seznamu, je nejjednodušší provést řazení v tomto bodě. Následně se sečtou počty nálezů hledaných entit a připojí se k seznamu. Poté je vypočítána maximální délka řádky, která je součtem nejdelší cesty prohledávaného souboru (případně jen jeho názvu při zvoleném přepínači –p), a největší délky (počet znaků) počtů výskytů. Tím je známa maximální délka řádku.

Následně jsou výsledky zapisovány do řetězce, přičemž je pro každou dvojici cesta-počet vypočítána mezera mezi těmito dvěma údaji, která je rovna maximální délce, od které je odečtena délka cesty a počtu pro aktuálně vypisovaný soubor. Aby nedošlo k „slepení“ údajů u nejdelších údajů, je každá dvojce oddělena ještě jednou mezerou. Následně se tento řetězec zapíše do souboru, nebo na standardní výstup.