Dokumentace úlohy CHA: C Header Analysis v PHP 5 do IPP 2013/2014
Jméno a příjmení: Michal Kozubík
Login: xkozub03

1 Popis řešení úlohy

Pro řešení registrovaného zadání projektu (analýza hlavičkových souboru programovacího jazyka C, vyhovujících standardu ISO C99) jsem zvolil kombinaci konečného automatu a regulárních výrazů.

1.1 Kontrola parametrů

Skript nejprve provádí kontrolu načtených parametrů. To je realizováno prostřednictvím funkce „getopt“, přičemž ještě probíhá testování, zda-li nebyl některý parametr zadán vícekrát nebo v chybném tvaru a jestli nejsou přítomny i parametry, které skript nezná.

1.2 Odstranění nepodstatných částí

Poté je načten celý obsah zadaného souboru (případně postupně každého souboru v zadaném adresáři a jeho podadresářích) do paměti jako jeden řetězec. Následně se předá řízení konečnému automatu, který z tohoto souboru odstraní sekvence komentářů, maker a řetězců, jelikož by se v nich mohly objevovat nechtěné shody s pozdějším vyhledáváním deklarací a definic funkcí. Samotný řetězec je ještě před zpracováním rozdělen do pole znaků proměnné délky, nad kterým je analýza prováděna (z důvodu proměnné velikosti znaků v kódování UTF-8).

Při procházení vstupu se ukládá pozice, kdy začala kódová část (potřebná pro pozdější analýzu) a při nalezení začátku nepodstatné části se tento úsek přenese do výsledného textu. Po ukončení nepodstatné části se začátek vyhovující pasáže aktualizuje za tuto pozici.

Konečný automat jsem pro tuto fázi zvolil z toho důvodu, že regulární výraz (příp. více výrazů) pro odstranění všech nechtěných částí by byl velice složitý náchylný k chybám (jestli by vůbec byl možný – z důvodů překrývání komentářů s řetězci atp.). Stejně tak pro další zpracování jsem zvolil regulární výrazy, jelikož tvar deklarace a definice funkce už nemá tak jasně daný tvar jako komentáře, makra a řetězce.

1.3 Vyhledání a zpracování funkcí

Po odstranění nadbytečných částí je tedy na zbytek souboru aplikováno vyhledávání shod s regulárním výrazem odpovídajícímu tvaru definice nebo deklarace funkce. Tento regulární výraz již v sobě obsahuje pojmenované podskupiny výrazu, díky čemuž jsou ve výsledném poli polí shodujících se částí souboru tyto dílčí shody indexovány názvy skupin. Následné rozpracování nalezené funkce je proto jednoduché rozdělit na název, návratový typ a parametry funkce.

Zpracování parametrů je podobné vyhledávání funkce v celém souboru. Nejprve je celá část rozdělena na podčásti oddělené znakem čárka. Na každou podčást je aplikován regulární výraz zajišťující roztřídění každého parametru na název a návratový typ. Není-li jedna z částí (název nebo typ) nalezen, znamená to, že celý parametr odpovídá pouze návratovému typu a jméno není uvedeno. Touto kontrolou je také realizováno rozšíření PAR (tedy chybějící název parametru). U každého parametru zároveň probíhá kontrola, zda-li se nejedná o identifikaci proměnného počtu parametrů (...). Pokud ano, je parametr přeskočen.

Při zpracování parametrů se také inkrementuje počítadlo parametrů, které se poté porovnává s hodnotou požadavku na maximální počet parametrů (pokud byl uveden přepínač „--max-par=n“). Při překročení této hodnoty je celá funkce přeskočena.

1.4 Výstup

Během zpracování funkcích a jejich parametrů jsou zjištěné informace ukládány do stromové struktury XML souboru pomocí rozšíření DOM. Je nutné zajistit správné formátování výstupu.. Proto je XML struktura nejprve převedena na řetězec a následně upravována funkcemi pro nahrazení částí textu. Při zadaném parametru pro zformátování výstupu je tak přidáno požadované odsazení každé nové úrovně (příp. defaultní, není-li uvedeno) a oddělení hlavičky dokumentu od kořenového elementu struktury znakem nového řádku. Takto upravený řetězec je teprve uložen do výstupního souboru.

2 Konečný automat pro odstranění nepotřebných částí souborů

pozn.:

- XX značí všechny ostatní znaky, které nejsou obsaženy v jiných výstupních hranách daného uzlu.
- Automat není implementován přesně jako je zobrazen níže, v kódu jsou použity pomocné proměnné pro zjednodušení (např. 1 stav pro řetězce, v proměnné je uložen ukončující znak atp.).