Zadání úlohy do projektu z předmětu IPP 2010/2011

Zbyněk Křivka a Dušan Kolář
E-mail: {krivka, kolar}@fit.vutbr.cz, {54 114 1313, 54 114 1238}

SMS: SMS Compress

Zodpovědný cvičící: Martin Čermák (icermak@fit.vutbr.cz)

1 Detailní zadání úlohy

Vytvořte skript pro kompresi zadané krátké textové zprávy (SMS), odstranění české diakritiky a případnou aplikaci pravidel ze slovníku zkratek.

Tento skript bude pracovat s těmito parametry:

- `--help` viz společné zadání všech úloh
- `--input=filename` zadaný vstupní textový soubor s krátkou textovou zprávou
- `--output=filename` textový výstupní soubor v kódování UTF-8
- `-r` provede odstranění české diakritiky a nahrazení stejnými znaky bez diakritického znaménka. V případě kombinace s `-c` nebo `-v` se provede odstranění diakritiky jako první.
- `-c` provede kompresi SMS podle dodatečných nastavení typu Camel, tj. převod každého slova[1] na Camel notaci[2] a vynechání přebytečných mezer (první písmeno slova je velké, všechna ostatní jsou malá).
- `-a` na Camel notaci převádí pouze slova[1] ze samých malých písmen (vyžadována kombinace s parametrem `-c`, jinak chyba). Pokud je tedy část SMS napsána již v Camel notaci, tak ze slova "ZeSlovaTakovehoto" nevznikne slovo "Zeslovatakovehoto", ale zůstane zachováno "ZeSlovaTakovehoto".[3]
- `-b` v kombinaci s `-c` (nepovoleno kombinovat s `-a`) vynechá z komprese slova[1] napsaná velkými písmeny (např. zkratky). Mezery před a za těmito slovy nezachovávejte.
- `--dict=filename` určení XML slovníku zkratek (obsahuje pravidla pro zkracování a expanzi zkratek, viz níže), pro jméno souboru platí stejná pravidla jako pro zadávání vstupního či výstupního souboru, implicitní hodnota však neexistuje. Tento parametr vyžaduje uvedení parametru `-v`.
- `-v` provede aplikaci pravidel ze zadaného slovníku zkratek (vyžadováno určení parametrem `--dict`, jinak chyba). Aplikace se provádí před samotnou kompresí, pokud je komprese parametrem `-c` vyžadována. Nejprve se provede expanze a poté zkracování. Je-li zadán parametr `-e` nebo `-s`, tak se může provést jenom jedna z aplikací slovníku zkratek (tj. buď expanzivní, nebo zkracující pravidla).

[1] *Slovo* je neprázdný řetězec složený pouze z písmen z obou stran oddělený nepísmenným znakem, tj. bez mezer, interpunkce, číslic a jiných znaků.

[2] Např. "Testovací text" převede na "TestovacíText" nebo "test, JAK38 jinak" převede na "Test,Jak38Jinak".

[3] Špatné použití Camel notace, jako např. začátek malým písmenem, neřešte.

- `-e` aplikuje pouze expanzivní pravidla ze zadaného slovníku zkratek (vyžaduje parametr `-v`; nelze kombinovat s `-s`).
- `-s` aplikuje pouze zkracující pravidla ze zadaného slovníku zkratek (vyžaduje parametr `-v`; nelze kombinovat s `-e`).
- `-n` vypíše minimální počet SMS, na které je nutno výstupní SMS rozdělit. Uvažujme, že jedna SMS může mít 160 znaků bez diakritiky či 70 znaků s diakritikou, ale v případě rozdělení na více zpráv již pouze 153 znaků bez diakritiky nebo 67 znaků s diakritikou na každou zprávu. Výstupem bude v tomto případě pouze číslo (počet SMS) a převedená SMS bude zahozena. Výpočet počtu rozdělených zpráv se provádí na základě výstupu po odstranění diakritiky, po aplikaci pravidel a po komprimaci (je-li to vyžádáno odpovídajícími parametry).

Slovník zkratek ve formátu XML je v kódování UTF-8. Po XML hlavičce `<?xml version="1.0" encoding="UTF-8"?>` následuje kořenová značka `<sms-abbreviation-dictionary>`, která obsahuje pravidla `<rule>` (párový element) s atributy `expansive="1"` (v případě expanzivního pravidla; jinak je pravidlo zkracující) a `casesensitive="1"` (v případě, že pravidlo přesně dbá na velikost písmen zkratek a textů; bez uvedení tohoto atributu nezáleží na velikosti písmen). Každé pravidlo obsahuje dva párové elementy `<abbrev>` a `<text>` (POZOR! Podle definice XML nezáleží na pořadí elementů.). Elementy `<abbrev>` a `<text>` mohou obsahovat pouze text a první definuje zkratku a druhý definuje text, ze kterého nebo na který se zkratka přepisuje.

Reference:

- `http://en.wikipedia.org/wiki/SMS`
- `http://www.dreamfabric.com/sms/default_alphabet.html`

2 Bonusová rozšíření

DIV (až 2 body): Rozdělte SMS na několik SMS tak, aby jejich počet byl roven nejmenšímu možnému počtu SMS (po zpracování podle zadaných ostatních parametrů) a zároveň aby byly splněny podmínky na nejlepší možné rozdělení, které nezvýší počet SMS po rozdělení. SMS zprávy jsou ve výsledku odděleny jedním prázdným řádkem (neboli dvěma konci řádku). Nejlepší rozdělení reflektuje věty SMS zprávy, takže rozdělení může nastat pouze za interpunkcí danou znaky `.`, `,` a `;`. Další o něco méně vhodné rozdělení reflektuje slova[4], takže nejsou rozdělena slova uprostřed. Pokud nelze zajistit ideální rozdělení u všech SMS, tak se snažte minimalizovat penalizační funkci, která při každém rozdělení nereflektující věty, ale reflektující slova dává postih 1 bod a při nereflektování ani slov dává postih 3 body.

3 Specifické požadavky na dokumentaci

Popište techniku odstranění české diakritiky.

4 Poznámky k hodnocení

Výsledný výstup skriptu je přesně porovnán nástrojem `diff` s očekávanými výstupy.

[4]Předchozí definici *slova* intuitivně rozšiřte i na provedenou kompresi do Camel notace.