

Praktický úkol č. 10

Návrh jednoduchého kompresního algoritmu pro soubor textu v Morseově abecedě.

Soubory:

morse-text.txt (4 930 B) - zdrojový anglický text
morse-code.txt (14 734 B) - anglický text přeložený do Morseovy abecedy
MorseTranslator.java - jednoduchý program pro překlad do Morseovy abecedy

Doplňující informace:

morse-code.txt.zip (2 621 B) - pokusná komprimace metodou ZIP
morse-text.txt.zip (2 268 B) - pokusná komprimace metodou ZIP

Popis:

Soubor morse-code.txt obsahuje text v Morseově abecedě. Tento text je kódovaný třemi možnými znaky. Ve znakové soustavě používané počítačem lze do 1 Byte zapsat 256 různých čísel (0-255). Jeden Byte znak lze chápat jako číslo v číselné soustavě se základem 256. Na morse-code.txt lze pak nahlížet jako na jedno velké číslo zapsané v trojkové soustavě (znaky "-./"). Do 14 734 znaků v trojkové soustavě lze napsat maximální číslo 3^{14734} . Takové číslo lze ale efektivněji přepsat do dvojkové (či 256tkové) soustavy a ze vztahu pro maximální číslo v jednotlivých soustavách můžeme získat nutný počet Byte pro jeho zapsání.

$$3^{14734} = 256^x \quad \text{z toho} \quad 14734 * \ln(3) = x * \ln(256) \quad \Rightarrow \quad x \approx 2920$$

Bez znalosti obsahu dat je tedy možné komprimovat pouhým převodem mezi číselnými soustavami text v Morseově kódu do 2920 B, což je rámcově porovnatelné s kompresní metodou ZIP. Je však nutné si uvědomit, že převod prostého textu do Morseova kódu není sám o sobě bezztrátový, ztrácí se při něm například velikost písmen, speciální a formátovací znaky.

Úloha:

Zkuste navrhnout algoritmus pro čtení textu v trojkové soustavě a jeho zápis do binární soustavy, ideální výstupní data v ukázkovém případě by měla zabírat zmiňovaných 2920 B.

Pokud byste dospěli k závěru, že je převod příliš komplikovaný, zkuste využít soustavu o základu 4 pro skupiny znaků (".", "-", "/", ""); koncový řetězec "////" lze vypustit a nahradit koncem souboru.

Dobrovolná úloha:

Zkuste navrhnout nějaký jiný postup kompresního algoritmu, vyzkoušejte, porovnejte výsledky s existujícími komprimačními programy.

Poznámka:

Úlohu jsem se snažil sestavit tak, aby pro Vás byla zajímavá, pokud by tomu tak nebylo, ozvěte se mi.