ADS 2023/2024

Príhody doktora Vojšiča, epizóda tretia: Jarné upratovanie

Doktor Vojšič sa rozhodol, že svoje každoročné jarné upratovanie tentoraz začne tým, že si konečne urobí poriadok vo svojom počítači. Aby ušetril miesto na disku, rozhodol sa, že si svoje cenné údaje s vedeckými experimentmi skomprimuje pomocou algoritmu LZW (Lempel-Ziv-Welch), v knihe kap. 5.5:

http://index-of.es/Varios-2/Algorithms%204th%20Edition.pdf

Po čase zistil, že ušetrená kapacita je pre niektoré súbory len veľmi malá. Preto sa rozhodol, že si pre každý súbor najprv vypočíta ***Vojšičov komprimačný koeficient*** (VKK) podľa vzorca:

$$VKK = \frac{VKS_{LZW}}{VPS},$$

kde $\boldsymbol{VKS_{LZW}}$ označuje veľkosť súboru po kompresii LZW a ***VPS*** označuje veľkosť pôvodného súboru (obidve veličiny sú vyjadrené zhodne v počte bajtov). Usúdil, že sa mu oplatí komprimovať iba tie súbory, pre ktoré je Vojšičov komprimačný koeficient lepší ako 0.9 (teda menej ako 90%).

Úloha. Naprogramujte kompresný algoritmus LZW aj Vy tak, aby ste pre každý vstupný súbor dokázali vypočítať aj Vojšičov komprimačný koeficient.

Váš algoritmus by mal vypočítať počet bajtov v pôvodnom súbore a po skomprimovaní metódou LZW zistiť aj počet bajtov v skomprimovanom súbore a následne aplikovať vzťah pre výpočet *VKK*.

Predpokladajte, že všetky vstupné súbory obsahujú znaky iba zo štandardnej sady ASCII (teda nie rozšírenej). Pri implementácii algoritmu LZW predpokladajte, že vytváraný slovník bude mať maximálnu kapacitu **4096** položiek. Pre kódovanie použite jednu z nasledujúcich alternatív:

- prvých **128** položiek je vyhradených pre jednotlivé bajty vstupného súboru (v ASCII),
- **(nepovinné)** použite kódovanie podľa textovej abecedy so znakmi ’a‘, .., ‚z‘, ‚A‘, ... ‚Z‘, ‚0‘, ..., ‚9‘ (spolu 62 znakov).

Algoritmus otestuje pre vstupný súbor ***udaje5a.txt*** a vyčíslite VKK pre tento súbor.

Algoritmus otestuje aj pre vstupný súbor ***udaje5b.txt***.

Nájdite text s aspoň 25 znakmi, pre ktorý algoritmus LZW bude mať *VKK* nanajvýš 0,9.