Programovanie (1) v C/C++ 2024/25
Cvičenia 10, príklad 3
Jazyk

V tejto úlohe napíšete program, ktorý sa bude snažiť zistiť, v akom jazyku je napísaný určitý text, na základe frekvencie výskytu jednotlivých písmen abecedy.

Na konzole dostanete názvy troch vstupných súborov (názvy súborov nebudú dlhšie ako 20 znakov a neobsahujú biele znaky). V prvom súbore je text v neznámom jazyku, v druhých dvoch sú ukážky textov v dvoch rôznych jazykoch. Na základe frekvencií písmen sa váš program pokúsi určiť, v ktorom z týchto dvoch jazykov je text v prvom súbore. Vo všetkých troch súboroch budeme spracovávať iba malé písmená anglickej abecedy. Veľké písmená sa v texte nenachádzajú a ostatné znaky (medzery, čísla, interpunkcia a podobne) budeme ignorovať.

Všetky tri vstupné súbory prečítajte (napríklad znak po znaku príkazom `getc`) a uložte si do tabuľky počty výskytov jednotlivých písmen `a` až `z`. Napríklad v prvom ukážkovom vstupe `slovak.txt` je 21 337 písmen, z toho $2\,618 \times$ `a`, $475 \times$ `b`, $721 \times$ `c`, ..., $728 \times$ `z`. Potom tieto počty prepočítajte na percentá tak, že každý počet vynásobíte číslom 100 a celočíselne vydelíte celkovým počtom (dostaneme tak percentá zaokrúhlené nadol na celé číslo). Napríklad písmeno `a` tvorí z korpusu približne `2618*100/21337=12`%.

Na konzolu vypíšte pre každý súbor najskôr jeho názov a na ďalší riadok tabuľku percentuálneho zastúpenia jednotlivých písmen. Každá tabuľka je riadok s 26 celými číslami oddelenými medzerami a predstavujúcimi percentuálne zastúpenie výskytov písmen `a` až `z`. Pre druhé dva vstupné súbory pod tabuľku vypíšte číslo vyjadrujúce, ako veľmi sa tabuľka ich frekvencií odlišuje od tabuľky pre prvý súbor. Odlišnosť dvoch tabuliek a a b spočítajte ako súčet absolútnych hodnôt rozdielov ich zodpovedajúcich prvkov $|a[i]-b[i]|$. Absolútnu hodnotu počíta funkcia `abs` z knižnice `cstdlib`. Napokon na posledný riadok výstupu vypíšte meno toho z druhých dvoch súborov, ktorý mal túto odlišnosť menšiu a teda potenciálne ide o jazyk zdrojového textu,

Na stránke nájdete štyri sady vstupných súborov. Uvádzame, ako má vyzerať výstup na konzolu pre prvú z nich.

Príklad výstupu na konzolu:

```
unknown.txt
10 1 2 4 7 0 0 1 6 2 3 2 3 6 9 3 0 5 5 5 5 3 0 0 2 1
slovak.txt
12 2 3 4 9 0 0 2 6 1 3 4 3 5 9 2 0 4 5 5 3 3 0 0 2 3
17
latin.txt
7 2 1 2 11 1 1 0 10 0 0 1 6 6 4 1 0 6 8 8 9 4 0 0 0 0
49
slovak.txt
```