2 Hľadanie reťazcov

Vytvoril/a Peter Koscelanský [Administrator]
Posledná aktualizácia: Oct 19, 2021

Vytvorte konzolovú aplikáciu, ktorá dostane tri pozičné argumenty

- cestu k textovému súboru *(ak súbor neexistuje alebo sa z neho nedá čítať skončite s nenulovým error kódom),*
- neprázdny reťazec znakov `X` *(akceptuje maximálnu dĺžku ostro menšiu 256, pre väčšie reťazce treba skončiť s nenulovým error kódom),*
- číslo `N` označujúce vzdiatenosť *(tu akceptujeme všetky* `uint32_t` *okrem nuly, inak nenulový error kód).*

Program následne na štandardný výstup vypíše v riadkoch pozície, kde sa nachádza výskyt aký chceme. Prvý znak v súbore má pozíciu 0, rovnako aj prvý riadok je 0). Whitespace sa ráta do dĺžky súboru a konce riadkov sú vždy jeden znak (teda s tým nemusíte nič robiť, to za vás vyrieši C++). Súbory budú obsahovať iba znaky s hodnotou 0 < c <= 127, ak by bol v súbore aj iný znak, tak správanie je nešpecifikované.

Hľadáme výskyty reťazca `X`, ktoré majú vo svojom okolí (definovnom pomocou `+-N`) ďalší výskyt reťazca `X`.

Ak máme súbor s obsahom:

```
aaabaaabbbbbbbbbaaa
```

`X = aaa` a `N = 4`, tak výsledkom budú pozície `0, 4` . Lebo prvý výskyt má vo vzdialenosti `4` iný výskyt reťazca, konkrétne na pozícií `4` a `|4 - 0| <= 4` druhý výskyt tam tiež zarátame, lebo pred ním je vo vzdialenosti `4` ten náš prvý výskyt. Posledný výskyt nepoužijeme, lebo nemá nikde v okolí ešte jeden výskyt.

Keďže v príklade vyššie je všetko na jednom riadku, vysledok bude teda

```
0 0
0 4
```

Samotná podmienka blízkosti neberie do úvahy riadky, koniec riadku je vlastne jeden znak. Riadky a pozície v nich sa berú do úvahy iba vo výpise.

Ďalšie príklady

`X = aaa, N = 2`

```
aaaaabbbbb
bbaaaaa
```

```
0 0
0 1
0 2
1 2
1 3
1 4
```

V príklade hore sa výskyty prelínajú, ale vôbec nevadí, je tam splnená podmienka blízkosti.

`X = aaabaaa, N = 2`

```
aaabaaabaaa
```

Tu bude výstupom nič, síce sa oba stringy prelínajú, ale začiatok prvého je `0` a začiatok druhého je `4` a teda ich rozdiel je väčší ako `2`.

Pre vstupný súbor input.txt a `X = bbb, N = 20` je výsledok

```
4 41
4 45
6 29
6 30
```

Súbor môže byť obrovský, pokojne aj niekoľko desiatok GB, takže neukladajte celý jeho obsah do pamäte (pozor na `std::getline`). Ako vždy sa hodnotí to aby boli všetky chybové stavy ošetrené (nenulový error kód). Tentokrát sa bonusový body získava za rýchlosť. Čiže aby to aj nejak bežalo. Prvých desať implementácií dostane bonusový bod.

Páči sa mi to Buďte prvý(-á), komu sa to páči Žiadne označenia

27 page comments

Peter Koscelanský [Administrator] Oct 19, 2021

Na cviceni bola otazka, ci novy riadok sa rata ako znak a teda vlastne “resetuje“ ten string co hladame. Ano novy riadok je normalny znak a ifstream::read vam ho vrati.

Odpovedať • Páči sa mi to

Anonymný Oct 19, 2021

Dobry den, mam otazku k argumentom, na potrvrdenie, ze som pochopil.

1) Takto zadane argumenty **su spravne**:

```
./main input.txt aaa 6
```

subor = input.txt

hladany retazec = aaa

vzdialenost = 6

2) Takto zadane argumenty **su tiez spravne**

```
./main aaa input.txt 6
```

subor = aaa

hladany retazec = input.txt

vzdialenost = 6

3) menej ako 3 alebo viac argumentov a ak treti argument nie je uint32_t je **nenulovy error**, napr:

```
./main input.txt aaa 6 bbb
./main input.txt aaa
./main input.txt aaa bbb  ||  ./main input.txt aaa -6  ||  ./main input.txt aaa 0
```

Dakujem.

Odpovedať • Páči sa mi to

Peter Koscelanský [Administrator] Oct 19, 2021

1. ano

2. nie, prvy musi ist subor

3. ano vsetky su, ze treba nenulovy error kod

Odpovedať • Páči sa mi to

Anonymný Oct 19, 2021

> subor = aaa
>
> hladany retazec = input.txt
>
> vzdialenost = 6

no v tej 2) som to tak napisal, ze bude otvarat subor “aaa” a ako retazec zoberie “input.txt”

Odpovedať • Páči sa mi to

Peter Koscelanský [Administrator] Oct 20, 2021

Aha, ano mas pravdu, nevsimol som si to. Je to tak ako pises. Prvy je subor, potom retazec a na zaver vzdialenost.

Odpovedať • Páči sa mi to

Anonymný Oct 19, 2021

Ak som nacital zo suboru string `"abc\naabc"` a hladam retazec **"abc"** s **N = 4**, tak v tomto pripade `\n` sa pocita ako znak pri vypocte vzdialenosti a preto na vypis nic nedostanem, kedze `|5-0| > 4`?

Odpovedať • Páči sa mi to

Peter Koscelanský [Administrator] Oct 20, 2021

Ano znak novy riadok (\n) sa pocita do znakov.

Odpovedať • Páči sa mi to

Anonymný Oct 22, 2021

Aká je maximálna dĺžka riadka? Teda ak by som našiel zhodu v riadku na pozícii za veľkosťou uint32_t, tak by som nevedel vypísať asi číslo, že na akej som pozícii 😅.

Odpovedať • Páči sa mi to

Peter Koscelanský [Administrator] Oct 26, 2021

tak ja odporucam pouzit aspon int64_t, to uz isto nepresiahneme.

Odpovedať • Páči sa mi to

Anonymný Oct 26, 2021

ok, vďaka 🙂

Odpovedať • Páči sa mi to

Anonymný Oct 30, 2021

Ja by som sa k tomuto tiez chcela spytat - ja to nacitavam do size_t, prislo mi to ako najvhodnejsie riesenie. Avsak ked sa to spusti na 32 bitovom systeme, tak to bude malo :(. Mozem ratat s tym, ze take cosik sa nestane, alebo tam teda mam podavat fixne tych 64? :D

Odpovedať • Páči sa mi to

Peter Koscelanský [Administrator] Oct 30, 2021

NIe je to idealne ale asi to na tuto ulohu bude stacit.

Odpovedať • Páči sa mi to

Anonymný Oct 22, 2021

Moze hladany substring obsahovat \n?

Odpovedať • Páči sa mi to

Anonymný Oct 22, 2021

> Súbory budú obsahovať iba znaky s hodnotou 0 < c <= 127, ak by bol v súbore aj iný znak, tak správanie je nešpecifikované.

\n je znak ako kazdy iny, c++ v tom nerobi rozdiel, preto nevidim problem v tom, aby sa mohol vyhladavat. Ak si to chces vyskusat, tak daj argument do uvodzoviek.

Odpovedať • Páči sa mi to

Anonymný Oct 22, 2021

No ja sa snazim o implementaciu pomocou KMP kvoli rychlosti a efektivite. Tazko sa to vysvetluje v komentari, ale bolo by dost komplikovane prepocitavat poziciu zaciatku substringu (teda stlpec aj riadok) - musel by som si ukladat pocet znakov v kazdom riadku, nasledne odpocitavat znaky z daneho stlpca aby som sa dostal k jeho pozicii a teda by to nebolo velmi pamatovo efektivne. Preto by som rad vedel s istotou ci toto mam riesit alebo nie.

Odpovedať • Páči sa mi to

Peter Koscelanský [Administrator] Oct 26, 2021

Ano moze byt aj \n.

Odpovedať • Páči sa mi to

Anonymný Oct 28, 2021

Ako máme rozlišovať medzi tým či substring obsahuje \n ako znaky \ a n a \n ako newline?

Odpovedať • Páči sa mi to

Anonymný Oct 28, 2021

Myslím teraz pri načítavaní z argumentov, keďže tam nevieme načítať \n ako jeden symbol

Odpovedať • Páči sa mi to

Peter Koscelanský [Administrator] Oct 29, 2021

Ako? Nerozumiem, vy mate pouzit to co vam pride v `argv` a tam uz nic neinterpretujte. Ono akym sposobom vam tam ten parameter pride je jedno, vy iba zoberte ten parameter a jeho sucastou moze byt aj `\n` teda new line, ale nie ako dva znaky `\` a potom `n`, ale ako jeden znak, ktory reprezentuje LINE FEED, teda `'\n'`.

Odpovedať • Páči sa mi to

Anonymný Oct 22, 2021

1. Mozeme pouzivat vsetky kniznice na pracu so stringami? Lebo ak by som nechcel ist na rychlost, tak s pomocou std::find je riesenie velmi jednoduche. A myslim si, ze nebol zamer tohto zadania, aby bolo lahke. Alebo tu skor ide o to, aby sme sa zamerali hlavne na osetrenie vsetkych moznych chybovych a nedefinovanych stavov?
2. Co sa tyka poctu nacitanych znakov na jeden raz, tak kolko by to cca malo byt? Je nejake odporucane maximum?
3. Ked pracujem na linuxe, nemusim sa bat, ze mi na testoch pustite CRLF subor?

Odpovedať • Páči sa mi to

Peter Koscelanský [Administrator] Oct 26, 2021

1. Zo standardnej kniznice mozes pouzivat vsetko az na niektore viacvlaknove konstrukcie (tie nepojdu vybuildovat, takze to si rychlo vsimnes).
2. To je tazko povedat, o tom si na internete diskusie, idealne asi v radovo tisickach a nie viac ako megabajt.
3. Nie, resp. ak ano tak \r je pre teba normanej znak.

Odpovedať • Páči sa mi to

Anonymný Oct 28, 2021

Zdravím, je potrebné nejako ošetrovať obsah parametru X?

Odpovedať • Páči sa mi to

Peter Koscelanský [Administrator] Oct 28, 2021

Iba velkost, budu tam len znaky ktore su ascii <= 127, ale ak tam bude nieco ine, tak je spravanie nespecifikovane, teda urobite hocico, okrem nedefinovaneho spravania.

Odpovedať • Páči sa mi to

Anonymný Oct 29, 2021

Vie niekto ako poslat \n ako vstupny argument vo visual studiu?

Odpovedať • Páči sa mi to

Anonymný Oct 29, 2021

Ja som to testoval tak, že som ho natvrdo vložil do kódu

Odpovedať • Páči sa mi to

Peter Koscelanský [Administrator] Oct 29, 2021

+1 to je super pristup. My pouzivame `fork` a `exec` na spustenie procesu a tam sa da dat vsetko.

Odpovedať • Páči sa mi to

Anonymný Oct 29, 2021

500 IQ. Dakujem, fakt mi to nenapadlo :DD.

Odpovedať • Páči sa mi to