

Tarefa 15 - Nuvem de palavras

Publicado em 04/11/2021

Prazo de entrega recomendado: 14/11/2021

Você deve selecionar elementos para uma nuvem de palavras. Será necessário implementar uma tabela hash.

Juan é um compositor e está interessado em saber de que falam as músicas. Assim, ele decidiu criar uma nuvem de palavras. Uma nuvem de palavras é um agrupamento de palavras-chaves representadas visualmente de forma que a relevância de uma palavra é refletida no seu tamanho. Eis um exemplo de nuvem de palavras:



Juan gostaria de criar uma nuvem de palavra para cada artista musical, mas como existem muitos artistas, ele precisa de sua ajuda realizar essa tarefa. Para cada artista, ele conseguiu de um amigo produtor um arquivo com várias letras de músicas. Ele gostaria de contar quantas vezes cada palavra é repetida em cada arquivo e descobrir quais mais aparecem, independentemente da capitalização.

Uma dificuldade é que normalmente as palavras mais frequentes são artigos, preposições, interjeições, como um, uma, de, ah em português, the, on, uh em Inglês, etc. Em processamento de texto, essas palavras são conhecidas como *stop words*. Elas não são tão interessantes para inspirar um compositor. Além disso, letras sozinhas devem ser ignoradas. Juan não gostaria de ficar conhecido como o compositor do próximo hit do o, o, o.

Outra dificuldade é que as letras podem ter pontuação ou caracteres especiais como parênteses (), aspas ", etc. Assim, uma palavra é uma sequência de letras somente e esses caracteres devem ser removidos. Além disso, duas palavras podem ser contraídas em uma usando um sinal de apóstrofo ' ou rotelro -, que devem ser removidos. Por exemplo, d'e'le deve ser interpretado como dele e can't deve ser interpretado como cant ou li-i-i-ike como liiiiike.

Sua tarefa é escrever um programa `nuvem.c` que descubra as 50 palavras de interesse mais frequentes.

Entrada

A primeira linha contém uma estimativa para o número de palavras distintas nos idiomas do arquivo. A segunda linha contém o número de stop words que estão depois desse número. Todas as linhas seguintes contém as letras de várias músicas diferentes. A maior palavra tem menos que 50 caracteres.

Exemplo de entrada

```
171146
142 a am an and any are arent as at be been but by cant cannot could couldnt di

Her name is Lara, we learned a lot, ah
How to do it, like we do it like we wanna
We just know, we just know
I ain't one-sided, I'm open-minded
I'm 50/50 and I'm never gonna hide it
You should know, eh, you should know, ayy

All summer, we've been in the 'Bu
'68 Chevy with nothin' to do
Just rollin' J's, kush lovin'
And last night, yeah, we got with the dude
I saw him, he was lookin' at you
So I said, "Hey," kush lovin'

Sometimes, I just wanna kiss girls, girls, girls
Red wine, I just wanna kiss girls, girls, girls
Sometimes, I just wanna kiss girls, girls, girls
Red wine, I just wanna kiss girls, girls, girls
Girls, girls, girls, girls, girls

Yeah, you know I tamed it, and then I named it
I put the lion in the cage and then I laid with
Her all night (All night), her all night, yeah
I'm the hunter and she the prey, yeah
I'm the thriller and the killer and the saviour
Up all night, we up all night, yeah (Do it one more time)
All summer, we've been in the 'Bu (Oh)
'68 Chevy with nothin' to do
Just rollin' J's, kush lovin' (Roll it out, roll it out)

Sometimes, I just wanna kiss girls, girls, girls
Red wine, I just wanna kiss girls, girls, girls
(Yeah-yeah, yeah-yeah)
Sometimes, I just wanna kiss girls, girls, girls
(You know that I do)
Red wine, I just wanna kiss girls, girls, girls
Girls, girls, girls, girls, girls

She gettin' down with me, yeah-ah
She gettin' down with me, yeah-ah
She gettin' down with me, yeah-ah
Oh, we can go up
She gettin' down with me, yeah-ah
She gettin' down with me, yeah-ah
She gettin' down with me, yeah-ah (Rita, look, Cardi)

Now I could be your lipstick, just for one night (One night)
Girls just wanna have fun and have their funds right (Yeah)
I mean, say my name, say my name, say my name (Say my name)
It tastes good just rollin' off your tongue, right? (Hurr)
I put this MAC on your lips, so pucker up (Mwah)
We ain't never heard of you 'cause you ain't done enough (No)
And I don't gotta introduce myself (Cardi!)
I'm too sexy, I seduce myself (Bardi!)
Seven-figure, never need a nigga (Nope)
I steal your bitch, have her down with the scissor
Tonight, I don't want a dog, I want a kitten (Eeeow)
I might French a girl from Great Britain
Sometimes, I just wanna kiss girls, girls, girls (Oh yeah)
Red wine, I just wanna kiss girls, girls, girls (Uh, yeah)
Sometimes, I just wanna kiss girls, girls, girls (Red wine, red wine)
Red wine, I just wanna kiss girls, girls, girls (Oh, oh, oh)
Girls, girls, girls, girls, girls

She gettin' down with me, yeah-ah
(Down with me)
She gettin' down with me, yeah-ah
(Shе gettin' down on me, ahh)
She gettin' down with me, yeah-ah (Yeah)
(Down with me, yeah)
She likes what she likes (Aha)
She gettin' down with me, yeah-ah (Yeah)
She gettin' down with me, yeah-ah (Ah-ha)
She gettin' down with me, yeah-ah
Girls, girls, girls, girls, girls
Hehehehe-aha (Hahahaha)
```

Dica

Você pode usar a biblioteca `ctype.h` para converter letras maiúsculas em minúsculas.

Saída

A saída contém as 50 palavras que mais aparecem em minúsculas, em ordem decrescente de frequência. Ao lado de cada palavra deve estar indicada a frequência. Palavras com a mesma frequência devem ser ordenadas lexicograficamente.

Exemplo de saída

```
girls 57
just 19
down 16
wanna 14
gettin 13
kiss 12
yeahah 12
yeah 11
night 8
red 8
wine 8
all 7
know 6
sometimes 6
say 5
name 4
your 4
aint 3
kush 3
lovin 3
never 3
one 3
rollin 3
bu 2
cardi 2
chevy 2
js 2
like 2
likes 2
myself 2
nothin 2
put 2
right 2
roll 2
summer 2
want 2
yeahyeah 2
aha 1
ahh 1
ahha 1
ayy 1
bardi 1
bitch 1
britain 1
cage 1
can 1
cause 1
dog 1
done 1
dude 1
```

Crerérios

Deve-se utilizar uma tabela de hashing para descobrir a frequência de cada palavra. Você pode escolher uma função de hashing de preferência, mas o uso de funções de hashing inadequadas pode levar a perda de pontos. Para encontrar as palavras mais repetidas, utilize um algoritmo de ordenação eficiente (*HeapSort*, *MergeSort* ou *Quicksort*).

Correção

Esta tarefa será corrigida automaticamente sempre que você realizar um `git push`. Depois de terminada a tarefa, deve-se utilizar o botão na interface de notas para solicitar a correção de um monitor.

Turma AB: O peso desta tarefa será 5.

Turma E: Você deverá apresentar esta tarefa a um monitor PED. Para isso, você deve procurar atendimento em algum horário com monitor PED e digitar *apresentar 15* no canal *fila-apresentar*.