

You are to write an interactive program that, given a sequence of Wikipedia excerpts (see example below), guesses the language of each, in turn. After each guess, your program is given the correct answer, so that it may learn to make better guesses the longer it plays.

Each language is represented by a number L between 0 and 55. Each excerpt has exactly 100 symbols, represented as an array E of 100 integers between 1 and 65 535. These integers between 1 and 65 535 have been assigned arbitrarily, and do not correspond to any standard encoding.

You are to implement the procedure $\text{excerpt}(E)$ where E is an array of 100 numbers representing a Wikipedia excerpt as described above. Your implementation must call $\text{language}(L)$ once, where L is its guess of the language of the Wikipedia edition from which E was extracted. The grading server implements $\text{language}(L)$, which scores your guess and returns the correct language. That is, the guess was correct if $\text{language}(L) = L$.

The grading server calls $\text{excerpt}(E)$ 10 000 times, once for each excerpt in its input file. Your implementation's accuracy is the fraction of excerpts for which $\text{excerpt}(E)$ guessed the correct language.

You may use any method you wish to solve this problem. *Rocchio's method* is an approach that will yield accuracy of approximately 0.4. Rocchio's method computes the similarity of E to each language L seen so far, and chooses the language that is most similar. Similarity is defined as the total number of distinct symbols in E that appear anywhere amongst the previous excerpts from language L .

Note that the input data have been downloaded from real Wikipedia articles, and that there may be a few malformed characters or fragments of text. This is to be expected, and forms part of the task.

Example.

For illustration only, we show the textual representation of excerpts from 56 language-specific editions of Wikipedia.

0. Yshokkie word meestal in Kanada , die noorde van die VSA en in Europa gespeel. Dit is bekend as 'n b
1. وهو المنتج الذي يجعل المنظم لا يكسب ربحا ولا يخسر ويحصل على دخل يكف ، Marginal Producer المنتج الحدي
2. "BAKILI" Futbol Klubu 1995-ci ildə Misir Səttar oğlu Əbilov tərəfindən yaradılmış və həvəskar futbol
3. Квинт Фулвий Флак (Quintus Fulvius Flaccus; † 205 пр.н.е.) е политик и генерал на Римската републик
4. ইন্ডিয়ান ইনস্টিটিউট অফ সোশ্যাল ওয়েলফেয়ার অ্যান্ড বিজনেস ম্যানেজমেন্ট (সংক্ষেপে আইআইএসডব্লিউবিএম)
5. 5. juni (lipanj) (5.6.) je 156. dan godine po gregorijanskom kalendaru (157. u prestupnoj godini
6. La Caunette és un municipi francès , situat al departament de l' Erau i a la regió de Llenguadoc-Ros
7. Praha je malé městečko v Texasu , které leží cca 85 km na jihozápad od Austinu . Bylo založeno
8. Graeme Allen Brown (født 9. april 1979 i Darwin , Northern Territory , Australien) er en australsk
9. Der Plattiger Habach (3.214 m ü. A. , nach anderen Angaben nur 3.207 m [1])
10. Το Νησί Γκρέιτ Μπάρριερ (Αγγλικά : Great Barrier Island , Μαορί : Motu Aotea) είναι νησί στα βόρει
11. Sid Bernstein Presents... is a 2010 feature-length documentary film by directors Jason Ressler and E
12. El término latino lex loci celebrationis aplicado al derecho internacional privado quiere decir: "le
13. Apollo 5 oli kosmoselaev , mis sooritas Apollo programmi teise mehitamata lennu. Lennu käigus testit
14. 1350 : هزار و سیصد و پنجاهمین سیارک کشف شده‌اس (TA نامگذاری : 1934 ، Rosselia ، سیارک ۱۳۵۰ (به انگلیسی : 1350
15. V. I. Beretti (myös Vikenty Ivanoviš Beretti , alk. Vincent Beretti ; 1781 Milano Italia – 18. elok
16. Le 5 e bataillon de parachutistes vietnamiens (ou 5 e BPVN ou encore 5 e Bawouan) est une unité par
17. Amina Sarauniyar Zazzau,, wadda ta rayu daga shekarar 1533 zuwa 1610, daya ce daga cikin 'ya'ya biyu
18. ה מתמטיקה , השערת רימן היא השערה שהציע בשנת 1859 ה מתמטיקאי ברנרד רימן , מגדולי המתמטיקאים של אותה ע
19. Sudski proces Doe protiv Boltona je sudski proces iz 1973 . godine kojim je američki Vrhovni sud uki

Day 1 Task4

Languages

- a viewable representation (in UTF-8) of the 100 symbols that you can read in your text editor or Firefox web browser. This viewable representation is for your convenience only, and is not intended to be used as input for your program.

The official grader uses 10 000 different excerpts, selected in the same way from the same 56 Wikipedia editions. However, the grader assigns a different number between 0 and 55 to each language, and a different number between 1 and 65 535 to each symbol.

1 Subtask [30 points]. Your submission must achieve accuracy of 0.3 or better on the grading server.

2 Subtask [up to 80 points]. Your score will be $114(\alpha - 0.3)$, rounded to the nearest integer, where α is the accuracy of your submission.

Implementation Details.

- Use the *RunC programming and test environment*
- Implementation folder: `/home/ioi2010-contestant/language/` (prototype: *language.zip*)
- To be implemented by contestant: `lang.c` or `lang.cpp` or `lang.pas`
- Contestant interface: `lang.h` or `lang.pas`
- Grader interface: `grader.h` or `graderlib.pas`
- Sample grader: `grader.c` or `grader.cpp` or `grader.pas` and `graderlib.pas`
- Sample grader input: `grader.in.1`
Note: Each line of input contains: a two-character language code; an excerpt represented as 100 numbers separated by spaces; the text representation of the excerpt.
- Expected output for sample grader input: *If the implementation calls language as specified for each of the 10 000 examples, the sample grader will output `OK alpha` where α is the accuracy.*
- Compile and run (command line): `runc grader.c` or `runc grader.cpp` or `runc grader.pas`
- Compile and run (gedit plugin): *Control-R*, while editing any implementation file.
- Submit (command line): `submit grader.c` or `submit grader.cpp` or `submit grader.pas`
- Submit (gedit plugin): *Control-J*, while editing any implementation or grader file.