ОДСЕК ЗА СОФТВЕРСКО ИНЖЕЊЕРСТВО

АЛГОРИТМИ И СТРУКТУРЕ ПОДАТАКА 2

2020-2021

- трећи домаћи задатак -

Опште напомене:

1. Пре одбране сви студенти раде тест знања који се ради на рачунару коришћењем система Moodle (http://elearning.rcub.bg.ac.rs/moodle/). **Сви студенти треба да креирају налог и пријаве се на курс пре почетка лабораторијских вежби.** Пријава на курс ће бити **прихваћена и важећа** само уколико се студент региструје путем свог налога електронске поште на серверу **mail.student.etf.bg.ac.rs**.
2. Домаћи задатак 3 састоји се од једног програмска проблема. Студенти проблем решавају **самостално**, на програмском језику C++. **Дозвољено је коришћење готових структура података из STL програмског језика C++.**
3. Реализовани програм треба да комуницира са корисником путем једноставног менија који приказује реализоване операције и омогућава сукцесивну примену операција у произвољном редоследу.
4. Унос података треба омогућити било путем читања са стандардног улаза, било путем читања из датотеке.
5. Решења треба да буду отпорна на грешке и треба да кориснику пруже јасно обавештење у случају детекције грешке.
6. Приликом оцењивања, биће узето у обзир рационално коришћење ресурса. **Примена рекурзије се неће признати као решење проблема које може освојити максималан број поена.**
7. За све недовољно јасне захтеве у задатку, студенти треба да усвоје разумну претпоставку у вези реализације програма. Приликом одбране, демонстраторе треба обавестити која претпоставка је усвојена (или које претпоставке су усвојене) и која су ограничења програма (на пример, максимална димензија низа и слично). Неоправдано увођење ограничавајуће претпоставке повлачи негативне поене.
8. Одбрана трећег домаћег задатка ће се обавити према распореду који ће накнадно бити објављен на сајту предмета. Пријава за одбрану биће омогућена преко Moodle система. Детаљније информације биће објављене на предметном сајту.
9. Предаја домаћих ће бити омогућена преко Moodle система до **уторка, 29.12.2020. у 23:59**. Детаљније информације су објављене на предметном сајту.
10. За решавање задатака који имају више комбинација користити следеће формуле. (**R** – редни број индекса, **G** – последње две цифре године уписа):

$$\mathbf{i = (R + G) \bmod 2 + 1}$$

11. Предметни наставници задржавају право да изврше проверу сличности предатих домаћих задатака и коригују освојени број поена након одбране домаћих задатака, као и да пријаве теже случајеве повреде Правилника о дисциплинској одговорности студената Универзитета у Београду Дисциплинској комисији Факултета.

18.12.2020. године *Са предмета*

Задатак 1 – предвиђање текста коришћењем *trie* стабала [100 поена]

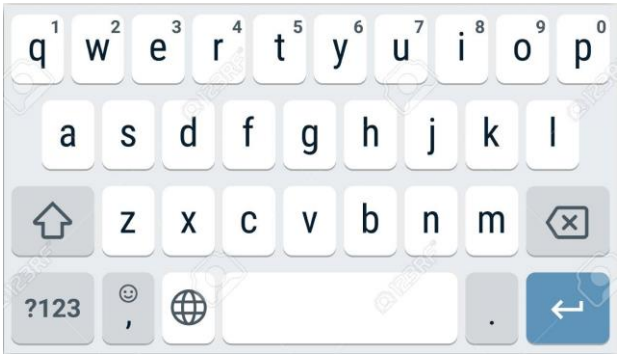
Системи за предвиђање (предикцију) текста се користе код многих преносних уређаја са ограниченом величином тастатуре или екрана да би се убрзао унос текста приликом куцања порука и сл. Корисник најчешће уноси део текста (речи), а систем за предвиђање му нуди предлоге речи. Такође, код неких технологија уноса, као што су *Swype* или *Microsoft SwiftKey*, корисник има могућност да „црта“ речи по виртуелној тастатури, а систем врши предикцију на основу тог уноса који може бити непрецизан.

Системи за предикцију текста су обично базирани на речнику који се попуњава речима из задатог језика и врше сугерисање речи или корекција заснованих на том речнику. Речник се често имплементира помоћу префиксних (*trie*) стабала.

Системи за предикцију текста обично постају "паметнији" (прецизнији) приликом дужег коришћења од стране корисника. У зависности од уноса корисника са тастатуре, систем може предвидети већи број речи, а кориснику се онда нуди неколико речи са највећом вероватноћом појављивања што систем учи на основу претходног понашања корисника. То се на најједноставнији начин постиже памћењем фреквенције (броја) појављивања појединачних речи и уређивањем речи по тим фреквенцијама приликом давања предикције.

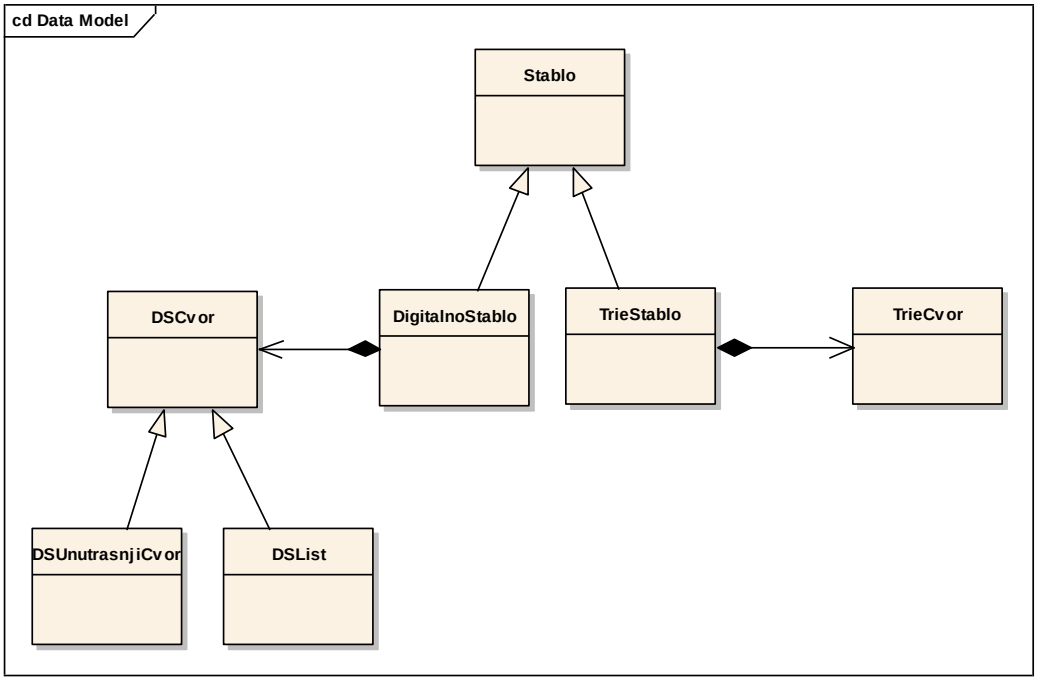
Користећи већ реализован пакет класа за рад са *trie* стаблом који је приказан на дијаграму у прилогу, модификовати одговарајуће класе тако да подрже предикцију текста уз памћење фреквенције појављивања појединачних речи.

Предиктор треба да прихвата стрингове који се састоје од корисниковог уноса који се може састојати од целе речи, дела речи или погрешно написане речи у неком делу као последица коришћене виртуелне тастатуре са одговарајућим распоредом слова (QWERTY, QWERTZ, AZERTY, DVORAK). Сматрати да се стрингови састоје само од малих и великих слова енглеског алфабета и да корисник употребљава стандардни QWERTY распоред, као на слици. Приликом предикције резултата за погрешно написане речи, претрагу ограничити на највише три погрешно написана слова у речи. Приликом замене потенцијално погрешних слова, ограничити се на слова у суседству на виртуелној QWERTY тастатури.



Након сваког претраживања, систем треба да врати највише три речи које су резултат предикције, поређане по вероватноћама појављивања. Уколико корисник унесе тачну реч или је задати префикс јединствен, довољно је исписати само један резултат предикције.

Изворни код, који одговара дијаграму са слике, доступан је у архиви која се налази на сајту предмета заједно са овим документом. Програм треба проширити тако да омогући читање и формирање речника из датотеке. За тестирање је доступан корпус текстова на енглеском језику у архиви `text.zip` који се такође налазе у поменутој архиви. Тај корпус треба користити за формирање иницијалног речника и фреквенција појављивања речи.



У зависности од редног броја проблема који се решава **i**, потребно је модификовати једну од следећих структура података из реализованог пакета класа и помоћу ње реализовати речник у оквиру система за предвиђање текста:

1. *Trie* стабло
2. Дигитално стабло

Реализовати следеће операције за рад са системом за предвиђање текста:

1. **[10 поена]** Стварање празног речника и уништавање речника
2. **[20 поена]** Стварање речника на основу задате датотеке или скупа датотека
3. **[30 поена]** Претраживање речи, уметање и ажурирање фреквенција
4. **[30 поена]** Предвиђање речи на основу задатог стринга
5. **[10 поена]** Главни програм који врши комуникацију са корисником

Главни програм треба да демонстрирати рад са реализованим системом за предвиђање текста путем интерактивног менија. Потребно је омогућити читање речника из датотеке или скупа (списка) датотека, а затим репетитивно уношење стрингова за претрагу и исписивати резултате предикције. За потребе тестирања омогућити формирање речника од најмање 100 000 речи из приложеног корпуса и то реализовати као посебну опцију за тестирање решења.