Zadanie č.1

15 bodov

Termín odovzdania: do 26.03.2023 do 23:59 hod.

Úloha:

Máme súbor *dictionary.txt*. Súbor obsahuje slová z nejakého anglického textu spolu s frekvenciou ich výskytu. Slová obsahujú iba malé písmená anglickej abecedy, t.j. ASCII znaky 97 až 122. Jeden riadok súboru obsahuje frekvenciu výskytu slova a samotné slovo. Frekvencia a slovo sú oddelené medzerou. Slová sú v súbore usporiadané podľa frekvencie výskytu: slovo s najvyššou frekvenciou je v prvom riadku, slovo s najnižšou frekvenciou výskytu je v poslednom riadku.

Vašou úlohou je zostrojiť **optimálny** binárny vyhľadávací strom pre vyhľadávanie slov s frekvenciou výskytu ostro väčšou ako 50 000. Ďalej budeme používať termíny a notáciu z kapitoly 15.5 z knihy Introduction to Algorithms od autorov Cormen, Leiserson, Rivest a Stein. Pri vytváraní stromu postupujte nasledovne:

- Kľúče budú slová s frekvenciou výskytu ostro väčšou ako 50 000.
- Na slovách uvažujte lexikografické usporiadanie.
- Pravdepodobnosť p_i, že vyhľadávame kľúč k_i, vypočítajte ako podiel frekvencie výskytu slova k_i a súčtu frekvencií výskytu všetkých slov v súbore *dictionary.txt*.
- Uvažujeme, že budeme vyhľadávať iba slová zo súboru *dictionary.txt*. Pravdepodobnosť q_i, že vyhľadávame slovo, ktoré je v lexikografickom usporiadaní medzi k_i a k_{i+1}, preto vypočítajte ako podiel súčtu frekvencií výskytu tých slov z *dictionary.txt*, ktoré sú v lexikografickom usporiadaní medzi k_i a k_{i+1}, a súčtu frekvencií výskytu všetkých slov v *dictionary.txt (pozri vzorec nižšie)*. Analogicky vypočítajte aj pravdepodobnosti q_0 a q_n.

$$q_{-i} = \frac{\text{súčet frekvencií výskytu tých slov, } ktoré\ sú\ v\ lex.usporiadaní\ medzi\ k_i\ a\ k_{i+1}}{\text{súčet frekvencií výskytu všetkých slov}}$$

Okrem toho, vytvorte funkciu **pocet_porovnani()**. Vstupom do funkcie bude reťazec. Funkcia vráti počet porovnaní, ktoré sa vykonajú počas hľadania vstupného reťazca v zostrojenom optimálnom binárnom vyhľadávacom strome.

Odovzdavanie:

Do vytvoreného miesta odovzdania odovzdajte zdrojové súbory.
Hodnotia sa len zadania odovzdané do AISu !!!

Pre získanie bodov zo zadania je potrebné riešenie odprezentovať (na cvičení) v termíne po dohode s cvičiacimi !!!

Hodnotenie:

15 bodov - správne vytvorený optimálny binárny vyhľadávací strom a správne fungujúca funkcia pocet_porovnani(). Študent(-ka) musí vedieť podrobne popísať postup, ktorý použil(-a) pri vytváraní stromu a funkcie.

V prípade, že študent(-ka) nevie vysvetliť správne a dostatočne fungovanie svojho riešenia, riešenie sa hodnotí nižším počtom bodov, možno až 0 bodmi !!!