TEHTÄVÄMÄÄRITTELY

Aihe: Sanaindeksoija

Sovellus lukee käyttäjän antamista tekstitiedostoista sanat hakupuurakenteeseen, jonka jälkeen käyttäjän on mahdollista kysellä sovellukselta sanojen tai niiden yhdistelmien (haluamiensa merkkijonojen) esiintymisestä kyseisissä tekstitiedostoissa.

Sovelluksen toiminnallisuus:

- tekstitiedoston sisäänluku
    1. sisäänluettavien tiedostojen maksimilukumäärän kysyminen
        a) jos lukumäärä on 0 tai pienempi, ohjelman päättäminen
        b) jos lukumäärä on suurempi kuin 0
            2. tiedoston nimen kysyminen
            3. tiedoston nimen tarkastus
                a) jos nimi on tyhjä ja sisäänlukuja on tehty, siirtyminen sanahakuun
                b) jos nimi on tyhjä, mutta sisäänlukuja ei ole tehty, siirtyminen tilaston tulostamiseen
                c) jos nimi ei ole tyhjä
                    4. tiedoston löytymisen tarkistus
                        a) jos tiedostoa ei löydy, ilmoitus käyttäjälle ja paluu kohtaan 2.
                        b) jos tiedosto löytyy
                            5. tiedoston läpiluku ja tallennus hakupuuhun
                            6. hakupuun tulostamisen kysyminen
                                a) jos vastaus on K, hakupuun tulostaminen
                            7. jos annettu tiedostolukumäärä ei ole vielä täynnä, paluu kohtaan 2.

- sanahaku
    1. haettavan merkkijonon pyytäminen
    2. haettavan merkkijonon tarkastus
        a) jos merkkijono on tyhjä, siirtyminen tilaston tulostamiseen
        b) jos merkkijono ei ole tyhjä
            3. annetun merkkijonon haku hakupuurakenteesta
            4. osumien tulostaminen käyttäjälle
            5. osumien lukumäärän tulostaminen käyttäjälle
            6. hakua tarkentavan merkkijonon pyytäminen
            7. hakua tarkentavan merkkijonon tarkastus
                a) jos merkkijono on tyhjä, paluu kohtaan 1.
                b) jos merkkijono ei ole tyhjä
                    8. annetun merkkijonon haku hakupuurakenteesta
                    9. aiemman hakutuloksen ja uuden hakutuloksen leikkauksen tekeminen
                    10. osumien tulostaminen käyttäjälle

11. osumien lukumäärän tulostaminen käyttäjälle
12. paluu kohtaan 6.

- tilaston tulostaminen
    1. sisäänluettujen tiedostojen, haettujen merkkijonojen ja osumien lukumäärien tulostaminen käyttäjälle
    2. käytön päättäminen

Toteutettavat algoritmit ja tietorakenteet:

- String-taulukko, johon tallennetaan tiedostojen nimet
- Trie-puu, johon tallennetaan maksimissaan yhden rivin pituisia merkkijonoja
    - rajaukset
        - tallennettavat merkkijonot aloitetaan sanojen alusta, joten haku ei löydä merkkijonoja sanan keskeltä
        - tekstitiedostoa käsitellään rivi kerrallaan, joten haettavan merkkijonon tulee sijaita yhdellä rivillä, jotta se löytyy
        - sana alkaa rivin alusta tai tyhjän merkin jälkeen
        - sana päättyy rivin loppuun tai tyhjään merkkiin
        - puuhun tallennetaan vain suomalaisen aakkoston kirjaimia ja numeroita sekä rivillä sanojen välissä olevia tyhjiä merkkejä, joten kelvollisia merkkejä on 40 kpl
    - Trie-puu valittu haun talletusrakenteeksi, koska se mahdollistaa helposti myös sanan aluilla ja vastaavasti myös pitkillä merkkijonoilla hakemisen. Lisäksi Trie-puu vaikuttaa mielenkiintoiselta rakenteelta.
- Trie-puun solmuolio, joka sisältää merkin, alun linkitettyyn listaan, jossa ovat kyseiseen solmuun päättyvät esiintymät tekstissä sekä viittaukset solmun lapsisolmuihin
- linkitetty lista, johon tallennetaan tieto tietyn merkkijonon esiintymistä tekstitiedostoissa
    - Linkitetty lista valittu tekstin esiintymien tallennusrakenteeksi, koska tallentaminen sujuu helposti sijoittamalla uuden solmun aina listan viimeiseksi ja koska yksittäistä tietoa ei ole tarvetta lukea listasta, vaan kun listan tiedot luetaan, ne luetaan alusta loppuun.
- 40-soluinen maksimitaulukko, jonka kukin alkio toimii aloitussolmuna seuraavalle linkitetylle listalle
    - Maksimitaulukko valittu lapsisolmujen tallennusrakenteeksi, koska tallennettavia arvoja on rajallinen määrä ja niille on helppo määrittää merkin Unicode-arvoon perustuva yksilöllinen sijainti maksimitaulukossa. Tallentaminen ja lukeminen on nopeaa.

Syötteet:

- ascii-muotoisia tekstitiedostoja
- käyttäjä antaa sovellukselle tiedoston nimen
- tiedoston on sijaittava samassa hakemistossa sovelluksen kanssa, jotta sovellus sen löytää

Tavoiteltava aika- ja tilavaativuus:

• yhden tekstitiedoston sisäänluku	aikavaativuus O(logn)	tilavaativuus O(n logn)
• yksi sanahaku	aikavaativuus O(logn)	tilavaativuus O(n logn)
• tilaston tulostaminen	aikavaativuus O(1)	tilavaativuus O(1)

Lähteet:

- kurssisivulla annetut materiaalit
- Tira-kurssin (kevät 2012) materiaalit
- Java 7 API Specification